Segmentation of Lung Tumor using Clustering Techniques

P. Thangaraju[1], N. Mala[2]
[1]Department of Computer Applications
[2]Department of Computer Science
[1, 2] Bishop Heber College (Autonomous), Tiruchirappalli, India

Abstract- *Lung cancer also known as carcinoma of the lung better known as pulmonary carcinoma in medical terms. In fact this is a malignant tumor of the lung characterized by uncontrolled cell growth, which is dangerous and abnormal. This takes place on the inner walls containing the tissues and cell parts forming lung. If left untreated, this growth can spread beyond the lung by process of metastasis in to nearby tissues or other organs of the body. Further it is identified that lung tumor causing cancer is the number one cause of cancer deaths in humans both men and women in the world. This paper proposes the study and classification of such lung cancer tumor using accurate image segmentation techniques. The proposed model compares lung tumor using three algorithms namely K Harmonic Means, Expectation Maximization and Hierarchical clustering using images. The various parameters like accuracy, time consumption, features extracted and iterations are studied and the best practices will be put into practice.*

Keywords- Lung Cancer, Image Classification, Tumor, EM, K-Harmonic, Hierarchical

I. INTRODUCTION

Cancer of the lung results from an abnormality in the cell which is the body's basic unit of life. Normally, every unit in the body maintains a system which checks the cell growth. The process is such that cells divide and produce different new cells, which are produced as and when needed. Any process damage to this system of checks causes cell growth retardation or abnormal tumor growth, which in turn results in uncontrolled cell multiplication forming a new mass called popularly as a tumor. Tumors are either good or bad. Bad tumor cells are called "cancer cells." Good tumors normally will not spread to the other parts of the body and may or may not be removed. These bad malignant tumors invade to other cells or other parts of the body. Next it enters into the bloodstream and the lymphatic system in turn causing degeneration. This is called lung cancer which tends to spread very early and very fast after its formation. Majority of the time it is a very life-threatening one and difficult to treat. While this type of lung cancer may spread to any cell in the body, only certain points are vulnerable. These vulnerable areas include the adrenal glands, the liver, the brain which are some of the most familiar places for lung cancer tumor.

The lung area has been identified as one of the most affected areas for tumors than other parts of the human body[1]. A lot of data has been collected and gone over. The tricky part is how to identify the lung cancer tumor into something meaningful for the patient or doctor or rather the persons involved in it or the persons who offer treatment. This is where data mining steps in and classification of the lung cancer data helps in such efforts.

The approach to find a suitable work related to the lung cancers segmentation. The proposed model aims and uses data image mining or segmentation features to find lung cancer tumors and remove them.

II. LITERATURE REVIEW

Lawrence A. Leob, et al.,[2] proved with confirmed evidence that tobacco smoking is the cause of 30 to 40% of deaths from lung cancer. Their focus is on lung cancer because of the sheer magnitude of this disease in males and the likelihood of a similar epidemic in females. Chemical analyses of cigarette smoke reveal a multitude of known mutagens and carcinogens. Moreover, these chemicals are absorbed, are metabolized and cause demonstrable genetic changes in smokers.

The social and economic costs of lung cancer and the smoking habit impinge on the productiveness of our society.

Higgins, I. T. T[3] presented to the evidence on balance is overwhelmingly against smoking as the most important cause of lung cancer, not to mention its important role in the genesis of obstructive airways disease. Excess mucus in the lungs is a condition to be rigorously avoided by any possible means.

Ahmed, Kawsar, et al.,[4] developed a significant pattern prediction tools for a lung cancer prediction system were developed. The lung cancer risk prediction system should prove helpful in detection of a person’s predisposition

for lung cancer. The early prediction of lung cancer should play a pivotal role in the diagnosis process and for ineffective preventive strategy.

Krishnaiah, V, et al.,[5] presented a model for nearly detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease.

Deoskar, Parag, et al.,[6] proposed to assorted data mining and ant colony optimization techniques for appropriate data sets for tumors in general.

Singh, Chinnappan Ravinder, et al.,[7] reviewed scientific evidence, particularly epidemiologic evidence of overall lung cancer burden in the world and molecular understanding of lung cancer at various levels by dominant and suppressor oncogenes.

Karthikeyan, T., and P. Thangaraju [8] apply classification techniques on a dataset of lung cancer patients based on smoking and non-smoking people.ACO employs artificial ants that cooperate to find good solutions for discrete optimization problems[8]. Proposed to This paper mainly deals with feature extraction algorithm used to improve the predicted accuracy of the classification. ACO employs artificial ants that cooperate to find good solutions for discrete optimization problems [9]. These software agents mimic the foraging behavior of their biological counterparts in finding the shortest-path to the food source. The first algorithm following the principles of the ACO met heuristic is the Ant System [10], [11]. where ants iteratively construct solutions and add pheromone to the paths corresponding to these solutions.

III. PROBLEM IDENTIFICATION

Lung cancer tissue and tumor segmentation in images have been an active research area. The extraction of essential features from the image is very important for successful tumor image segmentation. As a result of the complex extraction of different tumor cells such as the in the MR lung cancer images, the extraction of useful features is a challenging task.

The tumor location, shape and texture properties further complicates the search for robust features. Posteriorfossa(PF) tumor is usually located near the lung cancer stem and cerebellum. About 70% of the lung tumors arise in the narrow arteries. This narrow confinement at the base of the lungs and its complete removal poses some nontrivial challenges [12]. Therefore only accurate segmentation of the lung tumor is a must.

Detection of anatomical lung cancer structures with their exact location is important for treatments like radiation therapy and surgery. Radiologists perform the diagnosis of lung cancer tumour manually on MRI images but it being time consuming and error prone as large no of image slices and the large variations between them.

The brightness of the images, color also pose a problem during segmentation process. Any inherent noise should be removed using appropriate filters. Excess removal results in the image getting degraded. Also normally segmentation algorithms pose huge computational overheads resulting in huge processing time and accuracy loss. The above problems should be addressed.

IV. METHODOLOGY

The data for the lung tumors have been taken from the MRI image dataset of the Adayar Cancer Institute.

The study proposes the image segmentation classification model for studying lung cancer tumors and provides early incentives for easy detection of the lung cancer[13]. A tumour is an acronym for a neoplasm or a solid lesion formed by an abnormal growth of cells (termed neoplastic) which looks like a swelling.

MRI images to detect lung cancer tumour classifies the tumour depending on whether the lung cancer is an abnormal tissue containing normal volume lung cancer tissues like white matter, gray matter and CSF (cerebro-spinal fluid) but also have some slices contain pathology like edema and necrosis hence making them abnormal lung cancer tissues.

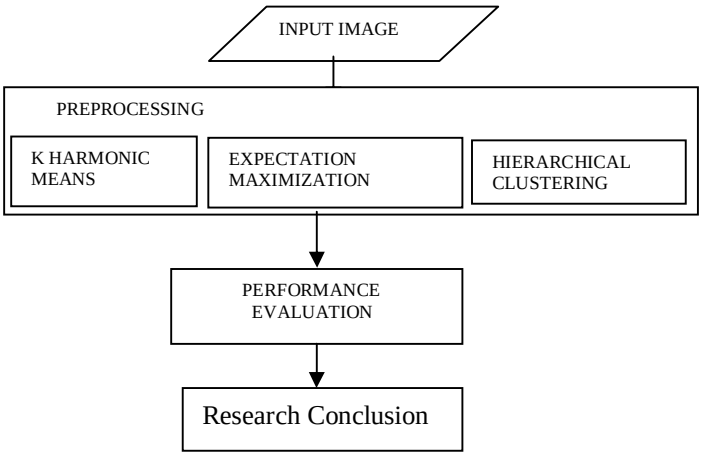
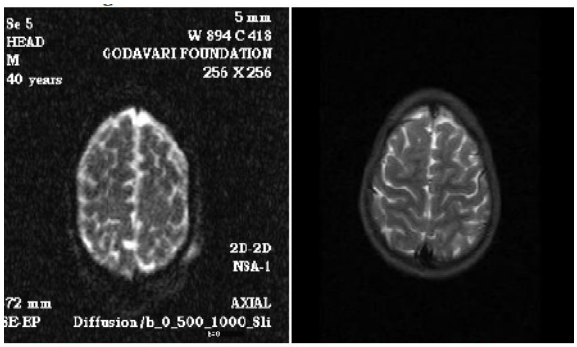


Figure 1: Tumour Detection Process

Based on the CSF Symmetry on the vertical axis through the lung cancer center a normal volume lung cancer tissue and an abnormal volume lung cancer tissue could be classified[14][15]. The MRI images can be of T1, T2 weighted type of which T2 weighted Images are being widely used in Medical Imaging as in this case of cerebral and spinal study, the CSF (cerebrospinal fluid) are lighter in T2 weighted images as they are acquired using fast echo spin sequence whereas the T1 weighted images are acquired using a spin echo sequence. Primary focus on exact lung cancer tumours location and its extraction with parameters like area and time to yield faithful and error free output.



Original Image Pre-Processed Image

V. CLUSTERING METHODS

The K-means algorithm is an iterative technique that is used to partition an image into K clusters. The basic algorithm is:

Step1: Pick K cluster centers, either randomly or based on some heuristic

Step2: Assign each pixel in the image to the cluster that minimizes the variance between the pixel and the cluster center

Step3: Re-compute the cluster centers by averaging all of the pixels in the cluster

Step4: Repeat steps 2 and 3 until convergence is attained (e.g. no pixels change clusters)

In this case, variance is the squared or absolute difference between a pixel and a cluster center. The difference is typically based on pixel color, intensity, texture and location or a weighted combination of these factors. K can be selected manually, randomly or by a heuristic.

VI. K HARMONIC MEANS

The computation starts with an initialization of the center positions and followed by iterative refinement of these positions. Many experimental results show that KHM is essentially insensitive to the initialization of the centers than KM and EM. The dependency of the K-Means performance on the initialization of the centers is a major problem; a similar issue exists for an alternative algorithm, Expectation Maximization (EM), although to a lesser extent. Many papers have been published to find good initializations for KM[16][19]. This paper takes a totally different approach by changing MIN() used in KM to HA() (Harmonic Average), which is similar to MIN() but "softer", to make the performance function "easier to optimize" by an algorithm that is essentially insensitive to initialization.

Algorithm:

Step1: Initialize the algorithm with guessed centers C

Step2: For each data point x i , compute its membership m(cj|xi) in each center c j and its weight w (xi).

Step3: For each center c j, recomputed its location from all data points x i, according to their membership and weights

Step4: The solution is then a set of k cluster centers, which is located at the centroid of the data for which it is the closest center

Step5: KM has a hard membership function, so each data point belongs only to its nearest cluster. This way a Voronoi partitioning of the data is created. The membership function is defined

Step6: KM has a constant weight function, that gives all data points equal importance in each iteration. The weight function is defined

VII. EXPECTATION MAXIMIZATION CLUSTERING

The EM algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates. The model depends on unobserved latent variables. The EM iteration always alternates between performing an expectation (E) step. This function creates a function for the expectation of the log-likelihood of the current estimate for the parameters. Second is a maximization (M) step, which computes the parameters by maximizing expected log-likelihood found on the E step[17]. Such parameter-estimates are subsequently used to determine the distribution of the latent variables.

Algorithm:

This is an iterative algorithm, in the case where both θ and Z are unknown:

Step1:First, initialize the parameters θ to some random values.

Step2: Compute the best value for Z given these parameter values.

Step3: Then, use the just-computed values of Z to compute a better estimate for the parameters θ. Parameters associated with a particular value of Z will use only those data points whose associated latent variable has that value.

Step4: Iterate steps 2 and 3 until convergence.

The algorithm as just described monotonically approaches a local minimum of the cost function and is commonly called hard EM. The k-means algorithm is an example of this class of algorithms.

VIII. HIERARCHICAL CLUSTERING

The hierarchical clustering is a novel method for cluster analysis. This model constructs a hierarchy of similar feature clusters. They are broadly split into two types:

Agglomerative Clusters: This is popularly known as "bottom up" approach with each observation starts its own feature cluster. The other similar pairs of clusters are merged into the parent and forms the top of the hierarchy.

Divisive Hierarchy Cluster: This is opposite of bottom up and is a "top down" cluster[20][21]. Here all the observations start from one cluster at the top. Next it is split into groups called cluster. This action is performed recursively by moving down the line in the hierarchy. The merges and splits are determined using a greedy approach. The results of hierarchical clustering are in dendogram format.

The algorithm criteria includes:

- The sum of all the intra (inside) cluster variance.
- The decrease results for the cluster being merged.
- The probability of candidate clusters from the same distribution function is called Vlinkage.
- The product called in-degree and out-degree is based on a k-nearest-neighbor model.

Each increment of cluster is by a defined quantity defined measured by the quality of the cluster formed after merging clusters.

IX. RESULTS AND DISCUSSION

The lung cancer tumor images have been segmented using the three algorithmsnamely KHM, EM and HC. The findings are as follows. All the three are good at identifying the tumor but when based on certain finer parameters the results vary slightly, which are discussed below. The first parameter is accuracy and hierarchical clustering outperforms the other methods as shown.

Table 1: showing the performance parameters of the three algorithms.

FACTORS	KHM	EM	HC
ACCURACY	90.2%	85.3%	98.2%
ITERATIONS	21	15	42
TIME TAKEN	9	6	15

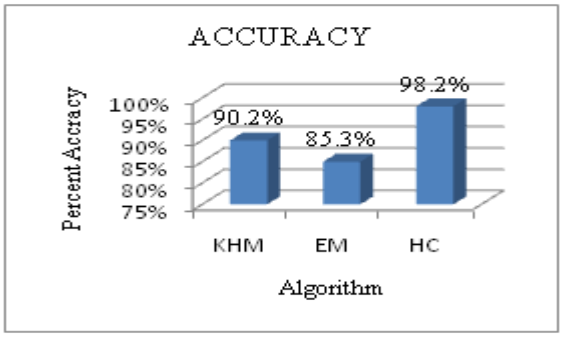


Figure2: Accuracy value

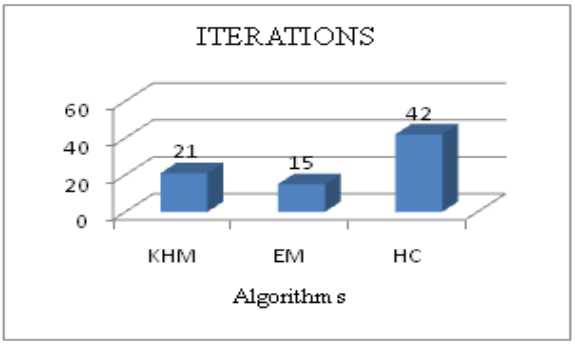


Figure 3: The next is about the convergence of the iterations and EM takes less iterations, while KHM and HC are more.

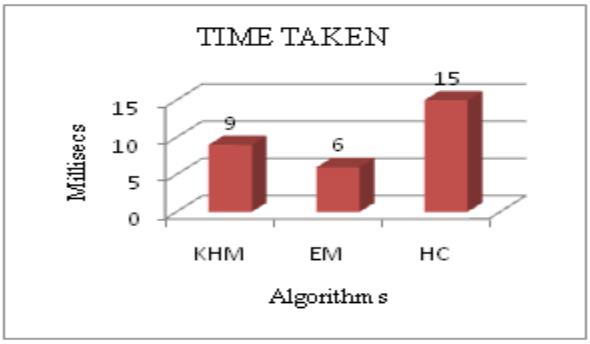


Figure 4: Therefore it may be concluded that EM takes lesser time due to the lesser number of iterations

X. CONCLUSION

Thus the study successfully introduces the cluster based segmentation of medical lung cancer images to identify tumors. Parameters like accuracy, iterations and time taken have been studied for all the lung cancer image sets. It has been successfully concluded that Hierarchical Clustering is most accurate for identifying lung cancer tumors, but EM is the most cost effective one and is also computationally inexpensive. As part of the future enhancements the lung cancer model may be implemented as web services.

REFERENCES

[1] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.

[2] Lawrence A. Loeb, Virginia L. Ernster, Kenneth E. Warner, John Abbotts and John Laszlo “Smoking and Lung Cancer”, on July 17, 2014.

[3] Higgins, I. T. T. "Commentary on “possible effects on occupational lung cancer from smoking related changes in the mucus content of the lung”, Journal of Chronic Diseases, 1983, Vol.36, no. 10 pp. 677-680.

[4] Ahmed, Kawsar, Abdullah-Al-Emran Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti, Md Rahman, and Farzana Ahmed. "Early detection of lung cancer risk using data mining", Asian Pacific Journal of Cancer Prevention, 2013, Vol. 14, no. 1, pp. 595-598.

[5] Krishnaiah, V., Dr G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques", International Journal of Computer Science and Information Technologies, 2013, Vol. 4, no. 1, pp. 39-45.

[6] Deoskar, Parag, Dr Divakar Singh, and Dr Anju Singh. "Mining Lung Cancer Data and Other Diseases Data Using Data Mining Techniques: A Survey", International Journal of Computer Engineering and Technology (IJCET), 2013, Vol. 4, no. 2.

[7] Singh, Chinnappan Ravinder, and Kandasamy Kathiresan. "Molecular understanding of lung cancers–A review", Asian Pacific journal of tropical biomedicine , 2014, Vol.4.

[8] Karthikeyan, T., and P. Thangaraju. "PCA-NB algorithm to enhance the predictive accuracy", Int. J. Eng. Tech, 2014, Vol. 6, no. 1, pp. 381-387.

[9] Karthikeyan, T., and P. Thangaraju. "Analysis of classification algorithms applied to hepatitis patients". International Journal of Computer Applications, 2013, Vol.62, no. 5

[10] Karabatak, Murat, and M. Cevdet Ince. "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, 2009, Vol. 36, no.2. pp 3465-3469.

[11] Cancer Research in ICMR Achievements in Nineties, ICMR Report 2006.

[12] Zaïane, Osmar R. "Principles of knowledge discovery in databases", Department of Computing Science, University of Alberta , 1999.

[13] Krishnaiah, V., Dr G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques", International Journal of Computer Science and Information Technologies, 2013, vol. 4, no. 1,pp. 39-45.

[14] Kaur, Harleen, and Siri Krishan Wasan. "Empirical study on applications of data mining techniques in healthcare", Journal of Computer Science, 2006, Vol. 2, no. 2, pp. 194-200.

[15] Kharya, Shweta. "Using data mining techniques for diagnosis and prognosis of cancer disease", arXiv preprint arXiv, 2012.

[16] Quinlan, J. Ross. "C4. 5: Programming for machine learning", Morgan Kauffmann , 1993.

[17] L. Breiman.” Random forests. Machine learning”, 2001,Vol. 45, no.1, pp.5–32.

[18] Breiman, Leo. "Random forests", Machine learning, 2001, Vol. 45, no. 1, pp. 5-32.

[19] R. D´ ıaz-Uriarte and A. de Andres. Gene selection and classification of microarray data using random forest. BMC bioinformatics, 2006, Vol.7, no.1.

[20] Michalski, Ryszard S., and Kenneth A. Kaufman."Learning patterns in noisy data: the AQ

approach". In Machine Learning and its Applications, Springer Berlin Heidelberg, 2001, pp. 22-38.

[21] Linder, Roland, Tereza Richards, and Mathias Wagner. "Microarray data classified by artificial neural networks". In Microarrays, Humana Press, 2007, pp. 345-372.