Lung Disease Prediction System Using Data Mining Techniques

S. Durga, Research Scholar, M.Phil (CS), Vels University, Chennai. E-mail:durgas20192@gmail.com

K. Kasturi, Assistant Professor, Dept of I.T, Vels University, Chennai. E-mail:kasturi.scs@velsuniv.ac.in

Abstract--- Data mining is defined as analyzing very large amount of data for getting some useful information. Data mining techniques like association rule mining, classification and clustering is implemented to analyze the different types of disease. Classification is an important problem in Data mining. Given a database contains a collection of records, each with a single class label, a classifier performs a brief and clear definition for each class that can be used to classify successive records. Data mining plays an important role in medical systems. It is used to discover the knowledge out of data and presenting it in the form that human can easily understand. It is a cooperative effort of humans and computers. There are two primary goals of data mining – Prediction and Description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. Description focuses on finding patterns describing the data that can be interpreted by humans. It is very useful for predicting diseases such as Heart disease, Lung disease. Lung cancer is one of the most dangerous diseases in the world. The early detection of lung cancer can cure the disease completely. Data mining plays an effective role by using Naïve Bayes and Artificial Neural Network to massive volume of healthcare of data. The health care industry collects huge amounts of data which unfortunately are not mined to find the hidden data. The Naïve Bayes aims at delivering robust classifications also when dealing with small or incomplete data sets. The aim of the paper is to detect and diagnose the lung diesases as early as possible which will help the doctor to save the patient's life. This paper describes how lung cancer was predicted and controlled, using data mining techniques.

Keywords--- Data Mining, Lung Cancer, Naïve Bayes, Classification.

I. Introduction

Lung cancer is also known as lung carcinoma is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung(1-2). If it is treated this growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. The majority factor of lung cancer are due to tobacco smoking. The other factors are the combination of genetic factors and exposure to radon gas, asbestos, second-hand smoke, or other forms of air pollution.

The two main types are:

- Small-cell lung carcinoma (SCLC)
- Non-Small-cell lung carcinoma (NSCLC)

The Symptoms of lung cancer are coughing, coughing up blood, wheezing, weakness, fever, bone pain etc. Many of the symptoms of cancer such as poor appetite, weight loss are not specific. In many people, the cancer has already spread beyond the original site by the time they have symptoms and seek medical attention. The lung cancer spreads on brain, bone, kidneys etc. About 10% people with lung cancer do not have symptoms at diagnosis. These cancers are found on routine chest radiography.

Treatment and long term outcomes depend on the type of cancer, stage, and person's health. The common treatments are surgery, chemotherapy, radiotherapy.

Smoking prevention and smoking cessation are effective ways of preventing the development of lung cancer.

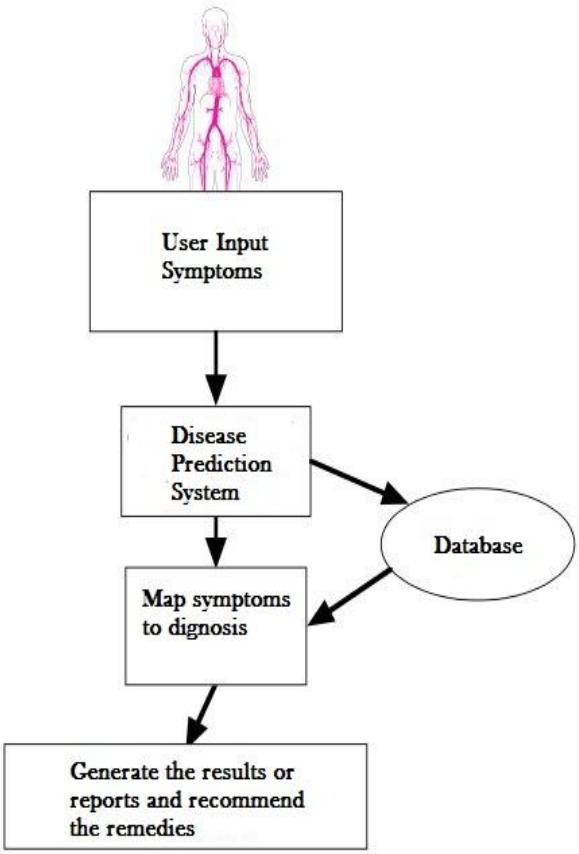
II. Diagnosis

A chest radiograph is one of the steps if a person reports the symptoms for lung cancer. This may reveal on widening of the media stinum, atelectasis, consolidation or pleural effusion. CT imaging is used to provide more information about the type and extent of the disease. Bronchoscopy or CT- guided biopsy is often used to sample the tumor for histopathology. The defective diagnosis of lung cancer is based on histological examination of the suspicious tissue in the context of the clinical and radiological features. CT imaging should not be used for longer or more frequently than indicated as extended surveillance exposes people to increased radiation.

Worldwide in 2012, lung cancer occurred in 1.8 million people and resulted in 1.6 million people deaths. This is the most common cancer- related death in men and 2nd most common in women after “breast cancer”. The most common age at diagnosis is 70 years.

III. Proposed System

In earlier days and still, predicting the disease as earlier is not possible. If disease is predicted earlier which will help the doctor to save the life of the people? This paper proposes to predict the disease as early as possible based on the symptoms. Data mining techniques like classification and clustering are helpful for predicting the disease. We can predict the disease by using the data mining hybrid approach. This paper will predict the disease state based on the symptoms.



Workflow for Disease Prediction System

The user will enter the symptoms according to the disease states he is suffering from the disease. The mapping of user symptoms and the prior database is once done; the result will be generated according to the disease state and level of affection.

IV. Classification

Classification is the process of finding a set of models (or functions) which describe and distinguish the data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown(5,6). The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae or neural networks. A decision tree is a chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. A neural network is a collection of linear threshold units that can be trained to distinguish objects of different classes. Classification can be used for predicting the class label of data objects. In many applications, one may like to predict some missing or

unavailable data values rather than class labels. When the predicted values are numerical data and are often specifically referred to as prediction. Prediction may refer to both the data value prediction and class label prediction; it is usually referred to data value prediction and thus is distinct from classification. Classification is a data mining machine learning technique used to predict group membership for data instances. Popular classification techniques include decision tree and neural networks. The Naïve Bayesian classifier is one of the classification algorithms and is based on Bayes theorem. A Naïve Bayesian algorithm is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x) and P(x|c). Naïve Bayes classifier assumes that the effect of the value of a predictor (x) on a given class(c) is independent of the values of other predictors.

$$P(c/x) = \frac{P(x/c)*P(c)}{P(x)}$$

Where,

P(c|x) is the posterior probability of class (target) given predictor(attribute)

P(c) is the prior probability of class

P(x|c) is the likelihood which is the probability of predictor given class

P(x) is the prior probability of predictor

V. Bayesian Classification

It is based on Bayes Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict the class membership probabilities such as the probability that a tuple belongs to a particular class.(3)

Bayes theorem is named after Thomas Bayes. There are 2 types of probabilities (4)

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

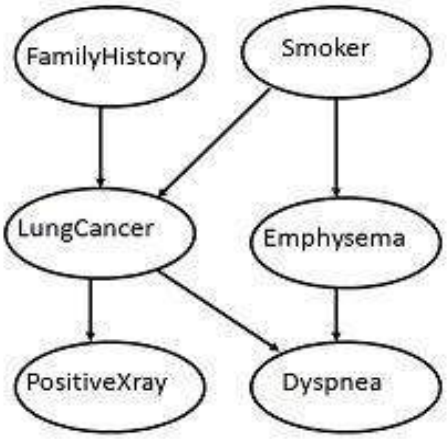
Where X is a data tuple and H is some hypothesis.

According to Bayes Theorem,

P(H/X) = P(X/H)P(H)/P(X)

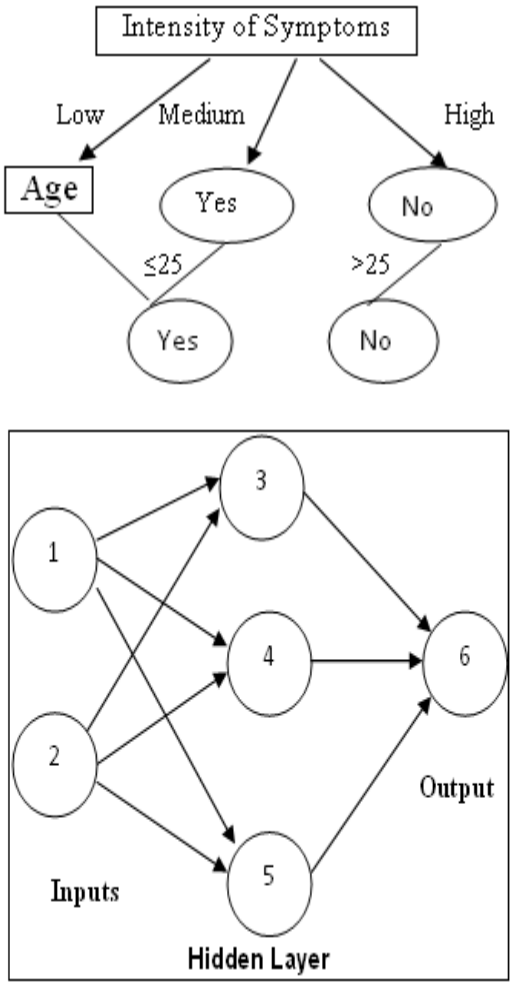
Bayes theorem is the method of finding the converse probability of the unconditional,

P(E/C)=P(C/E)P(E)/P(C) =P(C,E)/P(C)



VI. Decision Tree

A Decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test and each leaf node holds a class label.(7)



The top most node in the tree is the root node. Next to top most nodes is the leaf node. The user will enter the symptoms. It can be classified as low, medium, high Level based on age in the above decision tree structure.

VII. Conclusion and Future Work

Prevention of lung diseases is low in India, especially in rural, did not notice at early stage, because of lack of awareness. In this paper am proposing a system which can predict the diseases based on the input symptoms provided by the user and help them to analyze their health status so people can take some precautions as per the result. It could help doctors to know the health state of the patient and based on that manual diagnosis of the disease can also be easily possible. In Future work, have planned to conduct experiments on real time large health datasets to predict all the diseases and compare algorithm with other data mining algorithm. Continuous data can also be used.

References

[1] Banu, M.N. and Gomathy, B. Disease Predicting System Using Data Mining Techniques. *International Journal of Technical Research and Applications* **1** (5) (2013) 41-45.

[2] Ahmed, K., Abdullah-Al-Emran, A.A.E., Jesmin, T., Mukti, R.F., Rahman, M. and Ahmed, F. Early detection of lung cancer risk using data mining. *Asian Pacific Journal of Cancer Prevention* **14** (1) (2013) 595-598.

[3] Pradhan, M. and Sahu, R.K. Predict the onset of diabetes disease using Artificial Neural Network (ANN). *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)* **2** (2) (2011).
[4] Pattekari, S.A. and Parveen, A. Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences* **3** (3) (2012) 290-294.
[5] Vijayarani, S. and Divya, M. An efficient algorithm for generating classification rules. *International Journal of Computer Science and Technology* **2** (4) (2011).
[6] Agrawal, A. and Choudhary, A. Association rule mining based hotspot analysis on seer lung cancer data. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)* **2** (2) (2011) 34-54.
[7] Freund, Y. and Mason, L. The alternating decision tree learning algorithm. In *ICML*, 1999, 124-133.