Journal of Advanced Computing and Communication Technologies ***(ISSN:* 2347 - 2804*)***
Volume No.5 Issue No.3, June 2017

A comprehensive study of Data Mining techniques in prediction of Cancer

By

Harneet Kaur
USICT, Guru Gobind Singh Indraprastha University, New Delhi, India
harneetkaur98738@gmail.com

ABSTRACT

Cancer is one of the most common and fatal disease that results in death of large number of people across the world. Various Data Mining techniques are proven to be helpful in diagnosis of this fatal disease. In this report, an overview of various Data Mining techniques and Data Mining tools like WEKA, Orange, Tanagra, RapidMiner and KNIME is given. It also covers survey of various methodologies and techniques employed by different researchers in context of mining. An implementation on how Data Visualization and Classification can be done using Tanagra tool was shown. Lastly, challenges that are faced in the healthcare industries are discussed.

Keywords

Data Visualization, Benign Cancer, Malignant Cancer, and classification.

1. INTRODUCTION

1.1 Brief introduction

Cancer is one of the most common and fatal disease that results in death of large number of people across the world. It is a group of diseases in which there is an abnormal growth of cells in any tissue having the potential to invade or spread to other parts of the body. There are more than hundred types of cancer, including breast cancer, skin cancer, lung cancer, colon cancer, prostate cancer, leukemia and more. Types of Cancer found can be grouped into two categories:

- Malignant Cancer

Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss and a change in bowel movements. Most common causes that lead to cancer includes use of tobacco, obesity, poor diet, lack of physical activity, and excessive drinking of alcohol. Other factors include certain infections, exposure to ionizing radiation and environmental pollutants. Early detection and prevention of cancer plays an important role in reducing number of deaths caused due to it. Benign tumors do not spread to other parts of the body so if cancer is detected in benign phase, life expectancy of a patient can increase. Only early detection of cancer at the benign stage and prevention from spreading it to other parts of body in malignant stage can save the life of a person. Identification of genetic and environmental factors is an important in developing methods to detect and prevent cancer. Treatment of cancer may include chemotherapy, radiation, and/or surgery. Diagnosis of cancer can be done with the help of Data Mining Techniques.

1.2 Overview of Data Mining

Data mining is defined as a process of extracting previously unknown, implicit and useful information from the data and discovering unknown data, patterns, relationships and knowledge. Knowledge must be new and useful. Data mining is an important step of Knowledge Discovery in Database (KDD) process which is an iterative process of data cleaning, data integration, data selection, pattern recognition and knowledge recognition. It is a broad area that

involves integration of techniques from numerous fields like machine learning, statistics, pattern recognition, artificial intelligence, and database system for analysis of large volume of data. Besides the raw analysis step, it involves database management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining encompasses association, classification, clustering, statistical analysis and prediction. It has been widely used in the areas of communication, credit assessment, stock market prediction, marketing, banking, education, health and medicine, hazard forecasting, knowledge acquisition, scientific discovery, fraud detection but holds significant presence in every field of medical for the diagnosis of several diseases like diabetes, skin cancer, lung cancer, breast cancer, heart disease, kidney failure, kidney stone, liver disorder and hepatitis. There exist various interactive and scalable data mining methods to find latest patterns in the healthcare industry. It is used for analysis of data for making better policies, prevention of various errors in hospitals, detection of fraudulent insurance claims, early detection and prevention of various diseases, reducing costs and saving more lives by reducing death rates.

1.2.1 Components of Data Mining System Architecture

Components of Data Mining System architecture are:

- *Data Source or Database:* This includes data from Data Warehouse, flat files, Relational database, transactional database, spatial database, multimedia database, World Wide Web and more. In database, data needs to be cleansed, updated and modified before it is moved into Data Warehouse.

- *Database Server:* It fetches instructions from given database as per the request from the customer.

- *Data Mining Engine:* It is the most important component of Data Mining System and contains modules or algorithms to perform mining tasks such as Classification, Association Rule and Clustering.

- *Local Model Generator:* This component is used to generate a local model that consists of patterns and modules related to customer's query.

- *Final Model Generator:* Depending on implementation method used for generation of local model, it is used for interaction among customers and system by specifying data mining query or task. It leads to interpretation of data mining results evaluated to customers. Various visualization and GUI strategies are used in this step.

2. LITERATURE REVIEW

Jaimini Majali et al [1] recognised the need for early detection and cure of cancer and aimed to develop a system which would assist doctors in diagnosis decisions. Their system was based on Data Mining techniques like Association Rule Mining (ARM) and Classification Technique for prognosis and diagnosis of Cancer focussing mainly on Breast Cancer. They applied Frequent Pattern algorithm under Association Rule Mining (ARM) and ID3 algorithm in Decision Tree for classification on Wisconsin data set attributes. FP growth algorithm scanned whole data for various values of support and confidence and then mapped frequently found patterns to generate rules that indicated the general behaviour and range of values for malignant and benign tumor. A questionnaire was prepared to identify nature of user which was then applied to ID3 algorithm as input to classify whether user has possibility of cancer or not. They also compared range of values for malignant and benign tumor.

K.Arutchelvan et al [2] showed various Data Mining techniques used to predict different types of lung cancer. Lung cancer listed was divided into two main categories: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC) where NSCLC can further be classified into squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Various factors that raised the risk of lung

cancer were HIV infection, smoking, bacterial pneumonia. They used classification technique to predict the cost of treatment of various healthcare services. They discussed about clustering and partition clustering techniques for prediction. Various Data Mining challenges in the field of healthcare were discussed. They suggested using various data mining techniques in combination to improve survivability rate regarding serious death related problem and a better information system to achieve higher quality medical data.

P.Ramachandran et al [3] proposed a method that combines clustering and decision tree techniques to build cancer prediction system which predicts lung, breast, oral, cervix, and blood and stomach cancer. Data mining techniques such as classification, clustering and prediction were used to identify potential patients of cancer. They separated the results into three parts: First was frequent and sequential pattern discovery. Second was mapping of cancer to its cluster and third was prediction by giving risk score as output. They used Decision Tree algorithm under classification technique to mine frequent patterns from data set and K-means algorithm under clustering. Their prediction system provided an easy and cost effective way for screening of cancer which could help in early diagnosis process for different types of cancer and provided an effective strategy for prevention.

Ankit Agrawal et al [4] analysed the lung cancer data made available from SEER program with an aim to identify hotspots using Association Rule Mining (ARM) techniques where patient survival time was higher than and lower than the average survival time across the entire dataset. Such an analysis would help to identify factors that affected survival and aid doctors to avoid conditions that reduced survival time amongst patients.

3. TECHNIQUES USED

3.1 Association

Association (or relation) is a data mining technique where a simple correlation between two or more items is done which are of the same type in order to identify patterns. For example, while tracking people's buying habits, it may be identified that a customer who buys strawberries also purchase cream along with it, and therefore suggests that the next time that they buy strawberries they might also want to buy cream so both items are kept in close proximity of each other. Rules used to implement Association techniques are known as Association Rules. They are represented as: "If Then" rules and are used to depict relationship among various data items. To implement these rules, we use one of Data Mining algorithm called APRIORI algorithm.

3.2 Classification

Classification is used to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types by identifying different attributes like number of seats, car shape, driven wheels or customers by classifying them by age and social group. Classification can be used as an input to other techniques. Data Classification process includes two steps:

- *Building the Classifier or Model*- In this step, the classification algorithms build the classifier. This step is known as learning step or the learning phase. Classifier is built from the training set which is made up of database tuples and their associated class labels. Each tuple constituting the training set is referred to as a category or class. These tuples are also known as sample, object or data points.

- *Using Classifier for Classification*- In this step, the classifier built is used for classification. Test data is used to estimate the accuracy of classification rules which can then be applied to the new data tuples if the accuracy is considered acceptable. One of the basic Data Classification techniques is the use of Decision Tree.

3.3 Clustering

Clustering is a process of grouping abstract objects into classes of similar objects. A cluster of similar data objects is then treated as one group or one cluster. While performing cluster analysis, first we partition the dataset into different groups based on data similarity and then labels are assigned to the

groups. We can examine one or more attributes or classes and then group together individual pieces of data to form a structure opinion. Cluster analysis is broadly used in many applications such as market research, pattern recognition, data analysis and image processing. As a data mining function, it serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. There are various clustering methods like Partitioning Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method and Constraint-based Method.

3.4 Prediction

Prediction means analyzing trends, classification, pattern matching, and relation. By making an analysis of past events or instances, a prediction about an event can be done. It is generally used in combination with other Data Mining techniques. For example, Decision tree analysis of individual past transactions can be combined with classification and historical patterns in the case of credit card authorization to identify fraudulent transactions.

3.5 Sequential patterns

Sequential Patterns means the discovering frequently occurring ordered events or sub-sequences as patterns. They are a useful method for identifying trends or regular occurrences of similar events. They are generally used over long term data applications. For example, with the help of customer data we can identify that customers buy a particular collection of products together at different times of the year and we can use this information in a shopping basket application to automatically suggest that certain items can be added to a basket based on their frequency and past purchasing history.

4. TOOLS

4.1 RapidMiner

RapidMiner [5] is open source data mining software written in Java Language. It has template-based framework that lets the user do data analysis easily. It is a ready-made tool and requires no coding. It provides functionalities like data pre-processing, visualization, predictive analytics and statistical modelling, evaluation, and deployment. This tool can be integrated with tools like WEKA and R to directly give models from scripts written in the both of them. It helps to predict the future outcomes using various data mining and machine learning algorithms.

4.2 WEKA

Waikato Environment for Knowledge Analysis (WEKA) [6] is a suite of machine learning software developed at the University of Waikato, New Zealand. This is a Java-based customization tool and is free to use. It supports various standard data mining tasks including data pre-processing, clustering, classification, regression, visualization and feature selection. It provides access to SQL databases using Java Database Connectivity (JDBC) and can process the result returned by a database query.

4.3 KNIME

KNIME, the Konstanz Information Miner [7] has all the data mining tools that you require for data extraction, pre-processing, transformation and loading. It has a graphical user interface which helps users to easily connect the nodes for data processing. It combines various components for data mining and machine learning. KNIME has been extensively used in pharmaceutical research. It is also helpful in business intelligence and financial data analysis. KNIME can easily be extended by adding plugins.

4.4 Orange

Orange [8] is a Python-based powerful and open source tool used both by novice and experts. Its properties include visual programming and Python scripting. It is used for data analytics and includes components for machine learning, add-ons for bioinformatics and text mining.

4.5 Tanagra

Tanagra [9] is a free, open source, user friendly software developed for students and researchers for Data Mining. It proposes several data mining methods including exploratory data analysis, statistical learning, machine learning and database areas. It is programmed using Pascal language. It includes various components like Data visualization, statistics, clustering, association, factorial Analysis

and many more. It also includes basic clustering algorithm like K-Means, EM-Clustering. With the help of Tanagra, we can visualize our data. Data Visualization includes Viewing Dataset, plotting values on graphs and scatter plot. It shows relationship among attributes in 2D axes. It includes basic clustering algorithm like K-Means, EM-Clustering.

5. DATA VISUALIZATION & CLASSIFICATION USING TANAGRA

Dataset Considered here is Wisconsin Cancer Dataset [10] as shown in the table:

Table 1: Dataset showing different attributes

Attribute Name	Description	Category	Range Values
CT	Clump Thickness	Continuos	1-10
UCSIZE	Uniformity of Cell Size	Continuos	1-10
UCSHAPE	Uniformity of Cell Shape	Continuos	1-10
MAF	Marginal Adhesion Fibrous	Continuos	1-10
ECSIZE	Epithelial Cell Size	Continuos	1-10
BN	Bare Nuclei	Continuos	1-10
BC	Bland Chromatin: Evaluates the presence of bare bodies	Continuos	1-10
NN	Normal Nuclei	Continuos	1-10
MITOSIS	Cell Growth	Continuos	1-10
DIAG_CLASS	Diagnosis of Tumor	Discrete	2 values

5.1 Importing and viewing data in TANAGRA

To import and view data in Tanagra choose select the text file containing the dataset you want to explore by clicking on icon next to dataset tab. Choose file "Cancer.txt" and validate by clicking the OK. Following screen appears that show all the attributes along with the category they belong to as Discrete or Continuous.

5.2 Adding an operator to the diagram in order to visualize data

Add a "View dataset" component to the diagram by clicking on the "Data Visualization" tab of the components palette and then click on the "View dataset" node to select it and right-click on it to choose the "View" command. Data is displayed in the right frame.

5.3 Getting descriptive statistics

Basic statistics on each attribute includes min, max, average and standard deviation. Add a "Define Status" component under "Feature Selection" tab to the diagram and then choose some continuous variables and add a "Univariate continuous stats" component under "Descriptive Stats tab".

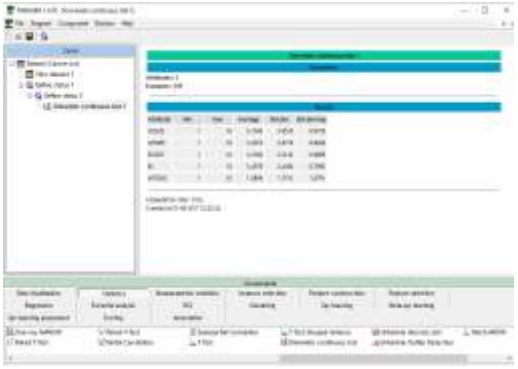

Figure 1: Descriptive stats on continuous variables

Add another Define status component to the "Dataset" node, and select the discrete attribute "Diag_Class". Add a Univariate discrete stats operator. Below is the result:

Figure 2: Descriptive stats on discrete variables

5.3 Statistics for each sub-population

Add "Define status" operator to the "Dataset" node and select some continuous attributes as Input, and the discrete attribute as Target. Add a "Group characterization operator" to view the statistics for each sub-population like to see number of women having benign or malignant cancer. Results are as shown:

Figure 3: Result of Group Characterization

Inspecting these results, it is noticed that on an average women with benign tumor present smaller values of "mitosis" (1.06 versus 1.59 for the complete population). On the other hand, the values of "ushape" attribute are still on an average higher for women with malignant tumor (6.56 versus 3.21).

5.4 Classification

- *Select the learning algorithm*- To use the Classification and Regression Tree (Breiman et al.) algorithm, there are two steps to define a supervised learning process in TANAGRA: Insert a "meta-supervised learning" component from the "Meta SPV Learning" tab and Embedding the learning algorithm- "C-RT" from the "SPV learning tab".

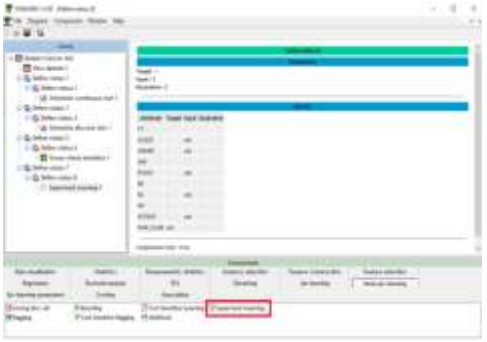

Figure 4: Applying C-RT algorithm for classification

- *Displaying the results*- In order to view the decision tree, click on the "View" menu of the last component and see tree 1. The tree has 3 leaves (3 rules).

Figure 5: Result of Classification

Figure 6: Decision Tree showing three rules

- *Cross-validation* - To compute the error rate with a cross-validation resampling method, add the "Cross-Validation" component from the "SPV learning assessment" tab and set the number of folds to 10, and the number of repetition to 1. The estimated error rate is 6.81%

Figure 7: Results of cross-validation technique

6. DATA MINING CHALLENGES IN HEALTHCARE

- Quality of data is a major challenge since meaningful information cannot be extracted from the data that has low quality. Meaningful information of data is necessary to improve the healthcare services. Quality of data depends on number of factors like removal of noisy data, missing data treatment, removal of outliers and more.

- Data sharing is another major challenge since neither patients nor the healthcare organizations are interested in sharing their private data. Due to this, the epidemic situations get worse, providing better treatment for a large population may not be possible and difficulty in the detection of fraud and abuse in healthcare insurance companies cannot be resolved.

- To build a Data Warehouse where data from all organisations could be stored and shared is a costly and time consuming process.

- It is possible that information collected and stored may not be available in a consistent format. Stored information becomes less useful if they are not available in easily apprehensible format.

- Healthcare associations in Data Mining require the use of big investment resources like time, effort and money. Data mining project can fail due to various reasons like lack of managerial support, inadequate DM expertise and more.

7. CONCLUSION & FUTURE SCOPE

In this report, an overview of Cancer, Data Mining and how Data Mining plays an important role in diagnosis and prognosis of Cancer was given. A study about various methodologies used by different authors in this field was done and an overview of their approach followed was given. Some of the Data Mining tools and techniques that are used were reviewed. An implementation on how Data Visualization and Classification can be done using Tanagra tool was shown. Data Mining can be used in numerous fields besides being used in healthcare

industry like Market Analysis and Management where it can be used for Customer Profiling, Identifying Customer Requirements, Cross Market Analysis , Target Marketing, Determining Customer purchasing pattern; Corporate Analysis & Risk Management where it is used for Finance Planning and Asset Evaluation, Resource Planning; Fraud Detection. Also, it can be used in areas like production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

References

1. Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, Omkar Tadakhe, "Data Mining Techniques for Diagnosis and Prognosis of Cancer", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2015.

2. K.Arutchelvan, Dr.R.Periasamy , "Analysis of Cancer Detection System using Data mining approach", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Issue 11, Volume 2 ,November 2015.

3. P.Ramachandran, N.Girija, T.Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", IJCA (0975- 8887), Volume 97– No.13, July 2014.

4. Ankit Agrawal, Alok Choudhary, "Identifying HotSpots in Lung Cancer Data Using Association Rule Mining", 11th annual Data Mining Workshops (ICDMW), IEEE, 11-11 Dec. 2011.

5. https://rapidminer.com/

6. http://www.cs.waikato.ac.nz/ml/weka/

7. https://www.knime.org/

8. https://orange.biolab.si/

9. https://eric.univlyon2.fr/ricco/tanagra/en/tanagra.html

10. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names.