St. Joseph's Journal of Humanities and Science (Volume 5 Issue 2 August 2018) 44-49

St. Joseph's Journal of Humanities and Science
ISSN: 2347 - 5331
http://sjctnc.edu.in/6107-2/

A Parametric Model for identifying Lung Cancer using Symptoms with Clustering Techniques

- R. Vidya[a*]
- V. Latha[b]

Abstract

Cancer is the most important ruin in worldwide. Early detection and prevention of cancer plays a essential role in decreasing death caused by cancer. This paper uses data mining techniques such as classification, clustering and prediction. To identify potential cancer patients by using decision tree algorithm in classification. For partitioning affected cancer and non affected cancer data by using K-means clustering algorithm. Prognosing the risk level is achieved by prediction.

Key words: Decision Tree, k-means, Prediction, Risk Levels, classification, clustering

INTRODUCTION

Data mining is the withdrawal of mysterious prognostic data and unwanted data, figures, correlation and observation by scrutinize the enormous data sets which are onerous to find and observe with conventional numerical methods. It is dynamic technology which will determine most related data from the data storehouse of the system. It's very imperative step that generally inspect huge amount of customarily data[1]. To find latest figures in healthcare commerce, there exist numerous bilateral and ascendable data mining methods. It is a perceptible access which is convenient in reviewing results and diminishing errors and restraint the quality more consistently.

Cancer is the most typical diseases in worldwide that results in majority of death. It is caused by spontaneous production of cells in any of the tissues or parts of the human structure. This may happen in any of the parts in human structure and will spread to several other parts. In early disclosure of cancer at the begin stage and avoidance from spreading to the other parts in malevolent stage could save a person's life[3].

[a] Assistant Professor, Department of Computer Science, St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Tamil Nadu, India.
[b] Research Scholar, Department of Computer Science, St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Tamil Nadu, India.
*E-mail ID: vidya.sjc@gmail.com
Mobile No.: +91 9443222181

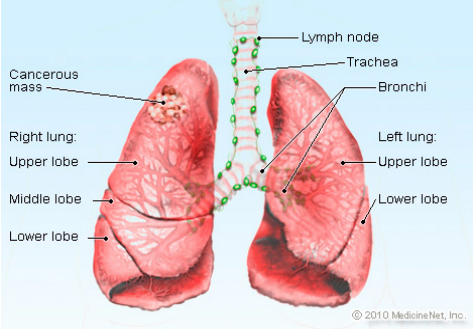

Fig. 1: Lung Cancer

Fig 1 explains the Lung cancer. It is the spontaneous production of aberrant cells that start off in one or both lungs generally in the cells that line the air passages[7]. Fig 2 relates among both men and women is the main act of cancer deaths. Small cell lung cancer and non-small cell lung cancer are two various types of lung cancer. These types are recognized based on how the cell seems with a microscope.

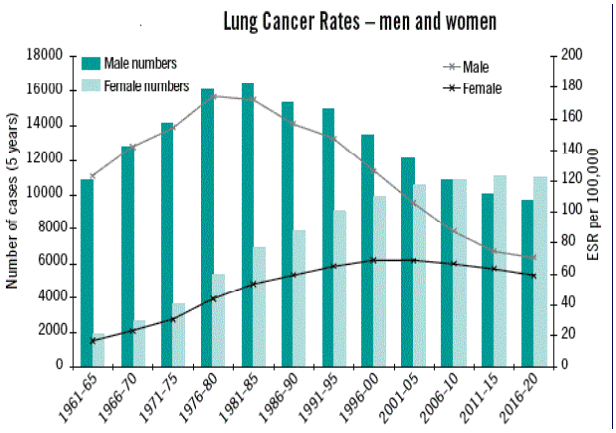

Fig. 2

DATA MINING TECHNIQUES

Data mining technique engage the use of refined data analysis tools to devise previously unknown, valid figures and relationships in a huge data set [2]. This implement can combine statistical models, mathematical algorithm and machine observing methods in early disclosure of cancer. In classification learning, the learning scheme is conferred with a set of classified examples from which it is to learn a form of classifying lurking examples. It can easily altered into classification rules. This decision tree is adopted to achieve continual figures in the data set.

Clustering is a progress of segregating dataset into many groups resultant to their aspects[1]. In K-means clustering, the numeral of clusters required is found out and then an algorithm is adopted to appropriately associate or disassociate instances with clusters until associations stabilize around k clusters. The extensive part of this study is to predicting the hazard level of lung cancer using MATLAB tool.

PROPOSED MODEL

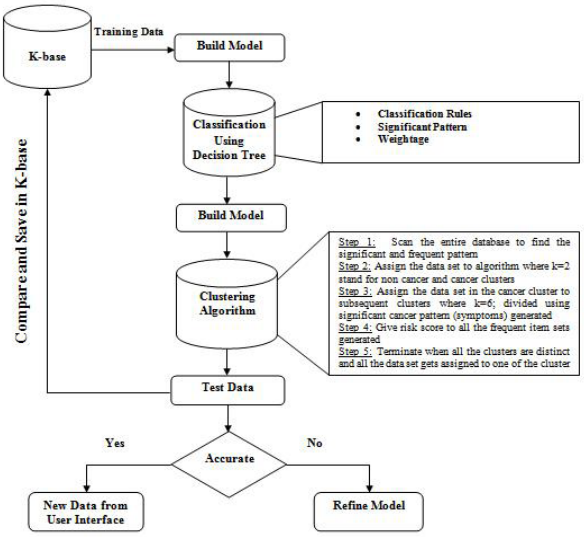

Fig. 3.1: Proposed Work

The above figure 3.1 is the proposed work in above model. The together data is pre-processed and saved in observation base to create the model. Seventy percent of incorporated data is taken as learning set to construct the classification and clustering model the remnant of which is taken for examination purpose. The decision tree model is create using the classification rules, the symptomatic frequent figures and its correlative weight. k-means clustering algorithm is adopted to construct clustering model. This model is then examined for correctness, sympathy and specificity using test data along with incorporate it to the knowledge base.

LITERATURE REVIEW

Ada et al [1] tried out to identify the lung tumours from cancer images and support tool is matured to identify the normal and abnormal lungs and to anticipate survival rate and years of an unbalanced patient so that patients live could be saved.

V.Krishnaiah et al [2] developed a sample lung cancer prognosis system by the use of data mining classification techniques.

CharlesEdeki et al[3]The dataset surpass the others in such way that it can be avowed the optimal algorithm

and not any of the algorithm worked poorly as to be concluded from future prediction model in lung cancer survivability tasks.

Rajashree Dash et al [4] a hybridized K-means algorithm have been suggested which integrate the steps of dimensionality contraction through PCA, a novel initialization access of cluster centrum and the steps of designated data points to opportune clusters. Using the proposed algorithm a provided data set was segregated in to k clusters.

Ritu Chauhan et al [5] focus on clustering algorithm such as HAC and K-Means are, HAC is to force on K-means to measure the numeral clusters. The value of the cluster is enhanced, if HAC is obligatory on K-means.

Dechang Chen et al [6] algorithm EACCD matured two step clustering method. In the first step, a variation dimension is studied by using PAM, and in the moment step, the studied variation is used with a hierarchical clustering algorithm to procure clusters of patients.

METHODOLOGY

In genral literature reviews, case studies and discussions with medical professionals mentioned that there are numeral factors influencing cancer. Those factors are taken as attributes for this learn.

Data Source

The source for this study was collected from a prominent Cancer registry Pondicherry, subsisting of non cancer and cancer patients data and they are pre-processed to serve this study. These attributes are aid to train and develop the arrangement and a part is used to test the connotation of the system. These attributes play a imperative role in diagnosticate cancer in all the cases. This data is stored in a observation base which has the capability to enlarge itself as new data enters the system over front end from which new knowledge is attained and thus the system becomes innovative.

Classification and Significant Pattern Generation

Decision tree algorithm is used to mine continual patterns from the data set. The continual item sets that result right through the data base and have a indicative link to cancer status are mined as significant patterns[2] .The data is fed into the decision tree algorithm to retrieve the significant patterns associate to non cancer and cancer data sets. In other words the example that are mined by the decision tree are well defined and dignified to be divided as non cancer and cancer datasets.

A batch of candidate attributes II, and **R,** a set of named instances is given as input. The algorithm to generate a decision Tree **U** is as proceed from Start

1. If (R is pure or empty) or (II is empty) Return U.
2. Compute Ps (Ci) on R for each class Ci.
3. For every attribute Y in II, compute IIG(R, Y) relay on equation 1 and 5.
4. Use the attribute Ymax with the highest IIG for the root.
5. Partition R into disjoint subsets Rx using Ymax.
6. For all values y of Ymax •Tx=NT (II-Ymax, Rx), •Add Tx as a child of Ymax.
7. Return U End.

Significant Pattern mined using Decision Tree Algorithm

1. Age - gender - living area - family history- anemia-symptoms -> none- Cancer Type -> None. Weight = 200.55
2. Age - gender- marital status-education-smoking-diet-symptoms-> Pain in chest, back, shoulder or arm->Shortness of breath and hoarseness-Cancer Type->Lung Weight = 300.50
3. Gender-Education-Occupational hazards- Alcohol-Family history- Weight loss- symptoms-> severe abdominal pain or bloating-> abdominal pain with blood in stool- Cancer Type ->Stomach Weight = 280.05.

Weight for Significant Pattern

The weight is computed for every single frequent pattern relay on the attributes to analyze its impact on the input. The frequent patterns mined which satisfies the below condition are taken as significant Frequent Pattern.

$$\mathbf{Rw(i) = \Sigma(Xi*Gi)}$$

(1).Where Xi is the weight of each attribute and Gi represents number of frequency for each rule and significant Frequent Pattern is selected by using the Equation (2) **S**FP=Rw (n) $\geq \varphi$ for all values of n (2). Where **S**FP denotes significant frequent pattern and φ denotes significant weight.

Table -1: Risk scores for the attributes that represent the significant patterns.

Attributes	Values	Risk score
Age	x < 20	4
	20 < x < 30	3
	30 < x < 60	2
Education	Uneducated	4
	School	2
	College	1
Living Area	Urban	3
	Rural	2
Habits	Smoking	1
	Alcohol	4
	Chewing	2
	Hot beverage	3
Occupational Hazards	Radiation Exposure	3
	Chemical Exposure	3
	Sunlight Exposure	2
	Thermal Exposure	2
Anemia	Yes	2
	No	1
Weight Loss	Yes	1
	No	2
Family History of Cancer	Yes	3
	No	2

Decision Tree Rule

If *symptoms* = none and *risk score* = *y* > 35 then *result* = you don't have cancer, *tests* = do simple clinical tests to confirm. If *symptoms* = none and *risk score* = 35 < y < 60 then *result* = you may have cancer, *tests* = do blood test and x ray to confirm

Else if *symptom* = related to stomach and *risk score* = *y* > 35 then *result* = you have cancer, *cancer type* = stomach, *tests* = endoscopy of stomach

Based on the above specified rules and the calculated hazard scores the severity of cancer is known as well as some tests were recommended to confirm the existence of cancer.

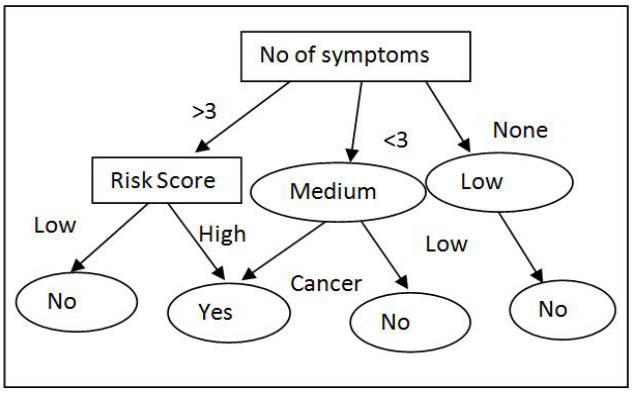

Fig. 4 : Decision Tree

Clustering using *k*-means

The k-means are now divided into numerous of classes where each class is identified by a unique feature based on the significant patterns mined by the decision tree algorithm. The partition of clustering is that the data object is designated to unknown classes that has a unique feature and hence greater the intraclass similarity and smaller the interclass similarity[9].The weight scores of the significant patterns mined are fed into K- means clustering algorithm to cluster and split it into cancer and non cancer cells. At the beginning the data is designated to a non cancer cluster and then based on the strength of the cancer given by its weight it is either moved to the cancer cluster or gets retained in the non cancer cluster, further the data object is moved between the many groups of the hierarchical cancer cluster based on the symptoms the data object contains. To compute the mean of the cluster center the symptoms are given certain values the average of which represents each distinguished cluster.

It also searches the whole database to detect a match to a one input data. The data is shifted to that particular cluster if and only if an exact match is found. This technic lower the error rate of clustering. The data in first cluster are all similar with small or no indications; no risk factors correlate with cancer and less risk scores. Hence the cluster is known as Non cancer cluster[7]. The peak cluster of the second hierarchical cluster contains all the data that has big risk factors correlate with cancer along with distinguished indications and high hazard scores[8].

Clustering Algorithm

Algorithm: The k-means clustering algorithm is used for segregating the data into cancer and non cancer clusters, where the initial cluster centers is expressed by the mean value of the weight of significant patterns.

Input: l: the number of clusters. E: data set consist of n objects.

Output: B set of hierarchical clusters

Start

1. Choose two mean values from weight of significant patterns as the initial cluster centers;
2. Assign each object to cluster to which it is most identical based on the mean value of the weight.
3. Amend the cluster means by manipulating mean value of all the substance in the cluster.
4. End

Now two clusters have been engendered based on the weight scores of the significant pattern. The two clusters are characterized as Non cancer and Cancer clusters. The mean weight of the cancer cluster is extremely lower than the non cancer cluster. Repeat partition the cancer cluster to generate six sub clusters each denoting a type of cancer.

Start

1. Chooses 1 objects from cancer cluster R with distinguished values for its indication.
2. Assume each object in R to the cluster whose mean value is closer to its indication
3. Refine the cluster means and
4. Repeat step 2 and 3 unless negative change
5. End.

The result is six clusters with each denoting a type of cancer.

EXPERIMENTAL RESULTS

The results are divided into three parts. The first is the frequent and significant pattern discovery. The second is mapping the cancer to its cluster and the third is prediction by giving hazard level as output. At the starting level all the input data is store in the non cancer cluster neither it gets classified and clustered by this model. A user input data is fed to the system and gets classified accordance to the significant pattern to which it matches through decision tree, gets analyzed for its risk level complex with either one cancer cluster either it gets clustered and classified.

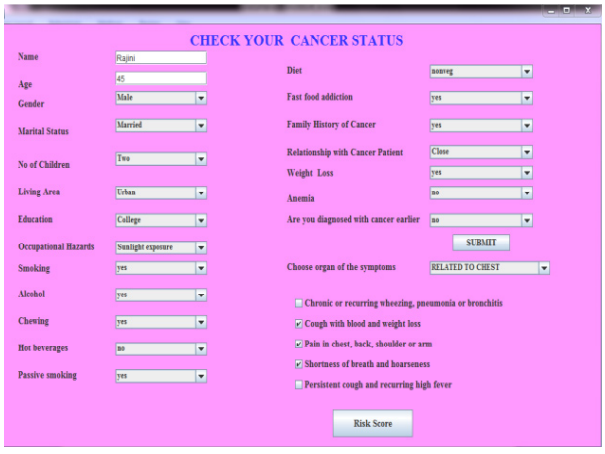

Fig. 5 : User Input Screen

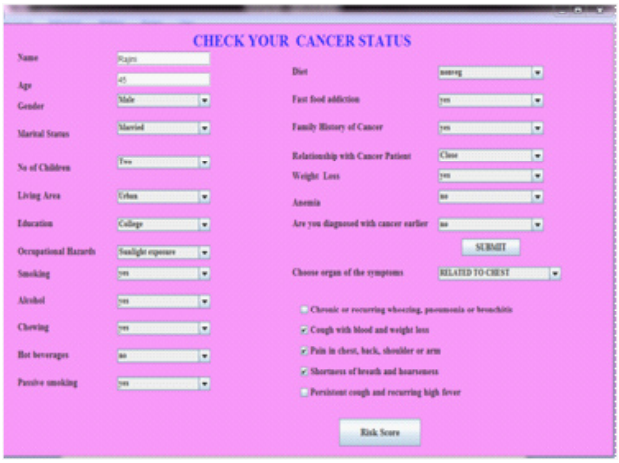

Fig. 6 : Report Screen with Prediction Results

The screen display the whether the patient having cancer or not. The person has cancer by matching his data with the entire database, the person's risk score developed by the significant pattern mined by decision tree, the type of cancer he has which is given as a cluster output, neither his risk status is medium or severe and finally some recommended tests by medical experts to ensure the survival of cancer.

CONCLUSIONS

In this learning various stratified method combining decision tree and clustering techniques to build a cancer risk prediction system is proposed. In worldwide cancer has become the leading causation of death . The most significant way to lower cancer deaths is to find it earlier. These prediction system may contribute easy and a cost effect way for display cancer and may play a crucial role in early diagnosis process for various types of cancer and provide effect anticipatory strategy.

REFERENCES

1. Ada and Rajneet Kaur "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6, ISSN 2320–088X
2. V.Krishnaiah "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 www.ijcsit.Com ISSN: 0975-9646.

3. Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" Mediterranean journal of Social Sciences Vol 3 (14) November 2012, ISSN: 2039-9340.

4. A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org

5. Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66

6. Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

7. Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

8. S M Halawani "A study of digital mammograms by using clustering algorithms" Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.

9. R.Vidya and G.M nasira "A novel medical support system for the social ecology of cervical cancer:A research to resolve the challenges in pap smear screening and prediction at firm proportion" advances in natural and applied science 9.6 SE(2015)