Research Paper Medical Science

Diagnosis and Prognosis of Breast Cancer Using Data Mining Techniques

Mrs.R.AMUTHA Assistant Professor, PSG College of Arts & Science, Coimbatore, India

Mrs.M.SAVITHRI Assistant Professor, Dr.NGP College of Arts & Science, Coimbatore, India

KEYWORDS diagnosis, classification, Breast cancer, malignant, benign.

ABSTRACT

Breast cancer is one of the most common cancers for women and cause of cancer death among women in worldwide [1]. In developed countries cancer rates are increased when compared with developing countries There are different reasons for this, one of the key factors are breast cancer is more common in elderly women. The major factor for this disease is women's eating habits & their lifestyles. In India, the risk factor is higher in urban areas compared to rural areas. The female risk group age is 43-46 years in India where as in the west the age group is 53-57 years are most prone to this kind of cancer. Risk factors for this type of cancer includes lack of physical exercise, hormone replacement therapy during menopause, obesity, ionizing radiation, drinking alcohol, early age at first menstruation, and having children late or not at all. In early stage, Breast self-exam and mammography may helps to diagnosis of breast cancer. The treatment is possible during early stage which consists of lumpectomy, radiation, hormone therapy and mastectomy. Detection of cancer in early stage can help the patients for proper treatment. Diagnosing & detecting the cancer in late phases is very crucial for survival. This paper deals with different types of classification algorithms of data mining used to diagnosis and prognosis of breast cancer.

INTRODUCTION

Breast cancer is cancer that develops from breast tissue. [2] Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin.[3] In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin[4].Cancer is a malignant cell which becomes a major cause of death which is difficult to prevent [5, 6]. In India younger women are becoming more susceptible for breast cancer. According to Globocan data (International Agency for Research on Cancer) in India the number of new cases of breast cancer will increase around 1, 80,000 by the year 2020. Breast cancer is the most common cancer diagnosed in Indian women. Due to the improvement effectiveness of classification and prediction systems in Data mining approaches in medical domain novel research directions are identified. Medical diagnostics includes valuable information and knowledge which is often hidden. Retrieving information and processing these data is a difficult task. Data Mining is a powerful tool to handle the task.

The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. There are many techniques to predict and classification breast cancer pattern. The early diagnosis helps to distinguish between benign tumor and malignant tumor without doing a surgical biopsy which is useful to assign the patients with either benign category (noncancerous) or malignant category (cancerous). In this paper, we examine the essential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to huge volume of healthcare data. Knowledge Discovery in Databases (KDD) process is to extract knowledge from huge data to identify and exploit patterns and relationships among large number of variables. This will predict the outcome of a disease using the historical case stored within datasets [7].

A. About Breast Cancer

The breast consists of billions of microscopic cells like other part of the body. These cells are growing; new cells are made to replace the old ones that died. Cancer begins when the cells in a part of the body grows uncontrollable. There are different kinds of cancer they all start when cells growth goes to abnormal. New cells are formed when they don't need and the damaged and old cell doesn't die as they should. These extra cells forms a tissue called a lump, growth or tumor. Breast cancer starts in the cells of breast. Tumors in the breast can be benign (not cancer) or malignant (cancer) which is mostly found in the women. Benign tumors are not harmful and rarely invade the tissues around them. This tumor doesn't spread to other parts of the body & it can be removed and usually don't grow back. Malignant tumors can invade nearby organs and tissues (such as the chest wall) which can spread other parts of the body. It can be removed but sometimes grow back [13].

A woman's breast is made up of glands that can make breast milk (lobules), small tubes that carry milk from the lobules to the nipple (ducts), fatty and connective tissue, blood vessels, and lymph vessels. Most breast cancers begin in the cells that line the ducts. Fewer breast cancers start in the cells lining the lobules. Cancers can also start in cells of the other tissues in the breast.

B. Risk Factors

The primary risk factors for breast cancer are female sex and older age. A risk factor is anything that affects your chance of getting a disease, such as cancer. Different cancers have different risk factors. Some of the risk factors are: Gender, Age, Genetic risk factor, woman have a family history of breast cancer, woman has a cancer in one breast has a higher chance to getting cancer in another breast , woman with certain benign breast changes may have an increased risk of breast cancer, . Having LCIS (Lobular carcinoma in situ) increases a woman's risk of getting cancer in either breast later, early menstrual periods, radiation treatment to the chest area, taking DES(diethylstilbestrol) drug, having children at later in life and using hormone therapy after menopause.

CLASSIFICATION TECHNIQUES

Data classification is a two-step process which consists of

learning step and a classification step. The aim of classification is to construct classifiers that predict categorical class labels. Thus the classifier is build based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. The data mining consists of several methods. Different methods serve different purposes, each method having its own advantages and disadvantages. Classification techniques are applied to assign patients either a benign (non cancerous) or a malignant (cancerous) group. Classification is the most important task to maps the data into predefined targets. The main aim of the classification is to build a classifier based on some attributes. The commonly used methods for data mining classification tasks can be classified into the following groups [8]. Decision Trees, Naive-Bayesian methods, Sequential Minimal Optimization (SMO), IBK, BF Tree etc.

A. Decision Trees

Decision trees are created by algorithms that identify different methods of splitting a data set into branch like segments. Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. Each internal node denotes a test attribute & each branch represents an outcome of the test and the terminal nodes hold the class label. The main aim of this algorithm is to create the classification model which predicts the target attribute value based on example sets input attribute. Using recursive partitioning decision trees are generated [10]. That is splitting the values of attributes recursively.

The decision tree classifier does not require any domain knowledge or parameter setting. This algorithm use parameters like data partition, attribute list & attribute selection method. Some of the commonly used algorithms are HUNTS algorithm, CART, ID3, SLIQ, SPRINT,C4.5.

B. IBK (K nearest Neighbors classifier)

K-Nearest Neighbor (KNN) classification [9] classifies instances based on their similarity. It compares a given test sample with training samples that are similar to it. The training sample consists of n attributes. In multi-dimensional space classification is done based on the nearest neighbors in which each sample is considered as a point. The K value for nearest neighbors may vary & it determines how many points are to be considered as neighbors to conclude how to classify an unknown instance. This classifier searches the pattern space for the K training samples that are nearest to the unknown samples in the given unknown sample. The unknown samples are assigned to the most common class among its neighbors of K-nearest. K-nearest is the machine learning algorithm, a sample is classified by a majority of its neighbors, with the sample being assigned to the class most common among its k nearest neighbors. (K is a positive integer, typically small). The sample is assigned to the class of its nearest neighbor when k=1.The basic k-Nearest Neighbor algorithm is involved two steps: Identify the k training samples which are closest to the unknown sample. Take the commonly occurring classification for these k samples.

C. Support Vector Machine

Support Vector Machines (SVMs) are supervised learning methods used for classification and regression tasks that originated from statistical theory. It is an algorithm that attempts to identify linear optimal separator (hyper-plane) in multi-dimensional space. Data points are categorized in SVM by mapping data to a high dimensional feature space. It uses a non linear mapping to convert training data into higher dimension. SVM is a suitable algorithm to deal with interaction among features and redundant features. SVM takes set of predicts and input data, for each data input which of the two possible classes comprises the input, making the SVM a non probabilistic binary linear classifier. Each marked as belonging into these categories; an algorithm creates a model that assigned into one category or other. In SVM model examples are represented space & the examples of separate categories are divided into wide gap. Find the new examples are belongs to which side of the gap and mapped into that. Support vector machine are used to construct a set of hyper planes with higher dimensional space which are used for the tasks like classification, regression and other tasks. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [12].

D. Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization is a new algorithm for training Support Vector Machines (SVMs). The Sequential Minimal Optimization (SMO) algorithm proposed by John Platt in 1998 [11], is a simple and fast method for training a SVM. The main idea is derived from solving dual quadratic optimization problem by optimizing the minimal subset including two elements at each iteration. The advantages of SMO are that it can be implemented simply and analytically. Training a support vector machine requires the solution of a very large quadratic programming optimization problem. SMO breaks this large quadratic programming problem into a series of smallest possible quadratic programming problems. These small quadratic programming problems are solved analytically, which avoids using a time-consuming numerical quadratic programming optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. Because matrix computation is avoided, SMO scales somewhere between linear and quadratic in the training set size for various test problems, while the standard chunking SVM algorithm scales somewhere between linear and cubic in the training set size. SMO's computation time is dominated by SVM evaluation; hence SMO is fastest for linear SVMs and sparse data sets.

E. Neural networks

Neural network is a type of artificial intelligence that imitates the way the human brain works. Human brain consists large number of neurons which are interconnected by synapses. A neural network is a collection of connected input/output units & every connection has associated with weight. Predict the correct class label of input by adjusting the weight [14]. Any neural network should be trained before it can be considered as an intelligent and ready to use. Using training data sets neural networks are trained.

CONCLUSION

This paper provides a study of breast cancer diagnosis and prognosis problems and how data mining techniques uncover patterns in the hidden data that helps the decision making. In medical field data mining classification techniques is highly acceptable and can help to take decision in early diagnosis and to avoid biopsy. In the above study, it proposes different classification techniques to find the accuracy within the prediction of positive and negative. By using these techniques doctors can take better decision and save several patients precious life.

REFERENCES

[1]A Novel Approach for Breast Cancer Detection using Data Mining TechniquesI VikasChaurasia, SaurabhPal IJIRCCE Vol. 2, Issue 1, January 2014. | [2] —Application of Data Mining Techniques to Model Breast Cancer Data S. Syed Shajahaan —,S. Shanthi , V. ManoChitra IJETAE Volume 3, November 2013. | [3] Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining TechniquesI ShomonaGracia Jacob, R. GeethaRamani Proceedings of the World Congress on Engineering and Computer Science Vol I October 2012. | [4]Jump up^ Saunders, Christobel; Jassal, Sunil (2009). Breast cancer (1. ed. ed.). Oxford: Oxford University Press. p. Chapter 13. ISBN 978-0-19-955869-8. | [5] Delen D, Patil N. Knowledge extraction from prostate cancer data. The 39thAnnual Hawaii International | Conference on System Sciences; 2006; 1-10. | [6] National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data(1973-2008). Cancer Statistics Branch; 2011. | [7] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. Data mining for indicators of early | mortality in a database of clinical records. Artif Intell Med 2001;22:215—31. | [8] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001 | [9] J. Han and M. Kamber, Data Mining—Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed.San Mateo, CA: Morgan Kaufmann, 2006. | [10] RapidMiner 5.0 Help (2013) | [11] Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98- 14,Microsoft Research, 1998. | [12] D. Wolpert and W. Macready, No Free Lunch Theorems for Search, Santa Fe Institute, Technical report no., No. SFI-TR-95-02-010,1995. | [13] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012. | [14]Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005. |