Early Detection of Lung Cancer using Data Mining Techniques: A Survey

G.VIJAYA, Research Scholar(Ph.D),

Annamalai University,

Annamalai Nagar.

DR.A.SUHASINI, Associate Prof.,

Annamalai University,

Annamalai Nagar.

Abstract:

Lung cancer is the leading cause of cancer death in the World States for both men & women. The early detection of lung cancer can be helpful in curing the disease completely. In general, a measure for early stage lung cancer diagnosis mainly includes X-ray chest films, CT scans, MRI scans, Biopsy etc. The present article surveys the role of different Data mining paradigms, like Decision Trees (DT), Artificial Neural Networks (ANN), Association Rule Mining (ARM) and Bayesian Classifier. The pros and cons of each Data Mining techniques were indicated and an extensive bibliography is also included.

Keywords: Decision Trees, ANN, ARM, Bayesian Classifier.

1. INTRODUCTION

Lung cancer remains a major cause of morbidity and mortality worldwide, accounting for more deaths than any other cancer cause. According to [1], new Lung cancer cases for 2012 among males 116,470 and among females 109,690. Deaths due to Lung cancer among males 87,750 and in females 72,590. In World, tobacco use is responsible for nearly 1 in 5 deaths.

1.1 Lung Cancer

Lung cancer is a disease of abnormal cells growing multiplying and growing into tumour. Cancer cells can be carried away from the lungs in
blood, or lymph fluid that surrounds lung tissues. Metastasis occurs when a cancer leaves the site where it began and moves into a lymph node or to another part of the body through the blood stream.

1.2 Lung Cancer Types

There are 3 types of Lung Cancer. They are:

a)Non – Small Cell Lung Cancer (NSCLC):

According to Cancer Society, about 85% to 95% of Lung Cancers are this type. This can be sub divided into:

i) ***Squamous cell carcinoma*** – about 25% to 30% of all lung cancers are squamous cell carcinomas. They are often linked to a history of smoking & tend to be found in the middle of the lungs, near a bronchus.

ii) ***Adenocarcinoma*** – about 40% of lung cancers are adenocarcinoma type of cancer. This is the most common type of lung cancer seen in non-smokers. It is more common in women than in men and it is more likely to occur in younger people than other types of lung cancer. It is usually found in the outer region of the lung. People with this type of lung cancer tend to have a better outlook (prognosis) than those with other types of lung cancer.

iii) ***Large cell (undifferentiated) carcinoma*** – this will carry about 10% to 15% of NSCLC cancers. It may appear in any part of the lung. It tends to grow and spread quickly, which can make it harder to treat.

b) Small Cell Lung Cancer (SCLC):

This type of lung cancer can occupy up to 10% to 15% of all lung cancers. It is very rare for someone who has never smoked. SCLC often starts in the bronchi near the centre of the chest, and it tends to spread widely through the body.

c) Other types of Lung Cancer:

1. "Carcinoid tumours" of the lung account for fewer than 5% of lung tumours. Most of them are slow growing & generally cured by surgery.
2. "Cancers that starts in other organs" (such as breast, pancreas, kidney or skin) can sometimes spread to the lungs. But these are not lung can**1.3. Detection of Lung Cancer:**

The prime method for cancer detection is through radiological imaging exams. There are many technologies used in the diagnosis of lung cancer like Chest X-ray, Lung CT scan, Lung PET scan etc.

The present article provides an overview of the available literature on the detection of lung cancer in the data mining framework. Section II describes the basic notions of Decision Trees and its application in Lung Cancer. This is followed by Section III by a survey explaining the importance of ANN in the detection of Lung cancer. The utility and applicability of Fuzzy and Rule Mining techniques are briefly discussed in Section IV & Section V. Section VI concludes the article. Some challenges to data mining and the application of Lung cancer are also indicated.

2. DECISION TREE

A *Decision Tree* [2] is a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each *branch* represents an outcome of the test, and each *leaf node* (terminal node) holds a class label. The topmost node in a tree is the *root* node.

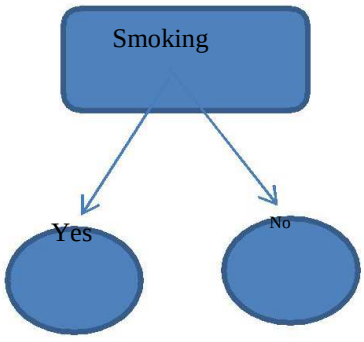

Fig 1. Attribute for Lung Cancer

Fig.1 represents the patients having Lung Cancer have the Smoking habits or not. The rounded rectangle represents the tuples (attributes) and the oval represents the leaf node (output).

2.1. C4.5 Decision Tree

C4.5 is a well-known decision tree induction learning technique. Decision tree J48 implements Quinlans C4.5 algorithm [3], for generating a pruned or unpruned C4.5 tree. The decision tree generated by J48 can be used for classification. J48 builds decision trees from a set of labelled training data using the concept of information entropy. It uses the concept that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

J48 can handle both continuous and discrete attributes, training data with missing attribute values, and attributes with differing costs. Further it provides option for pruning trees after creation.

2.2. CART

CART (Classification And Regression Tree) is a recursive and gradual refinement algorithm of building a decision tree, to predict the classification situation of new samples of known input variable value. Breiman et al 1984 provided this algorithm.

Algorithm:

Step 1: All the rows in a data set are passing onto the root node.

Step 2: Based on the values for the rows in the node considered, each of the predictor variables are splited at all its possible split points.

Step 3: At each split points, the parent node is split as binary nodes, by separating the rows with values lower than or equal to the split points and values higher than the split points, for the considered predictor variable.

Step 4: The predictor variable and split point with the highest value is selected for the node.

Step 5: Binary split of the parent node into two child node is performed based on the selected split point.

Step 6: Repeat Steps (2) to (5), using each node as a new parent node, until the tree has the maximum size.

Step 7: The regression tree is pruned to select the optimal

size tree.

2.3. Issues in Decision Tree

- Decision Trees can suffer from *Repetition* and *Replication*.
- *Repetition* occurs when an attribute is repeatedly tested along a given branch of the tree.
- In *Replication*, duplicate subtrees exist within the tree.
- These situations can impede the accuracy and comprehensibility of the decision tree.

How to Overcome?

- The use of multivariate splits (splits based on a combination of attributes), can prevent this.
- Use a different form of knowledge representation, such as rules, instead of Decision Trees.

3. ARTIFICIAL NEURAL NETWORK (ANN)

ANNs are networks of interconnected artificial neurons, and are commonly used for non-linear statistical data modelling to model complex relationships between inputs and outputs. The network includes a hidden layer of multiple artificial neurons connected to the inputs and outputs with different edge weights. The internal edge weights are „learnt during the training process using techniques like back propagation. Several good descriptions of neural networks are available in [7] & [8].

3.1. A Multilayer Feed – Forward Neural Network (MFFNN)

A Multilayer Feed – Forward Neural Network consists of an *input layer*, one or more *hidden layers*, and an *output layer*. An example of MFFNN is shown in Fig. 2.

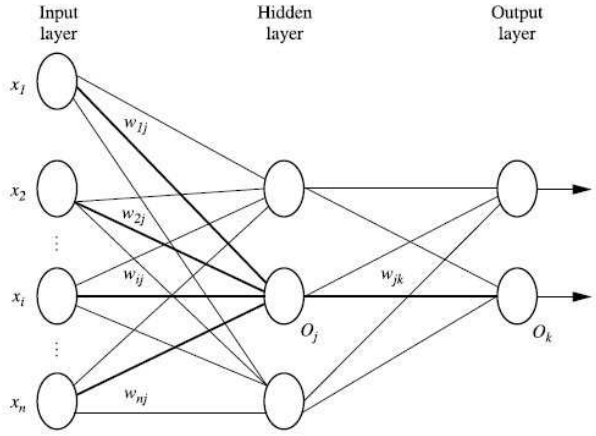

Fig 2. A Multilayer Feed – Forward Neural Network.

Each layer is made up of units. The inputs are fed simultaneously into the units making up the *input layer*. These inputs passing through the input layer and are then weight and fed simultaneously to a second layer of neuron-like units, known as *hidden layer*. The outputs of the hidden layer units can be input to another hidden layer, and so on. The weighted outputs of the last hidden layer are referred to as *output layer*.

The network diagram shown above is a full-connected, three layers, feed forward, perceptron neural network. – Fully connected means that the output from each input and hidden neuron is distributed to all of the neurons in the following layer. – Feed forward means that the values move from input to hidden layer to output layers; no values are fed back to earlier layers [9].

3.2. Backpropagation Neural Network (BPNN)

Backpropagation learns by iteratively processing a data set of training tuples, comparing the network s prediction for each tuple with the actual known *target* value. For each training tuple, the weights are modified so as to minimize the mean squared error between the network s prediction and the actual target value. These modifications are made in the "backward" direction, that is, from the output layer, through each hidden layer down to first hidden layer (hence the name *backpropagation*).

3.3. Support Vector Machines (SVM)

SVM introduced by Cortes is generally used for classification purpose. A detailed description of SVMs and SRM is available in [11]. In their basic form, SVMs attempt to perform classification by constructing hyperplanes in a multidimensional space that separates the cases of different class labels. It supports both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models like, linear, polynomial, radial basis function (RBF), and sigmoid. Of these, the RBF kernel is the most recommended and popularly used, since it has finite response across the entire range of the real x-axis.

SVM is particularly striking the biological analysis due its capability to handle noise, large dataset and large input spaces. The fundamental idea of SVM can be described as follows:

i) Initially, the inputs are formulated as feature vectors.

ii) Then, by using the kernel function, these feature vectors are mapped into feature space.

iii) Finally, a division is computed in the feature space to separate the classes of training vectors.

3.4. Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) derived from Single Hidden Layer Feed-Forward Neural Networks (SHLFNs) will randomly select the input weights and analytically determines the output weights of SLFNs. This algorithm tends to afford the best generalization performance at extremely fast learning speed.

The structure of the ELM network is shown in Fig.3 [14]. ELM contains an input layer, hidden layer, and an output layer. The ELM has several interesting and salient features from traditional popular learning techniques for feed forward neural networks. These include the following:

- The learning speed of ELM is very quick when compared to other classifier.
- ELM has enhanced generalization result when compared to the gradient-based learning technique.
- ELM will attain the results directly without any difficulties.

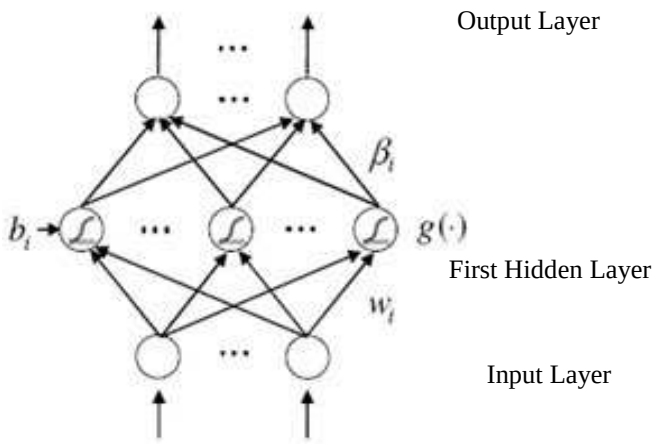

Fig. 3. Structure of Extreme Learning Machine (ELM)

3.5. Pros and Cons of Neural Networks:

Advantages of Neural Networks include:

- High tolerance of noisy data
- Ability to classify patterns on which they have not been trained.
- They are well suited for continuous valued inputs & outputs.

Challenges in Neural Networks:

- May not be applicable in some circumstances when the output is not clearly known.
- ANNs are more complex and robust when compared to other regressions.

How to overcome?

- It is planned to use reinforcement training model to overcome this problem.
- Reinforcement learning attempts to learn the input-output mapping through trial and error with a view to maximize the performance index.

4. ASSOCIATION RULE MINING

Association Rule Mining (ARM) [15], is often stated as follows: Let I be a set of n binary attributes called items. Let T be a set of transactions. Each transaction in T contains a subset of the items in I. A rule is defined as the implication of the form

$X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The set of items X and Y are called antecedent and consequent of the rule respectively. A commonly given example from market basket analysis rule is $\{Pencil\} \Rightarrow \{Eraser\}$, meaning that customer who buys pencil also buy eraser.

Association rule mining is popularly done with flag attributes, indicating the presence/absence of the item in the transaction. However, even from nominal attributes (having multiple but finite possible values), and numeric attributes, it is possible to derive flag attributes for the purpose of association rule mining.

Association rules are mined in two step process consisting of *frequent itemset mining* followed by *rule generation*:

Step 1: Searches for patters of attribute-value pairs that occur repeatedly in a data set, where each attribute value is considered an item. The resulting attribute value pairs form *frequent itemset.*

Step 2: Analyzes the frequent itemset in order to generate the association rule. All association rules must satisfy certain criteria regarding their

"accuracy" and the proportions of the data set that they actually represent.

4.1. Issues in ARM

- Associative classification offers a new alternative to classification scheme by building rules based on conjuctions of attribute-pair values that occur frequently in data.
- It shows much better efficiency with large sets of training data.

5. BAYESIAN CLASSIFIER

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.
The probability of a specific feature in the data appears as a member in the set of probabilities and is derived by calculating the frequency of each feature value within a class of training data set. The training data set is a subset, used to train a classifier algorithm by using known values to predict future or unknown values.

5.1. Naïve Bayesian Classifier

Naïve Bayesian classifiers [17], assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class conditional independence*. It is made to simply the computations involved and, is considered as,
"naïve".

The algorithm uses Bayesian theorem and assumes all attributes to be independent for the given values of the class variable. This conditional

– independence assumption rarely holds true in real-world applications, hence the characterization as Naïve yet the algorithm tends to perform well and learn rapidly in various supervised classification problems.

This "naivety" allows the algorithm to easily construct classifications out of large data sets without resorting to complicated iterative parameter estimation schemes.

Issues in Bayesian Classifier

- Despite its simplicity, the Naïve Bayesian Classifier is a robust method, which shows an average good performance in terms of classification accuracy, also when the independence assumption does not hold.

6. CONCLUSION

In this survey paper, different Lung Cancer prediction system is developed using the Data Mining classification techniques. The most effective model to predict patients with Lung Cancer Disease appears to be Naïve Bayes, followed by Association Rule Mining, Decision Trees and Neural Network. Decision Trees result are easy to read and interpret. Naïve Bayes are far better than Decision Trees as it could identify all the significant medical predictors. The relationship between attributes produced by Neural Network is more difficult to understand.

References:

[1] American Cancer Society. Cancer Facts & Figures 2012.

[2] J.R. Quinlan. Induction of decision trees. Machine learning,1(1):81–106, 1986.

[3] J.R. Quinlan. C4. 5: Programming for machine learning. MorganKauffmann, 1993.

[4] Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[5] V.Krishnaiah, Dr.G.Narasimha, Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol. 4(1), 2013, 39-45.

[6] B.Chandana, Dr. DSVGK Kaladhar, "Data Mining, Inference and Prediction of Cancer datasets using Learning Algorithms", International Journal of Science and Advanced Technology, Vol. 1, No. 3, May 2011, 68-77.

[7] C. Bishop, "Neural Networks for Pattern Recognition", Oxford: University Press, 1995.

[8] L. Fausett, "Fundamentals of Neural Networks", New York, Prentice Hall, 1994.

[9] Neha Sharma, "Comparing the Performance of Data Mining Techniques for Oral Cancer Prediction", ICCS□11, Feb 12-14, 2011, Rourkela, Odisha, India, 433-438.

[10] Penedo MG, Carreira MJ, MosquerraA, et al. "Computer – Aided Diagnosis: A Neural Network-based approach to Lung Nodule Detection", IEEE Trans Med Imag 1998; 17: 872-880.

[11] V.N. Vapnik. "The nature of statistical learning theory", Springer, 1995.

[12] Ankit Agrawal, Sanchit Misra et al., "A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data", BIOKDD 2011, Aug. 2011, San Diego, CA, USA.

[13] M.Gomathi, Dr.P.Thangaraj, "An Effective Classification of Benign and Malignant Nodules Using Support Vector Machine", Journal of Global Research in Computer Science, Vol. 3, No.7, July 2012, 6-9.

[14] M.Gomathi, Dr.P.Thangaraj, "A Computer Aided Diagnosis System for Lung Cancer Detection using Machine Learning Technique", European Journal of Scientific Research, Vol.51, No.2 (2011), 260-275.

[15] R.Agrawal, T.Imielinski, and A.Swami, "Mining association rules between sets of items in large databases", in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ser.SIGMOD□93, 1993.

[16] Ankit Agrawal and Alok Choudhary, "Identifying HotSpots in Lung Cancer Data Using Association Rule Mining", 11th IEEE International Conference on Data Mining Workshops , 2011, 995-1002.

[17] George Dimitoglou, James A.Adams, and Carol M.Jim, "Comparision of the C4.5 and a Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability", 2010, Frederick, USA.

[18] S.Vijayarani & S.Sudha, "Disease Prediction in Data Mining Technique – A Survey", International Journal of Computer Applications and Information Technology, Vol. II, Issue I, Jan 2013, 17-21.

[19] Muhammad Shahbaz, et al., "Cancer Diagnosis Using Data Mining Technology", Life Science Journal, 2012; 9(1), 308-313.