ANALYSIS OF CANCER DETECTION SYSTEM USING DATAMINING APPROACH

K.Arutchelvan[*]
Programmer (SS)
Department of Pharmacy
Annamalai University

Dr.R.Periasamy
Associate Professor
Department of Computer Science
Nehru Memorial College

***Abstract** - Data mining techniques are used for variety of applications. In health care industry, data mining plays an important role for predicting diseases. For detecting a disease number of tests should be required from the patient. Data mining is defined as shifting through very large amounts of data for useful information. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction and sequential patterns. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. This technique has an advantages and disadvantages. This research paper analyzes how data mining techniques are used for predicting different types of major life threatening diseases.*

***Keywords** - Cancer, Data Mining, Association rules, Clustering, Classification.*

I. INTRODUCTION

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and data-base systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it.

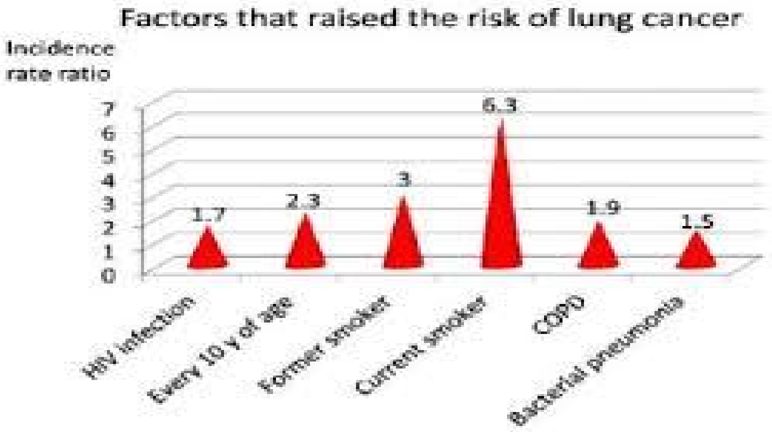
Data mining has been defined as “the nontrivial extraction of previously unknown, implicit and potentially useful information from data”. It is “the science of extracting useful information from large databases”. It is one of the tasks in the process of knowledge discovery from the database. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily under-stood to humans. It is a process to examine large amounts of data routinely collected.

Lung cancer is divided into two main categories: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is further classified into squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Because treatment varies greatly depending on the type and stage of lung cancer, the diagnostic workup is critical in terms of identifying the specific type of lung cancer, the stage of the disease, and the ability of the patient to tolerate treatment. NSCLC represents 80% of all lung cancers, with adenocarcinoma accounting for 40% of all cases of lung cancer. Squamous cell carcinoma occurs most frequently in the central zone of the lung whereas adenocarcinoma tumors are peripheral in origin, arising from the alveolar surface epithelium or bronchial mucosal glands.

Large cell carcinoma composes only 15% of all lung cancers and appears to be decreasing in incidence because of improved diagnostic techniques. The second major type of lung cancer is SCLC, in which there are also several histologic groupings: pure small cell, mixed small cell, and large cell carcinoma, as well as combined small cell. SCLC is usually more aggressive than NSCLC and presents as a central lesion with hilar and mediastinal invasion along with regional adenopathy. Distant metastasis at presentation is common in patients with SCLC. The most common sites of metastasis of lung cancer are the bones, liver, adrenal glands, pericardium, brain, and spinal cord. Staging for NSCLC is done using the internationally accepted TNM (tumor, node, metastasis) staging system Prognosis and treatment of SCLC are determined using a staging system developed by the Veterans Administration Lung Cancer Study Group, although some hospitals and cancer centers are beginning to apply the TNM system.

This educational activity is designed for nurses and other health care professionals who care for and educate patients and their families regarding lung cancer symptoms, pathophysiology and treatment. For those wishing to obtain CNE credit, an evaluation follows. After studying the information presented in this article, the nurse will be able to:

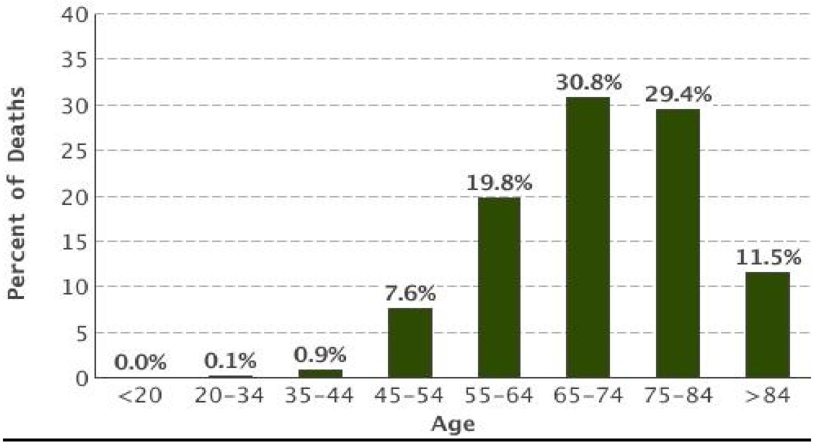
1. *List types of lung cancer.*
2. *Describe treatment options for non small cell lung cancer.*
3. *Discuss treatment options for small cell lung cancer.*



II. PROPOSED SYSTEM

CLASSIFICATION TECHNIQUES:

Classification techniques were used for predicting the treatment cost of healthcare services which was increased with rapid growth every year and was becoming a main concern for everyone.



CLUSTERING:

Clustering is defined as unsupervised learning that occurs by observing only independent variables while supervised learning analyzing both independent and dependent variables. It is different from classification which is a supervised learning method. It has no predefined classes. Because of this reason, clustering may be best used for studies of an exploratory nature, mainly if those studies encompass large amount of data, but not very much known about data.

The goal of clustering is descriptive while goal of classification is predictive. The main task of unsupervised learning method means clustering method is to form the clusters from large database on the basis of similarity measure. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, an important part of the assessment is extrinsic. Clustering partitioned the data points based on the similarity measure.

Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belongs to different groups. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type.

Therefore, it grasp various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological "taxonomies", to medical "syndromes" and genetic "genotypes" to manufacturing "group technology"— the problem is identical: forming categories of entities and assigning individuals to the proper groups within it. Following are the various clustering algorithms used in healthcare.

PARTITIONAL CLUSTERING:

The maximum number of data points in the datasets is 'n'. With the help of 'n' data points the maximum possible number of 'k' clusters is obtained. In order to obtained the 'k' clusters from 'n' data points partitional clustering method is used. In this method, each 'n' data points is relates to one and only 'k' clusters while each 'k' clusters can relates to more than 'n' data points. Partitional clustering algorithms require a user to input k, (which is the number of clusters). Generally, partitional algorithms directly relocate objects to k clusters.

Partitional algorithms are categorized according to how they relocate objects, how they select a cluster centroid (or representative) among objects within a (incomplete) cluster, and how they measure similarities between objects and cluster centroids. Before we obtained the clusters this method requires to define the required number of cluster which we may have to obtained from datasets. On the basis of similarities between objects and cluster centroids this method is partitioned into two categories. These are K-means and K-Mediods. One of the most popular algorithms of this approach is K-means. First of all it randomly selects k objects and then decomposes these objects into k disjoint groups by iteratively relocating objects based on the similarity between the centroids and objects. In k-means, a cluster centriod is mean value of objects in the cluster. The next algorithm is K-mediods.

The major advantage of partitional clustering algorithms is their superior clustering accuracy as compared with hierarchal clustering algorithms that is the result of their global optimization strategy (i.e., the recursive relocations of objects).. Another advantage is, partitional algorithms can handle large data sets which hierarchal algorithms cannot (i.e., better scalability) and can more quickly cluster data. In other words we can say that, partitional algorithms are more effective and efficient than hierarchical algorithms. One major drawback to the use of partitional algorithms is that their clustering results depend on the initial cluster centroids to some degree because the centroids are randomly selected.

III. DATA MINING CHALLENGES IN HEALTHCARE

The healthcare data is very useful in order to extract the meaningful information from it for improving the healthcare services for the patients. To do this quality of data is very important because we cannot extract the meaningful information from that data which have no quality. Hence, the quality of data is another very important challenge. The quality of data depends on various factors such as removal of noisy data, free from missing of data etc. All the necessary steps must be taken in order to maintain the quality in healthcare data.

Data sharing is another major challenge. Neither patients nor healthcare organizations are interested in sharing of their private data. Due to this the epidemic situations may get worse, planning to provide better treatments for a large population may not be possible, and difficulty in the detection of fraud and abuse in healthcare insurance companies etc. Another challenge is that in order to build the data warehouse where all the healthcare organizations within a country share their data is very costly and time consuming process.

IV. CONCLUSION

The privacy regarding to patient's confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. But in situations like epidemic, planning better healthcare services for a very large population etc. some confidential data may be provided to the researchers and government organizations or any authorized organizations.

In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination.

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry.

All the necessary steps are takes in order to minimize the semantic gap in data sharing between distributed healthcare databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services, to better detection of fraud and abuse, improved customer relationship management across the world.

REFERENCES

[1]. Taisia Huckle, and Karl Parker, Long-Term Impact on Alcohol-Involved Crashes of Lowering the Minimum Purchase Age in New Zealand, American Journal of Public Health, Vol. 104, pp.1087-1091,June 2014.

[2]. Messadi M, Ammar M, Cherifi H, Chikh MA and Bessaid A," Interpretable Aide Diagnosis System for Melanoma Recognition", 2014, Bioengineering & Biomedical Science.

[3]. Eduardo López-Caneda, Socorro Rodríguez Holguín, Montserrat Corral, Sonia Doallo, Fernando Cadaveira, Evolution of the binge drinking pattern in college students: Neurophysiological correlates, Alcohol (Elsevier) Vol. 48 , 2014.

[4]. Angelina Pilatti , Juan Carlos Godoy , Silvina Brussino , Ricardo Marcos Pautassi," Underage drinking: Prevalence and risk factors associated with drinking experiences among Argentinean children", Alcohol(Elsevier), Vol. 47, pp. 323-331, 2013.

[5]. Nadia Smaoui, Souhir Bessassi," A developed system for melanoma diagnosis",2013, International Journal of Computer Vision and Signal Processing, 3(1), 10-17(2013).

[6]. Teresa Mendonca, Pedro M. Ferreira,Jorge S. Marques, Andre R. S. Marcal, Jorge Rozeira," PH2 - A dermoscopic image database for research and benchmarking",2013, 35th Annual International Conference of the IEEE EMBS Osaka, Japan, 3 - 7 July, 2013.

[7]. Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto Gonz´alez Osorio," A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection",2013.

[8]. Mahmoud Elgamal ," AUTOMATIC SKIN CANCER IMAGES CLASSIFICATION", 2013, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 3, 2013.

[9]. G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani," An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET),Vol. 2 ,No. 4, pp139-144, August 2013.

[10]. Kawsar Ahmed, Abdullah Al Emran, Tasnuba Jesmin, Roushney Fatima Mukti,Md Zamilur Rahman, Farzana Ahmed," Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, Vol. 14,pp. 595-598,2013.

[11]. V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 4, No. 1 , pp. 39 –45, 2013.

[12]. Shweta Kharya," Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[13]. Naresh Nebhinani**,** Shubh M Singh**,** Gourav Gupta," Demographic and clinical profile of substance abusing women seeking treatment at a de-addiction center in north India," Industrial Psychiatry journal, Vol. 22, 2012.

[14]. Gabriella Fabbrocini, Giovanni Betta, Giuseppe Di Leo, Consolatina Liguori, Alfredo Paolillo,Antonio Pietrosanto, Paolo Sommella, Orsola Rescigno, Sara Cacciapuoti, Francesco Pastore,Valerio De Vita, Ines Mordente and Fabio Ayala," Epiluminescence Image Processing for Melanocytic Skin Lesion Diagnosis Based on 7-Point Check-List", 2010, The Open Dermatology Journal, 2010,Vol 4.

[15]. Suhail M. Odeh," Automatic Diagnosis of Skin Cancer", 2010, Journal of Communication and Computer 8 (2011) 751-755 at David Publishing.

[16]. Radu Dobrescu, Mateidobrescu, Stefan Mocnu, Danpopescu," Medical images classification for skin cancer", 2010, Wseas Transactions on Biology and Biomedicine;ISSN-1109-9518.