ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Progression Detection, Recognition and Prediction of Cancer Using CDS Framework

Gowsalya N [1],Divya A [2],Radha K [3],Sudha K [4]

[1]PG Student, Dept. of CSE, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

[2]PG Student, Dept. of CSE, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

[3]Associate Professor, Dept. of C.S.E, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

[4]Associate Professor, Dept. of C.S.E, Muthayammal Engineering College, Rasipuram, Tamilnadu, India

ABSTRACT: Clinical Decision Support (CDS) aids in early diagnosis of liver cancer, a potentially fatal disease prevalent in both developed and developing countries. Our research aims to develop a robust and intelligent clinical decision support framework for disease management of cancer based on legacy Ultrasound (US) image data collected during various stages of liver cancer. The proposed intelligent CDS framework will automate real-time image enhancement, segmentation, disease classification and progression in order to enable efficient diagnosis of cancer patients at early stages. The CDS framework is inspired by the human interpretation of US images from the image acquisition stage to cancer progression prediction. Specifically, the proposed framework is composed of a number of stages where images are first acquired from an imaging source and pre-processed before running through an image enhancement algorithm. The detection of cancer and its segmentation is considered as the second stage in which different image segmentation techniques are utilized to partition and extract objects from the enhanced image. The third stage involves disease classification of segmented objects, in which the meanings of an investigated object are matched with the disease dictionary defined by physicians and radiologists. In the final stage; cancer progression, an array of US images is used to evaluate and predict the future stages of the disease. For experiment purposes, we applied the framework and classifiers to liver cancer dataset for 200 patients. Class distributions are 120 benign and 80 malignant in this dataset.

KEYWORDS: CDS; Image Segmentation; Classification; SVM; LESH; WEKA; Ultrasound; Liver Cancer

I. Introduction

The Human body is made up of trillions of living cells. Normal body cells grow, divide into new cells, and die in an orderly fashion. Cancer begins when cells in a part of the body start to grow out of control. There are many kinds of cancer, but they all start because of out-of-control growth of abnormal cells. Liver cancer is the fifth most common cancer in men and the seventh in women. The regions of high incidence are Eastern and South-Eastern Asia, Middle and Western Africa. Low rates are estimated in developed regions, with the exception of Southern Europe where the incidence in men is significantly higher than in other developed regions. The liver is the largest gland and largest internal organ in the human body. Liver diseases can be classified into two main categories, focal diseases and diffused diseases. The focal diseases are where abnormalities are concentrated within a small area of liver parenchyma, whereas the diffused diseases are where the abnormalities are distributed over the whole extent of liver tissues. It has been proved by pathology and histology that the severity of liver diseases are closely related with liver fibrosis progressions. Thus most cases of liver diseases can be sorted into three classes, including hepatitis, fatty and cirrhosis, according to their different fibrosis stages. For a long time, Liver biopsy is the clinically golden standard for diagnosing chronic liver diseases and for guiding further medication. However, it is usually associated with high risks of infection and complication. So developing a non-invasive, reliable method for detecting the status of liver fibrosis is urgently required. B-mode ultrasonic imaging has been used to detect liver diseases because the change of fibrosis and abnormalities of liver tissues can be clearly reflected by the ultrasound images. The fact that the ultrasound images are usually examined by doctors` eyes makes the diagnosis quite subjective, because both the doctors` experience and the quality of images may vary in particular situation. So the high reliability and accuracy of algorithm analysis makes it attractive and reasonably becomes the research focus. Classifications of ultrasound liver lesion images are very difficult task in the image processing. In the medical field computer are now being used virtually in every aspect of modern medicine. Computers are used widely in medical research, where there is a vital need for better microelectronic sensors for data acquisition. Diagnosis by ultrasound imaging is a cost effective approach to ascertain the disease in earlier stage. Liver diseases are considered seriously because of the liver's vital importance to human beings. There are two classes of liver tumors: benign and malignant. Ultrasound image is a powerful tool for characterizing the state of soft tissues for medical diagnostic purposes. Ultrasound has been extremely valuable in differentiating a simple liver lesion from other liver masses. An approach has been made in this research to design a diagnostic classifier system for liver lesion in ultrasound images using image texture features in non-invasive manner. Image processing modifies pictures to improve them (enhancement, restoration), extract information (analysis, recognition), and change their structure (composition, image editing). Images can be processed by optical, photographic, and electronic means, but image processing using digital computers is the most common method because digital methods are fast, flexible, and precise.

II. RELATED WORK

Hepatocellular carcinoma (HCC) is one of the most deadly cancers worldwide. Current advances in proteomic approaches facilitate several proteome wide studies in identifying markers as well as insights into the mechanisms of HCC development. Facing a relatively large amount of data in the proteome, modern high-order data mining techniques may provide a systematic way in search for meaningful and biological significant pattern and trends hidden in the proteomic dataset. In this study, a proteomic dataset of 132 HCC related tumour and non-tumour samples, each consisting of 1433 variables was used for construction and evaluation of classification models based on artificial neural network (ANN) and classification and regression trees (CART) algorithm. Both algorithms successfully segregate samples into corresponding phenotypes with high sensitivities and specificities (ANN: 89.4%, 89.4%; CART: 80.3%, 80.3%), enlightening the usefulness and possibilities of data mining techniques in genomic and proteomic expression profiling studies[1].

Fuzzy enhancement is applied in computer aided diagnosis of liver cancer from B mode ultrasound images as a pre-processing procedure in this paper. It was evaluated with three classifiers including K means, back propagation neural network and support vector machine using 25 features from first order statistic (FOS), gray-level cooccurrence matrix (GLCM), gray-level run-length matrix (GLRLM), Grey level dependant matrix (GLDM) and LAWS. In the analysis of 166 normal liver tissues, 30 hemangioma and 60 malignant tumor, our method improved the classification accuracy of three classifiers (K means, BP neural network and support machine vector) in distinguishing liver cancer, hemangioma and normal liver cancer from B mode ultrasound images. It is proved that fuzzy enhancement as an efficient preprocessing procedure could be used in the computer aided diagnosis system of liver cancer[2].

A method for reducing speckle noise in medical ultrasonic images is presented. it is called the adaptive weighted median filter (awmf) and it is based on the weighted median, which originates from the well-known median filter through the introduction of weight coefficients. by adjusting the weight coefficients and consequently the smoothing characteristics of the filter according to the local statistics around each point of the image, it is possible to suppress noise while edges and other important features are preserved. application of the filter to several ultrasonic scans has shown that processing improves the detectability of small structures and subtle grey-scale variations without affecting the sharpness or anatomical information of the original image. comparison with the pure median filter demonstrate Segmentation through seeded region growing is widely used because it is fast, robust and free of tuning parameters. However, the seeded region growing algorithm requires an automatic seed generator, and has problems to label unconnected pixels (the unconnected pixel problem). This paper introduces a new automatic seeded region growing algorithm called ASRG-IB1 that performs the segmentation of color (RGB) and multispectral images. The seeds are automatically generated via histogram analysis; the histogram of each band is analyzed to obtain intervals of representative pixel values. An image pixel is considered a seed if itsgray values for each band fall in some representative interval. After that, our new seeded region growing algorithm is applied to segment the image. This algorithm uses instance-based learning as distance criteria. Finally, according to the user needs, the regions are merged using ownership tables. The algorithm was tested on several leukemia medical images showing good results[3].

Accurate diagnosis of cancer plays an importance role in order to save human life. The results of the diagnosis indicate by the medical experts are mostly differentiated based on the experience of different medical experts. This problem could risk the life of the cancer patients. From the literature, it has been found that Artificial Intelligence (AI) machine learning classifiers such as an Artificial Neural Network (ANN) and Support Vector Machine (SVM) can help doctors in diagnosing cancer more precisely. Both of them have been proven to produce good performance of cancer classification accuracy. The aim of this study is to compare the performance of the ANN and SVM classifiers on four different cancer datasets. For breast cancer and liver cancer dataset, the features of the data are based on the condition of the organs which is also called as standard data while for prostate cancer and ovarian cancer; both of these datasets are in the form of gene expression data. The datasets including benign and malignant tumors is specified to classify with proposed methods. The performance of both classifiers is evaluated using four different measuring tools which are accuracy, sensitivity, specificity and Area under Curve (AUC). This research has shown that the SVM classifier can obtain good performance in classifying cancer data compare to ANN classifier[5].

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Ultrasound image
- Preprocessing
- Segmentation using k-means algorithm
- Feature extraction
- Classification using ANN

B. *Description of the Proposed Algorithm:*

*K-means clustering algorithm***:**

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

where,

'$\|x_i$ - $v_j\|$' is the Euclidean distance between xi and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,v_c} be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, ‘c_i’ represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

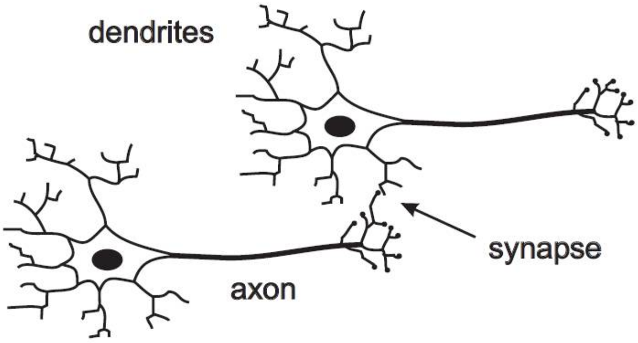
Neural networks

The ANNs consist of many connected neurons simulating a brain at work. A basic feature which distinguishes an ANN from an algorithmic program is the ability to gen- eralize the knowledge of new data which was not presented during the learning process. Expert systems need to gather actual knowledge of its designated area. However, ANNs only need one training and show tolerance for discontinuity, accidental disturbances or even defects in the training data set. This allows for usage of ANNs in solving problems which cannot be solved by other means effectively. These features and advantages are the reason why the area of ANN’s application is very wide and includes for example:

- Pattern recognition,
- Object classification,
- Medical diagnosis,
- Forecast of economical risk, market prices changes, need for electrical power, etc.,
- Selection of employees,
- Approximation of function value.

Biological neural networks:

The human brain consists of around 1011 nerve cells called neurons. The nucleus can be treated as the computational centre of a neuron. Here the main processes take place. The output duct of a neuron is called axon whereas dendrite is its input. One neuron can have many dendrites but only one axon; biological neurons have thousands of dendrites. Connections between neurons are called synapses; their quantity in a human brain is greater than 1014. A neuron receives electrical impulses through its dendrites and sends them to the next neurons using axon. An axon is split into many branches ending with synapses. Synapses change power of received signal before the next neuron will receive it. Changing the strengths of synapse effects is assumed to be a crucial part of learning process and that property is exploited in models of a human brain in its artificial equivalent.



Artificial Neural Network (ANN)

ANN is a parallel distributed processor that has a natural tendency for storing experiential knowledge. They can provide suitable solutions for problems, which are generally characterized by non-linear ties, high dimensionality noisy, complex, imprecise, and imperfect or error prone sensor data, and lack of a clearly stated mathematical solution or algorithm. A key benefit of neural networks is that a model of the system can be built from the available data. Image classification using neural networks is done by texture feature extraction and then applying the back propagation algorithm.The first layer is referred as input layer and the second layer is represents the hidden layer, has a tan sigmoid (tan-sig) activation function is represented by

$$\mathbf{Y(t)=tanh()}$$

DOI: 10.15680/IJIRCCE.2016. 0402150

This function is a hyperbolic tangent which ranges from -1 to 1, *yi*is the output of the *i*th node (neuron) and *vi* is the weighted sum of the input and the second layer or output layer, has a linear activation function. Thus, the first layer limits the output to a narrow range, from which the linear layer can produce all values. The output of each layer can be represented by

$$Y_{Nx1} = f(W_{NxM} X_{M,1} + b_{N,1})$$

where Y is a vector containing the output from each of the N neurons in each given layer, W is a matrix containing the weights for each of the M inputs for all N neurons, **X** is a vector containing the inputs, b is a vector containing the biases and f(·) is the activation function for both hidden layer and output layer.

The trained network was created using the neural network toolbox from Matlab9b.0 release. In a back propagation network, there are two steps during training. The back propagation step calculates the error in the gradient descent and propagates it backwards to each neuron in the hidden layer. In the second step, depending upon the values of activation function from hidden layer, the weights and biases are then recomputed, and the output from the activated neurons is then propagated forward from the hidden layer to the output layer. The network is initialized with random weights and biases, and was then trained using the Levenberq- Marquardt algorithm (LM). The weights and biases are updated according to

$$Dn+1 = Dn - [J^T J + \mu I]^{-1} J^T e$$

where*Dn*is a matrix containing the current weights and biases, *Dn+1* is a matrix containing the new weights and biases, e is the network error, *J* is a Jacobian matrix containing the first derivative of e with respect to the current weights and biases. In the neural network case, it is a K-by-L matrix, where K is the number of entries in our training set and L is the total number of parameters (weights+biases) of our network. It can be created by taking the partial derivatives of each in respect to each weight, and has the form:

$$J = \begin{bmatrix} \frac{\partial F(x_1, w)}{\partial w_1} \cdots \frac{\partial F(x_1, w)}{\partial w_w} \\ \frac{\partial F(x_2, w)}{\partial w_1} \cdots \frac{\partial F(x_2, w)}{\partial w_w} \end{bmatrix}$$

where F(xi,L) is the network function evaluated for the i-th input vector of the training set using the weight vector L and wj is the j-th element of the weight vector L of the network. In traditional Levenberg-Marquardt implementations, the jacobian is approximated by using finite differences, Howerever, for neural networks, it can be computed very effiecientlyby using the chain rule of calculus and the first derivatives of the activation functions. For the least-squares problem, the Hessian generally doesn't needs to be caclualted. As stated earlier, it can be approximated by using the Jacobian matrix with the formula:

$$H = J^T J$$

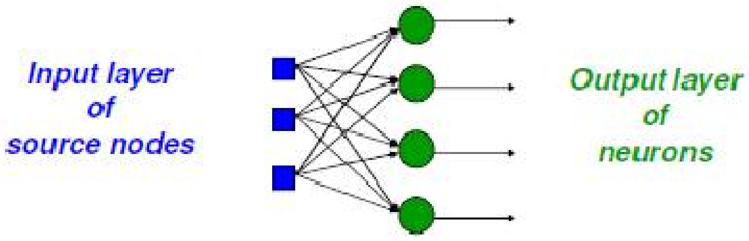
I is the identity matrix and μ is a variable that increases or decreases based on the performance function. The gradient of the error surface, g, is equal to *JTe*

Training of the Feed Forward Neural Network

Feed forward neural network is trained using back propagation algorithm. There are two types of training or learning modes in back propagation algorithm namely sequential mode and batch mode respectively. In sequential learning, a given input pattern is propagated forward and error is determined and back propagated, and the weights are updated.

Whereas, in Batch mode learning; weights are updated only after the entire set of training network has been presented to the network. Thus the weight update is only performed after every epoch. It is advantageous to accumulate the weight correction terms for several patterns. Here batch mode learning is used for training. In addition, neural network recognizes certain pattern of data only and also it entails difficulties to learn logically to identify the error data from the given input image. In order to improve the learning and understanding properties of neural network, noisy image data and filtered output image data are introduced for training. Noisy image data and filtered output data are considered as inputs for neural network training and noise free image is considered as a target image for training of the neural network. Back propagation is pertained as network training principle and the parameters of this network are then iteratively tuned. Once the training of the neural network is completed.

Two different classes of network architectures_
Single-layer feed-forward



IV. System Organization

Segmentation

Segmentation is a technique that subdivides a digital image into multiple segments. Segmentation is based on one of two basic properties of intensity: similarity and discontinuity. Detecting Similarities means to partition an image into regions that are similar according to a set of predefined criterion this includes image segmentation algorithms like thresholding, region growing, region splitting and merging. Detecting Discontinuities means to partition an image based on abrupt changes in intensity, this includes image segmentation algorithms like edge detection. Different Approaches for Medical Image Segmentation:

- Level set
- Edge Detection
- Threshold Based
- Clustering Based
- Region Based

Liver tumor segmentation is done with region based approach in this thesis work. Region based methods partition an image into regions that are similar according to a set of predefined criteria where as other segmentation approaches like edge detection methods partition an image according to rapid changes in intensity near the edges. Region based segmentation approach provides good results on contrast enhanced images and is immune to noise. Region growing methods can correctly separate the regions that have the same properties.

V. Results

The detection of cancer and its segmentation is considered as the second stage in which different image segmentation techniques are utilized to partition and extract objects from the enhanced image. Active Contour Model is used for the purpose of Region of Interest segmentation. The third stage involves disease classification of segmented

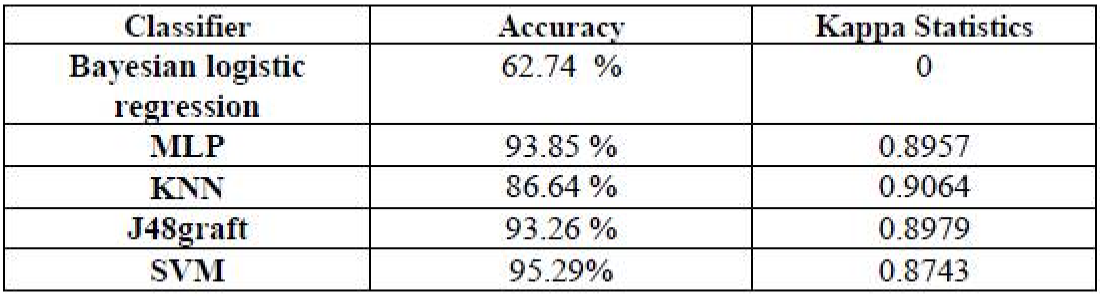
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

objects, in which the meanings of an investigated object are matched with the disease dictionary defined by physicians and radiologists. At this stage LESH features were obtained of normalized ROI.



Classifier	Accuracy	Kappa Statistics
Bayesian logistic regression	62.74 %	0
MLP	93.85 %	0.8957
KNN	86.64 %	0.9064
J48graft	93.26 %	0.8979
SVM	95.29%	0.8743

Classifiers performance is measured by introducing WEKA Explorer where several classifiers such as Bayesian Logistic regression, Multi-Layer Perception, KNN, J48graft and SVM classifier were tested on LESH features. SVM produced 95.29% accuracy results and performed better among the machine learning algorithms tested. In future work, disease prediction using US and Magnetic Resonance Imaging (MRI) Fusion along with cost-sensitive learning is hoped to further improve the effectiveness and value of the study.

VI. CONCLUSION AND FUTURE WORK

An efficient CDS framework inspired by the human interpretation of US images is presented. The proposed framework is composed of a number of stages where images are first acquired from an imaging source and pre-processed before running through an image enhancement algorithm. 2D Median Filter and CLAHE are employed for Image Normalization and Image Enhancement respectively. The detection of cancer and its segmentation is considered as the second stage in which different image segmentation techniques are utilized to partition and extract objects from the enhanced image. Active Contour Model is used for the purpose of Region of Interest segmentation. The third stage involves disease classification of segmented objects, in which the meanings of an investigated object are matched with the disease dictionary defined by physicians and radiologists. At this stage LESH features were obtained of normalized ROI. Classifiers performance is measured by introducing WEKA Explorer where several classifiers such as Bayesian Logistic regression, Multi-Layer Perception, KNN, J48graft and SVM classifier were tested on LESH features. SVM produced 95.29% accuracy results and performed better among the machine learning algorithms tested. In future work, disease prediction using US and Magnetic Resonance Imaging (MRI) Fusion along with cost-sensitive learning is hoped to further improve the effectiveness and value of the study.

REFERENCES

[1] S. Sahu, M. Dubey and M. I. Khan, “Liver Ultrasound Image Analysis using Enhancement Techniques,” International Journal of Advanced Computer Research, vol. 2, no. 6, pp. 2277-7970, 2012.

[2] U. Zakir, A. Hussain and L. Ali, “Improved Efficiency of Road Sign Detection and Recognition by Employing Kalman Filter,” 6th International Conference, BICS, vol. 7888, no. 1, pp. 216-224, 2013.

[3] W. Qiu, F. xiao, X. Yang, X. Zhang, M. Yuchi and M. Ding, “Research on Fuzzy Enhancement in the Diagnosis,” I.J. Image, Graphics and Signal Processing, vol. 3, pp. 10-16, 2011.

[4] M. M., M. A. Rajabi and J. ..Blais, “EFFECTS AND PERFORMANCE OF SPECKLE NOISE REDUCTION FILTERS ON,” [Online]. Available: http://www.isprs.org/proceedings/XXXVI/1-W41/makaleler/Rajabi_Specle_Noise.pdf.

[5] T. Loupas, W. ..McDicken and P. L. Allan, “An adaptive weighted median filter for speckle suppression in medical ultrasonic images,” Circuits and Systems, IEEE, vol. 36, no. 1.

[6] M. H. Yap, E. A, E. and H. E. Bez, “A novel algorithm for initial lesion detection in ultrasound breast images,” Journal of Applied Clinical Medical Physics, vol. 9, no. 4, 2008

[7] S. Saini, B. Kasliwal and S. Bhatia, “Comparative Study Of Image Edge Detection Algorithms,” 21 Nov 2013. [Online]. Available: http://arxiv.org/abs/1311.4963. G. Octavio, J. A., G. E. F and M. , “Image Segmentation Using Automatic Seeded,” Springer, vol. 4756, no. 1, pp. 192-201, 2007.