Lung Cancer Prediction using Data Mining techniques

Data Mining Project

Why do we do Cancer Prediction??

- The increase in population coupled with the growth in diseases has necessitated the inclusion of data mining in medical diagnosis to extract the underlying pattern. Of these, cancer claims over 7 million lives every year and lung cancer accounts for 18% of these mortalities.
- Existing medical techniques like X-Ray, Computed Tomography (CT) scan, sputum cytology analysis are of complex equipment with high cost and also proven to be efficient only in stage 4, when the tumor has spread to other parts of the body.
- The proposed system involves the development of a data mining tool that will help in the classification of patients into the category that could potentially test positive for lung cancer in stage 1.

About Lung Cancer (briefly....)

- Cancer is two types i.e., Non-Small Cell Lung Cancer (NSCLC), Small Cell Lung Cancer (SCLC) and Mixed Small Cell/Large Cell Cancer (if both).

Indications for Cancer disease

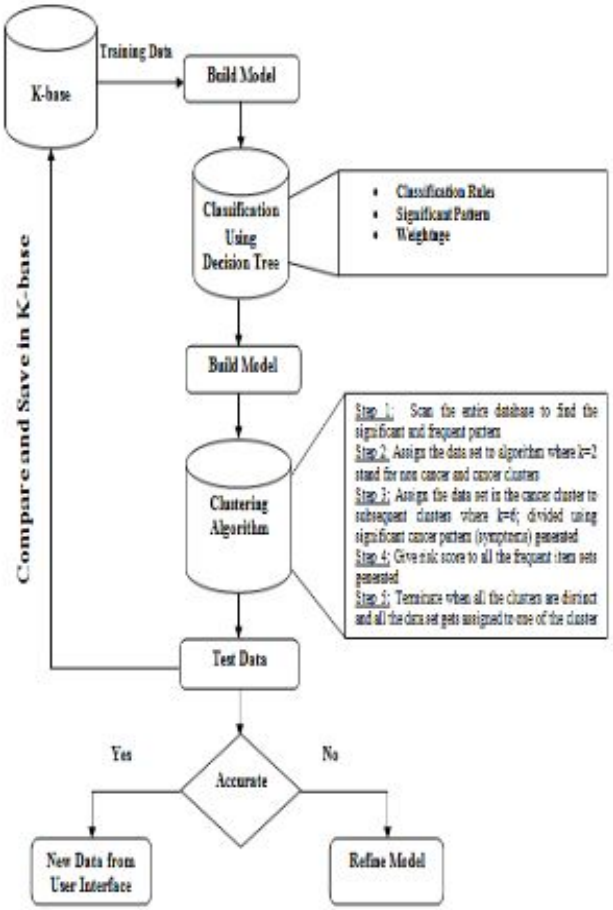
Symptoms

- Chest, shoulder, or back pain that doesn't go away.
- Weight loss and loss of appetite.
- Repeated respiratory infections, such as bronchitis or pneumonia.
- cough that does not go away and gets worse over time

Risk factors

- Smoking: i. Beedi ii. Cigarette iii. Hukka
- Second-hand smoke.
- Air pollution.
- Insufficient consumption of fruits & vegetables.
- Suffering with other types of malignancy

Model



Step 1

- The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm.

Step 2

- Then the data is clustered using K- means clustering algorithm to separate cancer and non cancer patient data.

Why Decision tree algorithm?

Why K-Means Clustering algorithm?

Datasets

- Kaggle Website

https://www.kaggle.com/c/data-science-bowl-2017

- Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/lung+cancer

Attributes can be “**Age, Marital status, Symptoms relating to cancer, occupational hazards, family history of cancer**” etc.

Classification and Significant Pattern Generation

- The frequent itemsets that occur throughout the database and have a significant link to cancer status are mined as significant patterns.
- The data is fed into the decision tree algorithm to obtain the significant patterns related to cancer and non cancer data sets.

Weightage to Significant Pattern

- Weightage is calculated for every frequent pattern based on the attributes to analyze its impact on the output.
- They should satisfy the condition **Sw(i)=Sum of (W(i)*F(i)) where Sw = Significant Weightage , W(i) = Weightage of each attribute , F(i)= Frequency of the rule** to take it as a Significant pattern.

Example of Significant pattern

- Age - gender - living area - family history- anemia-symptoms -> none- Cancer Type -> None.

  Weightage =100.55

Example of Decision rule

- If symptoms = none and risk score = x <45 then
  result = you don't have cancer,
  tests = do simple clinical tests to confirm.

Risk scores and their respective attributes

Age	x<30 30 < x < 40 40< x < 60	3 4 5
Habits	Smoking Alcohol Chewing Hot beverage	3 5 3 2
Family History of Cancer	Yes No	5 1

*All these attributes and their risk scores are tentative and they could be modified to increase accuracy.

K-Means Clustering

- The instances are now clustered into a number of classes where each class is identified by a unique feature based on the significant patterns mined by the decision tree algorithm.
- The aim of clustering is that the data object is assigned to unknown classes that has a unique feature and hence maximize the intraclass similarity and minimize the interclass similarity.

Algorithm

Input: K- the number of clusters, D- data set containing n objects.
Output: Clusters

Step 1:

- Choose two mean values from weightage of significant patterns as the initial cluster centers.

Step 2:

- Choose two mean values from weightage of significant patterns as the initial cluster centers.

Step 3:

- Assign each object to the cluster to which it is most similar based on the mean value of the weightage

Step 4:

- Update the cluster means by calculating mean value of all the objects in the cluster

Now two clusters have been generated based on the weightage scores of the significant pattern. The two clusters are named as **Non cancer** and **Cancer clusters**

Thank you...

COE15B019
COE15B029
CED15B010
CED15BI033