International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Data Mining Techniques to Predict Cancer Diseases

Mohd Thousif Ahemad
TSKC Faculty Nagarjuna Govt. College(A) Nalgonda, Telangana, India

ABSTRACT
Cancer is one of the most common diseases in the world that results in majority of death. The most important thing in medical field is early diagnosis of any disease which helps to cure it in early stage and increase the life expectancy chances. Cancer is one of such disease where early diagnosis can reduce the death rate in cancer patients. Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. Therefore a novel multi layered method combining clustering and decision tree techniques to build a cancer risk prediction system is proposed. This research uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming.
Keywords: Diagnosis, Prediction, Data Mining, Clustering

I. INTRODUCTION

Nowadays in all fields like education, science, genetics, agriculture and medicine the amount of data is increasing regularly. Analyzing this huge amount of data to extract the novel and usable information or knowledge is very complicated and time consuming task. Data mining techniques are useful for this matter.
Generally, in the medical world, there are two phases for making the decisions. These two phases are

Differential Diagnosis (DD): In this phase, all information of patients including their medical history, symptoms of disease, results of various testing such as blood testing and etc. are perceived by doctors as the input data. These data are processed by doctors based on their medical knowledge for disease diagnosis. Sometimes several diseases have some similar symptoms, therefore, medical doctors must be assign arbitrary weights to each one of inputs and make patterns, match these patterns with the patterns of various diseases and finally select the closest match and diagnosis the exact disease.

Final or Provisional Diagnosis (FD): In this phase, the preliminary recommendations and treatments would be start according to the identified disease. In this step, a physician with medical knowledge and his/her logic, continues checkups and records the results of continually perceives or tests, and decides the final treatments and prognosis.

Cancer is one of the most common diseases in the world that results in majority of death. Cancer is caused by uncontrolled growth of cells in any of the tissues or parts of the body. In the other words, whenever cells in part of the body divide uncontrollably and damage the other cells, cancer is occurred. Nowadays, more than 100 types of cancers based on the part of body where it's appeared, or cells that are affected, have been classified. Several factors affect the creation or spreading cancers including: gender, age, genetics, marital status, quality of life, living location and etc. Cancer may occur in any part of the body and may spread to several other parts.

The most important thing in medical field is early diagnosis of any disease which helps to cure it in

early stage and increase the life expectancy chances. Cancer is one of such disease where early diagnosis can reduce the death rate in cancer patients. There are generally two types of cancer:- 1) Benign cancer and 2) Malignant Cancer.

If cancer is detected in benign phase life expectancy of a patient increases. Breast cancer is one of the leading cancers for women in developed countries including India. It is the second most common cause of cancer death in women. The malignant tumor develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Hence, cancer on breast tissue is called breast cancer. Worldwide, it is the most common form of cancer in females that is affecting approximately 10% of all women at some stage of their life. Although scientists do not know the exact causes of most breast cancer, they do know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors include such attributes as age, genetic risk and family history.

Nowadays Data Mining is becoming a common tool in healthcare field. Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, 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 association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity.

Identifying of genetic as well as environmental factors is very important in developing novel methods of cancer prevention. However, this is a multi-layered problem. Therefore a cancer risk prediction system is here proposed which is easy, cost effective and time saving.

This study describes the association between cancer incidence pattern and risk levels of various factors by devising a risk prediction system for different types of cancer which helps in prognosis

II. LITERATURE REVIEW

Large numbers of studies have been conducted by the researchers in different contexts on the topic chose for the present study. Here a humble attempt is made to review few of them.

Wei-pin Chang et al. made a comparative study for predicting breast cancers by decision tree, neural network, genetic algorithm and logistic regression. They concerned on 10 variable/attribute for creating breast cancer classification model. These variables were included: Clump thickness, Bland chromatin, Uniformity of cell size, Uniformity of cell shape, Bare nuclei, Normal nucleoli, Marginal adhesion, Mitoses, Single epithelial cell size and class variable with two value (benign/malignant). Their experimental results revealed that, decision tree has lowest prediction accuracy and logistic regression model had higher accuracy rate among these applied techniques for predicting breast cancers. Further, genetic algorithm had highest accuracy in the classification of breast cancers and created acceptable classification rules.

Shrivastava et al. made a review of different classification techniques which have been done for diagnosis of breast cancers. Finally they showed that, Neural Network and decision tree are the most popular techniques which are used by various researchers to create decision rules or predictive models from the breast cancer data.

Sudhir D. Sawarkar et al. have used neural networks and SVM (Support Vector Machine) method for diagnosis of breast cancers and proposed a new

algorithm with implementing SVM by using kernel Adatron algorithm. This algorithm has capability for mapping inputs into a high-dimensional space. Further, it can be isolate inputs and separating data into their respective classes. Their experimental results revealed that, the proposed algorithm has high accuracy on diagnosis and detection of breast cancers. Based on their results, the accuracy of cancer diagnosis by surgeons, radiologists or physicians was nearly 85%, whereas, the accuracy of detections made by their proposed algorithm was 97%.

V.Krishnaiah et al developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees

III. BRIEF SURVEY OF RESEARCH PROBLEM

NEED FOR THE PRESENT INVESTIGATION

It is such a hard job to expect the state of the patient as it is in actuality hard to judge against the facts of the other patient and to guess the consequences. In the detection and prevention of cancer so far we lack in Accuracy in terms of predicting the class label, guessing value of predicted attributes, Speed in building required model, Robustness in correct predictions, Scalability of database size.

STATEMENT OF THE PROBLEM

In the background of the observations made and in the light of the literature review some of the research questions raised for the study are

- To identify the most common symptoms which can help for earlier diagnosis of cancers?
- How can improve the accuracy of diagnosis and decrease the number of Biopsy and error in detecting malignant cancers?
- Is it possible we develop a tool which can be automatically without human Interference diagnosis the cancer with analysing automatically results of mammography and etc.?

HYPOTHESES

For the present study the following hypotheses have been formulated:

- Cancer detection
- Classification of cancer:
- Prediction of clinical outcome of patients after cancer surgery
- Prediction of Survival in types of cancer
- Intelligent data analysis for cancer patient monitoring and home care

METHODOLOGY

Extensive literature reviews, case studies show that there are number of factors influencing cancer. These factors are identified and taken as attributes for this study.

Data Source: Initially cancer and non-cancer patients' data were collected from different diagnostic centres. Data of male and female patients whose age was between 20-70 years old are taken. 20 risk factors were considered for cancer assessment in population, which includes-age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, smoking, food habit, physical activity, obesity, tobacco, genetic Risk, environment, mental trauma, uptake of red meat, balance diet, hypertension, heart disease, excessive alcohol, radiation therapy and chronic lung diseases.

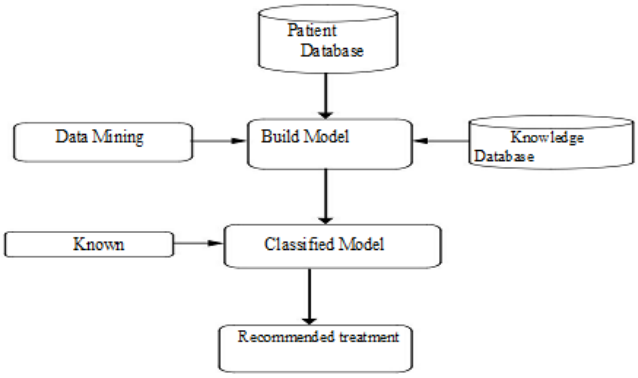
These attributes are used to train and develop the system and a part is used to test the significance of the system. These attributes play an important role in diagnosing cancer in all the cases. This data is stored in a knowledge base which has the ability to expand itself as new data enters the system through front end from which new knowledge is gained and thus the system becomes intelligent.

Classification and Significant Pattern Generation: Decision tree algorithm is used to mine frequent patterns from the data set. The frequent item sets that occur throughout the data base 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. In other words the patterns that are mined by the decision tree are well defined and distinguished to be separated as cancer and non cancer datasets. The following pseudo code is used to generate frequent pattern using decision tree.

Clustering Using K-Means: 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 interclass similarity and minimize the interclass similarity. The weightage scores of the significant patterns mined are fed into K- means clustering algorithm to cluster and divide it into cancer and non cancer groups. The cancer group is further subdivided into six groups with each cluster representing a type of cancer. At the beginning the data is assigned to a non cancer cluster and then based on the intensity of the cancer given by its weightage it is either moved to the cancer cluster or gets retained in the non cancer cluster, further the data object is moved between the subgroups of the hierarchical cancer cluster based on the symptoms the data object contains. To calculate the mean of the cluster center the symptoms 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.

DETAILS OF THE PROPOSED IMPLEMENTATION

The following is the model of the proposed work. The collected data is pre-processed and stored in the knowledge base to build the model. Seventy five percent of the entire data is taken as training set to build the classification and 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 and its corresponding weightage. The clustering model is build using the k-means clustering algorithm. The model is then tested for accuracy, sensitivity and specificity using test data along with merging it to the knowledge base. Finally the model is evaluated using Support Vector Machine.



IV. CONCLUSION

This paper focused on the possibility of building a model to predict cancer in early stage. In the process of cancer control plan and treatment, early detection of cancer is key component. If cases was early detected then treatment more effective and there is greater chances of reduce premature deaths and suffering due to cancer. By applying data mining k-mean algorithm to cluster and divide it into cancer and non cancer groups. Most people avoid screening due to the cost involved in tests for diagnosis . This prediction system may provide easy and cost effective way for screening cancer

V. REFERENCES

[1]. Ferlay J, Shin HR, Bray F, et al (2010). GLOBOCAN 2008: cancer incidence and mortality worldwide: IARC, 10, 220-7.

[2]. 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.

[3]. Amorim R, Mirkin B (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recognition, 45,1061-75. 2

[4]. 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.

[5]. Reeti Yadav "Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques" International Journal of Computer Applications (0975-8887), Volume 76-No.10, August 2013

[6]. In 2013 Divya Tomar et al proposed "Survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013),