Cancer Prediction System Using Data Mining

Ms. Amruta Yadav*, Ms. Shreya Hedaoo*, Ms. Konika Katkar*, Ms. Pooja Nikhare

Dept of Information Technology

Rashtrasant Tukadoji Maharaj Nagpur University

ABSTRACT:

It might have happened so many times that you or someone yours need doctors help immediately, but they are not available due to some reason. The Cancer Disease Prediction application is an end user support &online consultation project. Here, we propose a web application that allows users to get instant guidance on their cancer disease through an intelligent system online. The application is fed with various details &the cancer disease associated with those details. The application allows user to share their health related issues for cancer prediction. It then processes user specific details to check for various illness that could be associated with it. Here we use some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's details. Based on result, system automatically shows the result specific doctors for further treatment. The system can be use in case of emergency.

General Terms: Data Mining, Clustering, Classification

Keywords : **Decision Tree, k-means, Prediction, Risk Levels**

1.INTRODUCTION:

Cancer may occur in any part of the body & may spread several other parts. There are several factors that could affect a person, predisposition for cancer. Education is an important indicator of socio-economic status through its association with occupation & life-style factors. A number of study in developed countries have shown that cancer incidence varies between people with different levels of education. A high incidence of breast cancer has been found , among those with high levels of education, whereas an inverse association has been found for the incidence of cancers of the stomach, lung & uterine cervix. Such difference in cancer risks associated with education also reflect in the differences in life-style factors & exposure to both environmental work related carcinogens. This study describes the association between cancer incidence pattern & risk levels of various factor by a risk prediction system for different types of cancer which helps in prediction.

Data mining technique involves the use of data, analysis tools to discover previously unknown, valid result & relationships in large data set. These tools can include models, algorithm & mining methods in cancer prediction. In classification the learning scheme is presented with a set of classified examples, from which it is expected to learn a way of classifying unseen examples. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is numeric quantity. In this study, to classify the data & to mining frequent patterns in data set Decision Tree algorithm is used. A decision tree is a tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test & each leaf node holds a class label. The top most node is the root node. The attribute value of the data is tested against a decision tree. A path is traced from root to leaf node, which holds the class prediction for that data.

Decision trees can be easily converted into classification rules. It is used to generate frequent patterns in the dataset. The data & item sets that

occur's frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types & are helpful in predicting the cancer & its type is known as Significant frequent pattern. Using significant patterns generated by decision tree the data set is clustered & risk scores are given. Clustering is a process of associate or dissociate instances with clusters until associations stabilize around k-clusters. Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a

2.PROPOSED MODEL:

The collected data is pre-processed &stored in the knowledge, base to build the model.75% of the entire data is taken as training set to build the classification &clustering model the remaining of which is taken for testing purpose. The decision tree model is build using the classification rules, the significant frequent pattern & its corresponding weightage. The clustering model is build using the k-means clustering algorithm. The model tested for accuracy, sensitivity &specificity using test data along with merging it with knowledge base.

3.MATERIAL & METHODS :

3.1 DATA SOURCES:

This data consists of 30 attributes such as Age, Marital status, Symptoms relating to cancer, occupational hazards, family history of cancer etc. These attributes are used to train &develop the system & a part is used to test the significance of the system. These attributes play an important, role in diagnose of cancer in all the cases. This data is stored in a knowledge base which has ability to & itself as new data enters the system through front end from which new knowledge is gained &thus the system becomes intelligent.

3.2 CLASSIFICATION & SIGNIFICANT PATTERN:

Decision tree algorithm is used to mine frequent patterns from the data set. The data is fed into decision tree algorithm to obtain the significant patterns related to cancer & non-cancer data sets.The patterns that are mined by the decision tree are well defined &

separating dataset into subgroups with their unique features. A cluster is a collection of data objects that are similar to one another within the same cluster & are dissimilar to the objects in other clusters. In K-means clustering, the number of clusters needed is found out & then an algorithm is used to successively

patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances.

distinguished to be separated as cancer & non-cancer datasets. A set of candidate attributes II & S a set of labelled instances is given as input. The algorithm to generate a decision Tree T . If (S is pure or empty) or (II is empty) Return T. Compute Ps (Ci) on S for each class Ci. For each attribute X in II compute II G(S, X) based on equation. Use the attribute X-max with the highest II G for the root. Partition S into disjoint subsets Sx using X-max. For all values x of X-max •Tx=NT (II-X-max, Sx), •Add Tx as a child of X-max. Return T End.

3.2.1 SIGNIFICANT PATTERN USING DECISION TREE ALGORITHM

1) Age - gender - living area - family history-symptoms -> none- Cancer Type -> None. Weightage = 100.55 2) Age - gender- marital status-smoking-diet symptoms-> Pain in chest, back, shoulder or arm->Shortness of breath & hoarseness-Cancer Type->Lung Weightage = 200.50 3) Gender-Education-Alcohol-Family history- Weight loss- symptoms-> severe abdominal pain or bloating-> abdominal pain with blood in stool- Cancer Type >Stomach Weightage = 180.05 4) Age- gender- no of children- occupational hazards- Family history- relationship with cancer patient- symptoms-> swelling or mass in armpit -> discharge or pain in nipple -> Cancer Type -> Breast. Weightage = 170.55 5) Gender- education- living area- Smoking- Hot beverage- Diet- fast food addiction- Earlier cancer diagnosis- symptoms> Ulcers in mouth or pain of teeth & jaw-> White or red patches in tongue, gums- Cancer Type -> Oral. Weightage = 190.50 Numeric values are given as risk scores to the

attributes that have a direct link to the significant patterns .

Table 1. Risk score represent the significant patterns according to parameter.

3.2.2 WEIGHTAGE SIGNIFICANT PATTERN :

The weightage is calculated for every frequent pattern based on the attributes to analyze its impact on the output. The frequent patterns mined which satisfies the below condition are taken as significant Frequent Pattern.

Parameters	Values	Risk Score
Age	X<30	3
	30<x<40	4
	40<x<60	5
Living Area	Urban	5
	Rural	3
Occupational Hazards	Chemical	3
	Radiation	3
	Sunlight	2
	Thermal	2
Habits	Alcohol	5
	Smoking	3
	Chewing	3
Anemia	No	1
	Yes	3
Weight loss	No	1
	Yes	2
Family History	No	1
	Yes	5

$Sw(i)= \Sigma(Wi*Fi)$ 1.Where Wi is the weightage of each attribute & Fi represents number of frequency for each rule. & significant Frequent Pattern is selected by using the following Equation. $2.SF=Sw (n) \geq \varphi$ for all values of n 2. Where SF denotes significant frequent pattern & φ denotes significant weightage.

3.2.3. RULES FOR DECISION TREE :

If symptoms = none & risk score = x <44 then result = you don't have cancer, tests = do simple clinical tests to confirm. If symptoms = none &risk score = 44< x < 50 then result = you may have cancer, tests = do blood test & x ray to confirm Else if symptom= related to stomach & risk score = x > 44 then result = you have cancer, cancer type = stomach, tests = endoscopy of stomach If symptom= related to breast & shoulder & risk score =x > 44 then result = you have cancer, cancer type = breast, tests= mammogram & PET scan of breast If symptom= related to chest & shoulder & risk score =x > 40 then result = you have cancer, cancer type =lung tests = take CT scan of chest. If symptom= related to pelvis & lower hip & risk score= x >55 then result = you have cancer, cancer type = cervix, tests = do pap smear test. If symptom= related to head & throat & risk score =x > 40 then result = you have cancer, cancer type = oral, tests = biopsy of tongue & inner mouth. Else symptom= other symptoms & risk score =x > 40 then result = you have cancer, cancer type = leukemia, tests = biopsy of bone marrow.
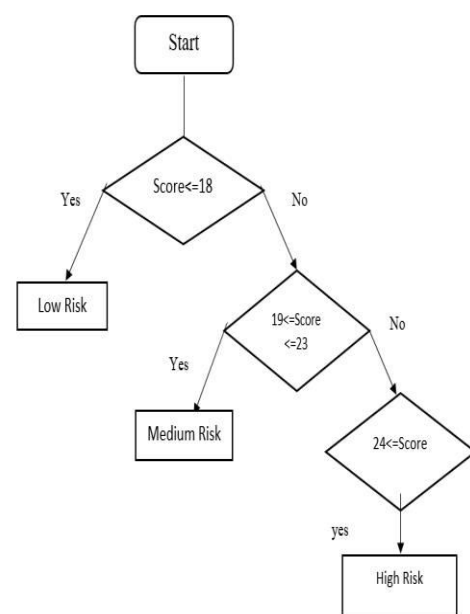
Figure 2. Flow Diagram of decision tree algorithm

4.CLUSTERING USING K-MEANS:

The instances are now cluster into a number of classes , where each class is identified by a unique feature based on the significant pattern mined by the decision tree algorithm. Aim of clustering is that the data object is assigned ,to unknown classes that has a unique features & hence maximize the interclass, similarity & minimize the interclass . The weightage scores of the significant pattern mined are fed into K-means clustering algorithm to cluster & divide it into cancer & non-cancer groups. The group is further subdivided into six groups with each cluster representing a type of cancer. The data is assigned to a non-cancer cluster & then based on the intensity of the cancer given by its weightage either moved to the cancer cluster or gets retained in the non-cancer cluster, further the data object is moved between the sub-groups of the hierarchical cancer cluster, based on the symptom the data object contains.

Calculate the mean of the cluster center the symptom are given certain values the average of which represents each distinguished cluster. The data objects are distributed to the cluster based on the cluster center to which it is nearest. It searches the entire database to find a match to a single input data. The data is moved to that particular cluster if & only if exact match is found. This technique minimizes , the error rate of clustering. The data in the first cluster are all similar with little or no symptoms; no risk factors associated with cancer & low risk score. Hence cluster is labeled as Non-cancer cluster. The top cluster of the second hierarchical cluster contains all the data that has high risk factors associated with cancer along with distinguished symptom & high risk scores. The data in the cluster is again fed into k-means clustering algorithm to further subdivide it. The result six clusters are separated based on particular symptom associated with any one type of cancer i.e. lung, cervix, breast, stomach, oral & leukemia. All the data is partitioned into two types of clusters & six sub clusters of the cancer cluster.

4.1CLUSTERING ALGORITHM:

Algorithm: The k-means clustering algorithm is used for partitioning the data into cancer & non-cancer clusters, where the initial cluster centers is , represented by the mean value of the weightage of significant patterns.

Input: The number of clusters (k). D: data set containing n objects. Output: A set of hierarchical clusters Begin 1] Choose 2 mean values from weightage of significant patterns as the initial cluster center; 2] Assign each object to the cluster to which it is most similar ,based on the mean value of the weightage. 3] Update the cluster means by calculating mean value ,of all the objects in the cluster. 4] End. Now 2 clusters have been generated ,based on the weightage scores of the significant pattern. The 2 clusters are named as Non-cancer &Cancer clusters. The mean weightage of the non-cancer cluster is significantly lower than the cancer cluster. Again partition the cancer cluster to generate, 6 sub clusters each representing a type of cancer.

Begin 1] arbitrarily chooses k objects from cancer cluster S with distinguished values for its symptoms. 2] Assign each object in S to the cluster whose mean value is closer to its symptom. 3] Update the cluster means &4] Repeat step 2 &3 until no change 5] End. The output is 6 clusters with each representing a type of cancer.

5. RESULT:

Finally using the significant pattern prediction tools for a cancer prediction system were developed. Table 1 represents the frequent pattern parameters and & their corresponding score & Figure 2 represents the risk level of Cancer which is implemented using Table 1. Fig 3 show the status of cancer patient & fig 4 shows the status of report i.e. output screen . The results are separated into 3 parts. First is the frequent & significant pattern discover. The second is mapping the cancer to cluster & third is prediction giving risk score as output. At the beginning all the input data is stored in the non-cancer cluster further classified & clustered by the model. A single user input data is fed into the system and gets classified according to the significant pattern to which it matches through decision tree,

analyzed for its risk score merged with either one of the Non-cancer & cancer clusters. This gives the result whether the patient has cancer or not.
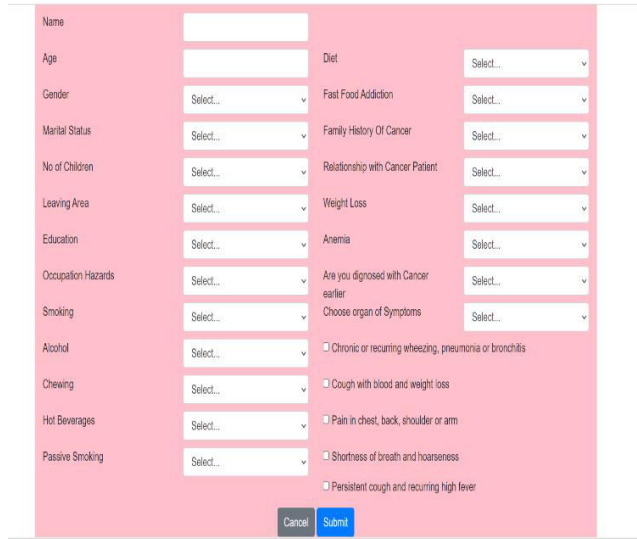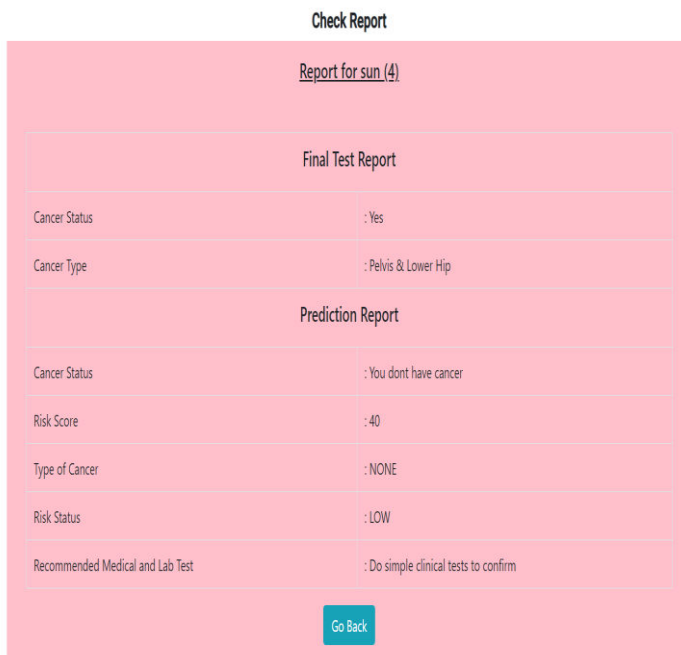

Fig 3: Screen User Input (check cancer status)

Fig.4OutputScreen(prediction)

6. CONCLUSION & FUTURE SCOPE :

In this paper a clustering & decision tree techniques to build a cancer risk prediction system is proposed. Cancer has become the leading cause of death. The most effective way to reduce cancer deaths is to detect as soon as possible . Many people avoid cancer screening due to the cost involved in taking several tests for diagnosis. This system may provide easy & cost effective way for screening cancer & may play a important role in earlier diagnosis process for different types of cancer & provide effective preventive strategy. This prediction system can also be used as a source of record with detailed patient history in hospitals as well as help doctors to concentrate on particular treatment or therapy for any patient. . The main focus is on using different algorithms for cancer prediction using data mining. In future the work may be extended & improved for the automation of breast cancer, lung cancer, blood cancer prediction. We will use the other types of data mining techniques to predict .

6.REFERENCE:

[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] Reeti Yadav "Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques" olume 76-No.10, August 2013 [

International Journal of Computer Applications (0975-8887),   V