Data Mining Techniques for Early Cancer Diagnoses

Abstract

The increase in population along with the growth in diseases has made the inclusion of data mining in medical diagnosis to extract the underlying pattern. Cancer is one of the widespread diseases that claims 7 million lives every year and also lung cancer accounts for 18% of these mortalities.Earlier researches and case studies indicate that the survival rate of the patients suffering from cancer is higher when the disease is diagnosed at an early stage.

Existing medical techniques like X-Ray, Computed Tomography (CT) scan, sputum cytology analysis and other imaging techniques not only require complex equipment and expensive but is also proven to be efficient only in stage 4, when the tumor has attacked all parts of the body.

Introduction

Main characteristic of disease is, uncontrolled growth of cells. If it is not diagnosed and treated early, the tissues could spread to other parts of the body such as the brain, bone, liver and adrenal gland.

There are two major types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). (85% of NSCLS and 13% (SCLC). Because the lungs are big, tumors can grow in them for a long time before they are found. Even when symptoms like coughing and fatigue occurs, people think that they are due to other causes. Because of this reason, early-stage of lung cancer (stages I and II) is difficult to detect. Most people with NSCLC are diagnosed only at stages III and IV.

In spite of the available existing techniques, most of the time lung cancer is detected only after crossing stage 1. The lung cancer five-year survival rate is 16.3% where 52.6% for cases that are detected early, when the disease is still localized (within lungs). However, only 15% of lung cancer is diagnosed at this early stage. So, over half of the people with lung cancer die within one year of being diagnosed.

Lung cancer, a disease which is highly dependent on historical data can make use of data mining for its early detection.This abstract proposes a model for measuring if applying data mining techniques to lung cancer dataset can provide reliable performance in the detection of lung cancer at Stage I.

Survey

Ahmed et al, implemented a model to diagnose Lung cancer at an early stage using k-means clustering. The significant patterns are then discovered using Apriori Tid and Decision Tree algorithm. Apriori Tid, which is an extension of Apriori algorithm is one of efficient algorithms to mine frequent item sets. Apriori algorithm called two sub-processes which are Apriori-gen() and subset(), Apriori-gen() process produces a candidate for lung cancer, then use the Apriori property to delete those candidates of the non-frequent subsets. Once candidates gets generated, the database will be scanned and for each transaction, the Subset() sub procedure is used to identify all the candidate subsets, and make cumulative count for each of these candidates The major drawback of this system is that it requires lot of database scans as the number of attributes increases and it could occupy more space and memory for candidate set generation.

Oh et al, proposed a method for predicting local failure lung cancer post radiation therapy using Bayesian network. Bayesian networks encode the relations between variables using probability theory. They are used predictions based on the encoded relations. The attributes of the patient records are assigned to the nodes of the graph. The joint probability distribution function is then encoded by the network as per equation

$$P(X) = \prod_{i=1}^{n} P(X_i|par(X_i)$$

This Bayesian model makes use of both the physical attributes and the biological attribute, i.e, the blood biomarkers, for predicting the local failure.

Palanisamy et al made a model to detect significant genes of leukemia cancer using k-means clustering algorithm. We know that k-means clustering always converges to a local minima. Classification accuracy depends on the starting cluster centroid selection which was one difficulty faced with the K-means algorithm.

Bayesian method proposed by Oh et al. gave a better accuracy because of inclusion of physical parameters in combination with the biological parameters.

Procedure:

Recently, digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases and made available globally. Data mining techniques applied on these databases discover relationships and patterns.

Data Integration

Heterogeneous data from health organizations are collected from multiple sources and made into a common source.

Data Selection

Dataset containing irrelevant and missing data, are discarded. Relevant data is retained.

Data cleaning

Few patient record contain errors, noise or missing information. Data are corrected and those that cannot be corrected are discarded. Fuzzy Self Organising Maps can be used to filling the missing values. Some useful attributes could be Smoking habit, Radon gas, Organ transplantation, AIDS.

Data Transformation

Large values are normalized so that the calculations can be made faster. After the normalization and the discretization process, the records of the patients are represented in the form of a matrix.

Data Classification

Interesting patterns are discovered in this step. Lot of techniques have been discussed are applied here.

Conclusion

A comparison of different data mining techniques could produce an efficient algorithm for lung cancer classification for multiple classes.