A REVIEW PAPER ON DATA MINING CLASSIFICATION TECHNIQUES FOR DETECTION OF LUNG CANCER

Supreet Kaur[1], Amanjot Kaur Grewal[2]

[1]Research Scholar, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

Abstract— *Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. A disease that is commonly misdiagnosed is lung cancer. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as BFO, SVM, LDA and Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease. . In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the lung cancer. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.*

Keywords—Data Mining, classification, KDD, Data Mining methods.

1. Data Mining

There are several applications for Machine Learning (ML), the most significant of which is data mining. Data Mining (The analysis step of the knowledge discovery in data base) a powerful new technology improved and so fast grown. It is a technology used with great potential to help business and companies focus on the most important information of the data that they have to collect to find out their customer's behaviors .Intelligent methods are applied in order to extracting data pattern, by many stages like" data selection, cleaning data integration, transformation and pattern extraction". Many methods are used for extraction data like" Classification, Regression, Clustering, Rule generation, Discovering, association Rule etc. each has its own and different algorithms to attempt to fit a model to the data. The field of data mining developed as a means of extracting information and knowledge from databases to discover patterns or concepts that are not evident.

2. KDD

Data mining (DM), also known as ‘‘knowledge discovery in databases" (KDD), is the process of discovering meaningful patterns in huge databases .The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process.

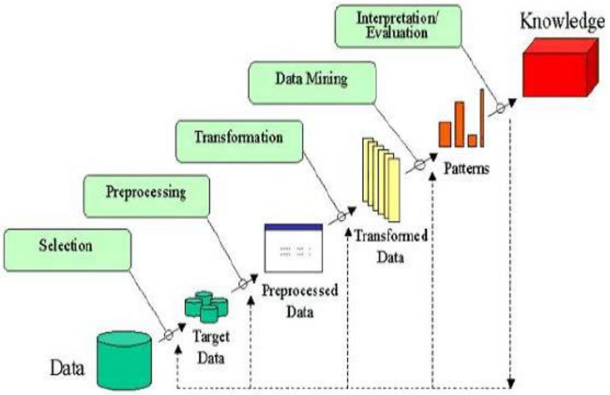

Figure 1: The KDD Process

- Stores data: Where data are stored in any data repository.
- Data integration: Multiple or heterogeneous data sources are integrated into a single unit.
- Data selection: Data retrieve from the database as needed for the KDD method.
- Data transformation: It is a process of data normalization where data are transformed and joined together into a form that is appropriate for mining process. Sub stages of this data transformation are,
- Data cleaning: It handles noisy, erroneous and irrelevant data.
- Improve data: Improve quality of data by adding new information, missing values to available data.
- Generalizing data: Applying operations on data in order to prepare for a machine learning approach.
- Data mining: This is a very important step in mining process. Here intelligent methods are applied in order to extracts data patterns.
- Pattern evaluation: This process is to identify the truly interesting patterns that are presented in knowledgebase.
- Knowledge presentation: It is final reporting of KDD process. Where visualization and knowledge representation techniques are used to represent the mined knowledge to the user.

3. DATA MINING CLASSIFICATION METHODS

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for this review are of classification category as the applied prediction techniques assign patients to either a ”benign” group that is non-cancerous or a ”malignant” group that is cancerous and generate rules for the same. Hence, the Lung cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups [4].

3.1. Support Vector Machine (SVM)

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

3.2. Bacterial Foraging Optimization

The bacterial foraging optimization algorithm proposed by Passion is an optimization technique which derives its idea from foraging behavior of Bacteria *E. coli*. The *E. coli* bacteria grow at a very fast rate if given suitable condition and sufficient food to grow. The *E.coli* bacteria moves very fast into nutrient areas and tries to go away from noxious substances. The motions of bacteria are known as taxes. In

Foraging theory, the objective is to search for and obtain nutrients in a fashion that energy intake per unit time (E/T) is minimized. An *E.coli* bacterium has 8-10 flagella placed randomly on its body with a speed of 100-200 rps. Its movement and behavior is characterized by the spinning flagella which acts as a Biological motor and helps bacteria to swim. The bacterial foraging process consists mainly of four sequential mechanisms, namely chemotaxis, swarming, reproduction and elimination dispersal

3.3 Neural Networks

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input

4. Data Mining Classification Techniques for lung Cancer detection

Yongqian Qiang, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, &Duwu Cuic [1] conducted clinical and imaging diagnostic rules of peripheral lung cancer by data mining technique, and to explore new ideas in the diagnosis of peripheral lung cancer, and to obtain early-stage technology and knowledge support of computer-aided detecting (CAD). The data were imported into the database after the standardization of the clinical and CT findings attributes were identified. The diagnosis rules for peripheral lung cancer with three data mining technology is same as clinical diagnostic rules, and these rules also can be used to build the knowledge base of expert system. The demonstrated the potential values of data mining technology in clinical imaging diagnosis and differential diagnosis.

Cheng-Mei Chen and Chien-Yeh Hsu [2] established a survival prediction model for liver cancer using data mining technology. They collected the data from the cancer patient's registration database of a medical center in Northern Taiwan between the years 2004 and 2008. A total of 227 patients were newly diagnosed with liver cancer during this time. They extracted nine variables pertaining to liver cancer survival were analyzed using ttest and chi-square test through literature survey and expert consultation. Six variables showed significant. Artificial neural network (ANN) and classification and regression tree (CART) were adopted as prediction models.

Dai and Xu [3] proposed dimension reduction method with respect to fuzzy gain ratio based on fuzzy rough set theory. Three data sets for real world tumors in gene expression were used. The paper proved the efficiency of their model with classification accuracy.

Thangaraju P, Karthikeyan T, Barkavi G [4] conducted smoking is the biggest risk factor of lung cancer. The more years and larger number of cigarettes smoked the greater the risk of developing lung cancer. The average age of someone diagnosed with lung cancer is 65 to 70 years old, but people who are younger can develop lung cancer. Young adults who have never smoked also can develop lung cancer.

Jennifer et.al [5] proposed a system that utilizes gene expression data from oligonucleotide microarrays to predict the presence or absence of lung cancer, predict the specific type of lung cancer should it be present, and determine marker genes that are attributable to the specific kind of the disease. The proposed system would help in the faster diagnosis and serve as a reliable adjunct approach to current lung cancer classification methods.

Julliet et.al [6] presented that earlier researches and case studies indicate that the survival rate of the patients suffering from cancer is higher when the disease is diagnosed at an early stage. Lung cancer, a disease highly dependent on historical data for early diagnosis, has influenced researchers to pursue the data mining techniques for the pre-diagnosis process. The five year survival rate increases to 70% with the early detection at stage 1, when the tumour has not yet spread. Existing medical techniques like X-Ray, Computed Tomography (CT) scan, sputum cytology analysis and other imaging techniques not only require complex equipment and high cost but is also proven to be efficient only in stage 4, when the tumors has metastasized to other parts of the body.

Wenyan et.al [7] presented that Support Vector Machine (SVM), Random Forests algorithm (RF) and Fisher discriminate model are good methods for auxiliary diagnosis efficiency for lung cancer with the models. The diagnosis indexes of the SVM and RF algorithm are higher than Fisher discriminate analysis, and it can be thought that they are judging the optimal classification model of lung cancer. Compared with the healthy people, the results show that the study on diagnosis of the lung cancer by SERS on data mining can be a new type of the lung cancer diagnosis tool.

Iyrena et.al [8], investigates application of novel Bidirectional Data Partitioning Technique (BDP) to cancer survival analysis. Author has developed this technique for classification problems with unstable feature relevance and SEER Cancer Data illustrates this machine learning concept. BDP is applied for survival analysis in order to find groups of patients with different key factors that determine survival time.

5. METHODOLGY

To start with, different classification algorithms are chosen from data mining and implemented in a programming language. MATLAB will be the implementing language here. Each algorithm will be tested with lung cancer data as input data. I am applying different- different classification algorithms as well as optimization algorithms.

6. Conclusion

The presented discussion on knowledge extraction from medical databases is merely a short summary of the ongoing efforts in this area. It does, however, point to interesting directions of our research, where the aim is to apply hybrid classification schemes and create data mining tools well suited to the crucial demands of medical diagnostic systems. It is proposed to develop a substantial set of techniques for computational treatment of these data. The approaches in review are diverse in data mining methods and user interfaces and also demonstrate that the field and its tools are ready to be fully exploited in biomedical research.

References

[1] YonqianQia,ng, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, & DuwuCuic, “The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique”, Journal of Nanjing Medical University, Vol. 21(3), pp. 190-195.

[2] Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen andChien-Yeh Hsu,” Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees”, Seventh International Conference on Natural Computation 2011.

[3] Jianhua Dai, Qing Xu," Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification", Applied Soft Computing, Vol.13(2013), pp. 211–221.

[4] Thangaraju P, Karthikeyan T, BarkaviG, ”Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014.

[5] J. Cabrera, A. Dionisio and G. Solano, "Lung cancer classification tool using microarray data and support vector machines," Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, Corfu, 2015, pp. 1-6.

[6] J. R. Rajan and C. C. Chelvan, "A survey on mining techniques for early lung cancer diagnoses," Green Computing,

Communication and Conservation of Energy (ICGCE), 2013 International Conference on, Chennai, 2013, pp. 918-922.

[7] W. Liu et al., "Data mining methods of lung cancer diagnosis by saliva tests using surface enhanced Raman spectroscopy," 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, 2014, pp. 623-627.

[8] I. Skrypnyk, "Finding Survival Groups in SEER Lung Cancer Data," Machine Learning and Applications (ICMLA), 2012 11th International Conference on, Boca Raton, FL, 2012, pp. 545-550.

[9] Shelly Gupta, Harminder Kumar, Anand Sharma, "Data Mining Classification Techniques applied for breast cancer diagnosis and prognosis", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 2, Apr-May 2011.