Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining Techniques in Multiple Cancer Prediction

Dr. A. R. Pon Periasamy
Associate Professor of Computer Science
Nehru Memorial College
Puthanampatti, Trichy (DT)
Tamilnadu, India

K. Arutchelvan
Assistant Professor / Programmer
Department of Pharmacy
Annamalai University, Chidamparam
Tamilnadu, India

Abstract— Data mining intends to endow with a systematic survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research. Discussion is made to enable the disease diagnosis and the breakthrough of hidden healthcare patterns from related databases is offered. Also, the use of data mining to discover such relationships as those between health conditions and a disease is presented. It further discusses about the tools that can be used for the processing and classification of data. This paper summarizes various technical articles on medical diagnosis and prognosis. Defective software modules cause software failures, increase development and maintenance costs, and decrease customer satisfaction. It strives to improve software quality and testing efficiency by constructing predictive models from code attributes to enable a timely identification of fault-prone modules. The main objective of paper is to help developers identify defects based on existing software metrics using data mining techniques and thereby improve the software quality. In this paper, we will discuss data mining techniques that are association mining, classification and clustering for lung cancer prediction.

Key Words: Lung cancer, data mining, clustering, classification and association rule mining.

I. INTRODUCTION

The quick growth in data and databases has created a pressing need for new tools and techniques that can quickly and efficiently process raw data into useable information. In the last years the development of information technology has motivated a parallel growing of facilities to store and manage database. The largest amount of stored data is more important for the demand of extracting the implicit information they contained to aid the decision-making in business, health care services, research…etc. Thus, to obtain useful knowledge from the data stored in large repositories (i.e. the "knowledge discovery"), which is recognized as a basic necessity in many area. Since nineties of the last century, the research area named "data mining" has become central topic in databases and Artificial Intelligence. The task of discovering association rules was introduced by [Agrawal, Imieliński and Swami 1993]. In its original form, the task was defined for a special kind of data, often called basket data, where a tuple consists of a set of binary attributes called items. Each tuple corresponds to a customer transaction, where a given item has a value of true or false, depending on whether or not the corresponding customer bought the item in that transaction. This kind of data is usually collected through bar-code technology; the typical example is a supermarket scanner.

There are various algorithms have been proposed to discover the frequent item sets (Sarawagi, Thomas and Agrawal 1998), (Agrawal, Imieliński and Swami1993). The Apriori algorithm is one of the most popular algorithms in the mining of association rules in a centralized database, which will explained broadly later.

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large database in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation (Tan P., Steinbach M. and Kumar V. 2006). Data mining is a method of extracting what is useable within a database and separating it out from what is unusable. Such methods are necessary because, as human being, we lack the capacity to sort and organize such large volumes of data.

II. DATA MINING

Data mining (DM) is the process of discovering meaningful correlation, patterns, and trends by sifting through large data, using recognition technologies. DM emphasizes on making and testing algorithms that can assist the process of classification, prediction, and pattern recognition. This process uses computer models obtained from existing data (previous data) with limited human interaction. The idea is to increase accuracy and reduce human biases by using automatic pre-programmed methods. As a result, a solid and reliable functional data mining algorithms can be developed to classify objects or predict new cases of diseases.

In **[Frawley, 2012]** describes DM as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In **[Baxt, 1990]** defines DM as the process of automating information that has been discovered. In **[Moxon, 2012]** states that data mining is the process of discovering meaningful new correlation, patterns and trends.

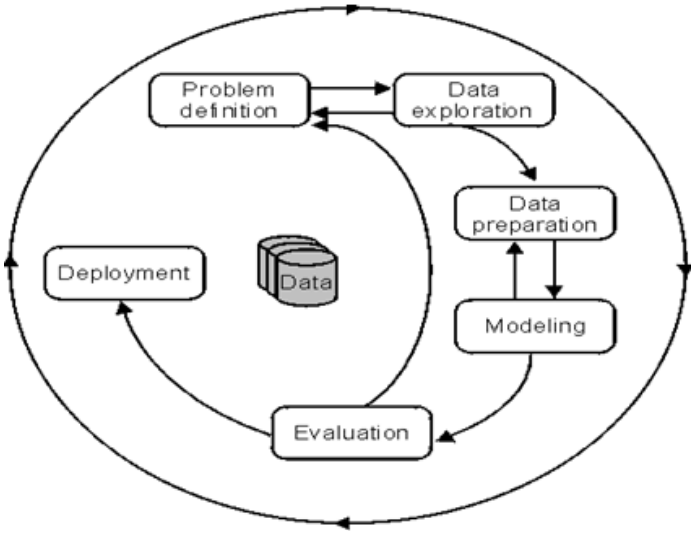

Figure 1.1 Data Mining Process Representation

By sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. In **[Han and Kamber, 2012]** argues that DM techniques can be considered to be predictive.

2.1 Association Rule Mining

One of the main and important topic of data mining is Association Rule Mining. Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amount of data continuously being collected and stored in databases, many industries are becoming interested in mining association rules from their databases. For example the discovery of interesting association relationships among huge amounts of business transaction records can help catalogue design, cross-marketing, loss leader analysis, and other business decision making process (Al-hamami Alaa 2008).

An association rule is an expression X ⇒ Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of such a rule might be that 70% of customers who purchase bread also purchase butter. Mining association rules is a widely used in data mining technique. Association analysis is an unsupervised form of data mining that looks for links between records in a data set. Association analysis is referred to the most common application as market basket analysis.

2.2 Classification

Classification is the process of learning the target function that maps between a set of features (inputs) and a predefined class labels (output) i.e. it puts data in single groups that belongs to a common class, inferring the defining characteristics of a certain group done by Regression algorithms which attempt to map input to domain values. For instance, a regression can forecast certain goods sales by considering the goods features. At the same time, classifiers can map the input space into pre-defined classes. Consequently, a classifier can predict a new case of patient whether benign (healthy) or malignant (suffer from a certain disease).

Kotsiantis et al, 2007, describes supervised ML as the search for algorithms that reason from externally supplied instances to produce general hypotheses; the general hypotheses are then used to make predictions about future instances. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, and the value class label is unknown.

The input data for the classification is a set of instances. Each instance is a record of data in the form of (,y) where is the features set and is the target variable (class label). Classification model is a tool that is used to describe data (Descriptive Model) or a tool to predict the target variable for a new instance (Predictive Model). The decision tree, artificial neural network, Naïve Bayes, and k-nearest neighbour's classifier are some of the examples of classification models.

The training data consists of instances whose class labels are known. The classification model can be built based on the training data. The model then can be evaluated and tested by using the testing data which contains records with missing class labels. The evaluation of model performance is based on the number of testing instances that are correctly forecasted. The result of performing the model on the testing data produces the confusion matrix.

2.3 Decision Tree (DT)

DT is a classification method which contains nodes, branches, and leafs. The first node on the tree or the top node is called the root node. Each node in the tree is connected with one or more nodes using branches, the last node in the tree that contains no outgoing branches is called leaf node. The leaf node indicates to termination or the outcome value. The terminology of such classification method is to keep asking question until conclusion is reached. The set of questions and answers could form a decision tree with set of nodes: first, root node having a zero or more outgoing nodes

and no incoming nodes, as well as containing the testing condition that separate the records; second, Normal nodes, those nodes are internal nodes and each has one and only one incoming node and two or more outgoing edges. It also contains the testing condition that separate records and thirdly, Leaf nodes, those nodes hold the class labels, have no outgoing edges, and only one incoming edge.

2.4 Machine Learning (ML)

ML is a scientific discipline responsible for recognizing complex patterns and making intelligent decisions based on data. Emphasizing on making and testing algorithms, ML can assist the process of classification, prediction, and pattern recognition using computer models. ML provides limited human involvement and uses the automatic pre-programmed methods that reduce human biases. The process of proposing the algorithm and its functionality to classify objects or predict new cases are to be built on solid and reliable data, Mitchell, 1997. The database contains a collection of instances (records or case). Each instance used by ML algorithms is formatted using same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning. Whilst the process of ML without knowing the class label of instances is called unsupervised learning.(Ozgür, 2004), clustering is a common unsupervised learning method. The goal of clustering is to describe data. However, classification and regression are predictive methods. This research will focus on supervised machine learning.

III. BACKGROUND OF WORK

In context of software engineering, software quality refers to software functional quality and software structural quality. Software functional quality reflects functional requirements whereas structural quality highlights non-functional requirements. Software quality measurement [1] is about quantifying to what extent a system or software possesses desirable characteristics namely Reliability, Efficiency, Security, Maintainability and (adequate) Size. This can be performed through qualitative or quantitative means or a mix of both. In both cases, for each desirable characteristic, there are a set of measurable attributes like Application Architecture Standards, Coding Practices, Complexity, Documentation, Portability and Technical & Functional Volumes. The existence of these attributes in a piece of software or system tends to be correlated and associated with this characteristic. A software defect is an error, flaw, mistake, failure, or fault in a computer program or system that produces incorrect or unexpected results, or causes it to behave in unintended way. Software defect prediction is the process of locating defective modules in software. It helps to improve software quality and testing efficiency by constructing redictive models from code attributes to enable a timely identification of fault-prone modules, it also helps us in planning, monitoring and control and predict defect density and to better understand and control the software quality. The Software Defect Prediction result, that is the number of defects remaining in a software system, it can be used as an important measure for the software developer, and can be used to control the software process [2].

In this paper, we will discuss Data mining techniques for software defect prediction. Data mining is a process of analysing data from different perspectives and summarizing it into useful information. It allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data. It is viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large databases. The techniques of data mining for software defect prediction are: clustering, association mining, and classification. In rest of the paper Section 2 presents the related work on the topic, Section 3 presents the data mining techniques for Defect Prediction model and at last Section 4 presents the conclusion and future work.

In [3] Q. Song, et. al. presented an application of association rule mining to predict software defect associations and defect correction effort with SEL defect data. This is important in order to help developers detect software defects and project managers improve software control and allocate their testing resources effectively. The objective of the study is to discover software defect associations from historical software engineering data sets, and help determine heather or not a defect is accompanied by other defects. If so, we attempt to determine what these defects are and how much effort might be expected to be used when we correct them.

In [4], they applied K-Means and Neural-Gas techniques on different real data sets and then the representative module of the cluster and several statistical data are explored in order to label each cluster as fault-prone or not fault-prone. In their study they have presented comparative results performed on same data sets. They have applied unsupervised learning approach for fault prediction in software module. The false negative rates (FNR) for the clustering-based approach are less than that for metrics-based approach, while the false positive rates (FPR) are better for the metrics-based approach The overall error rates for both approaches remain the same.

In [5], predictive models are estimated based on various code attributes to assess the likelihood of software modules containing errors. Many classification methods have been suggested to accomplish this task. In this paper, they assess the use of classification method, CBA2, and compare it to other rule based classification methods. They have investigated the performance of an association rule based classification method for software defect prediction problems Data experiments were conducted to compare the CBA2 classifier with two other rule/tree based classifiers showing

IV. CONCLUSION

In this paper, we have discussed that how data mining techniques are used for software defect prediction. In order to improve the efficiency and quality of software development, we can make use of the advantage of data mining to analysis and predict large number of defect data collected in the software development. We have also studied in previous

papers that how these techniques have performed better results when performed on different data sets. The available raw medical data are widely distributed, different and voluminous by nature that must be collected and stored in data warehouses in organized form. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires and conditions of the patients and to make adequate and optimal decisions about their treatments. It is believed that the data mining can significantly help in the Lung Cancer research and ultimately improve the quality of health care of Lung Cancer patients. It can also be implemented using several classification techniques. Future models can be used in the design of clinical decision support system for mining Lung Cancer.

REFERENCES

[1] The Global Conference for Wikimedia,(2014);

[2] P.J. Kaur ,Pallavi , " Data Mining Techniques for Software Defect Prediction ", International Journal Of Software and Web Sciences 3(1), December, 2014 February, 2013, pp. 54-57. on software engineering, Vol. 32, no. 2, February 2016

[3] Qinbao Song, Martin Shepperd, Michelle Cartwright, and Carolyn Mair, "Software Defect Association Mining and Defect Correction Effort Prediction", IEEE Transactions on software engineering, Vol. 32, no. 2, February 2016.

[4] Baojun Ma1 Karel Dejaeger2 Jan Vanthienen2 Bart Baesens2, "Software Defect Prediction Based on Association Rule Classification", The 2014 International Conference on E-Business Intelligence.

[5] M.C.M. Prasad, L.Florence,A.Arya," A Study on Software Metrics based Software Defect Prediction using Data Mining and Machine Learning Techniques International Journal of Database Theory and Application Vol.8, No.3 (2015).

[6] R.Goyala, P.Chandraa, Y. Singha, "Suitability of KNN Regression in the Development of Interaction ased Software Fault Prediction Models" IERIProcedia International Conference on Future Software ineering and Multimedia Engineering, Elsiever, vol 6, pp 15-21, (2016).

[7] Automated Software Engineering (ASE)", 2013 IEEE/ACM 28th International Conference, (2013).

[8] M. Jureczko, "Significance of Different Software Metrics in Defect Prediction", Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wybrzeże Wyspiańskiego vol. 27, pp.50- 370.

[9] Michael Laszlo and Sumitra Mukherjee, Member IEEE, "A Genetic Algorithm Using Hyper- Quadtrees forCow-DimensionalK- meansClustering",IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28 April 4, 2017

[10] B. Liu, Y. Ma, C.K. Wong, "Improving an association rule based classifier," In Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, 2013.

