ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

A Novel Approach for Breast Cancer Detection using Data Mining Techniques

Vikas Chaurasia[1], Saurabh Pal[2]

Research Scholar, Sai Nath University, Ranchi, Jharkhand, India[1]

Head, Dept. of MCA, VBS Purvanchal University, Jaunpur, UP, India[2]

ABSTRACT: Breast cancer is one of the leading cancers for women when compared to all other cancers. It is the second most common cause of cancer death in women. Breast cancer risk in India revealed that 1 in 28 women develop breast cancer during her lifetime. This is higher in urban areas being 1 in 22 in a lifetime compared to rural areas where this risk is relatively much lower being 1 in 60 women developing breast cancer in their lifetime. In India the average age of the high risk group is 43-46 years unlike in the west where women aged 53-57 years are more prone to breast cancer.

The aim of this paper is to investigate the performance of different classification techniques. The data breast cancer data with a total 683 rows and 10 columns will be used to test, by using classification accuracy. We **analyse** the breast Cancer data available from the Wisconsin dataset from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. In this experiment, we compare three classification techniques in Weka software and comparison results show that Sequential Minimal Optimization (SMO) has higher prediction accuracy i.e. 96.2% than IBK and BF Tree methods.

Keywords: Breast cancer, Classification techniques, Sequential Minimal Optimization (SMO), IBK, BF Tree.

I. INTRODUCTION

Globally, the rising breast cancer incidence and mortality represent a significant and growing threat for the developing world. Breast cancer is on the rise across developing nations, mainly due to the increase in life expectancy and lifestyle changes such as women having fewer children, as well as hormonal intervention such as post-menopausal hormonal therapy. In these regions, mortality rates are compounded by the later stage at which the disease is diagnosed, as well as limited access to treatment, presenting a 'ticking time bomb' which health systems and policymakers in these countries need to work hard to defuse. A recent study by the Asian Pacific Journal of Cancer Prevention indicated that in the urban areas of Delhi, only 56% women were aware of breast cancer; among them, 51% knew about at least one of the signs/symptoms, 53% were aware that breast cancer could be detected early, and only 35% mentioned about risk factors. In rural Kashmir, only 4% of women had received any training or education about the purpose and technique of breast self exam.

In the recent years the data from several domains including banking, retail, telecommunications and medical diagnostics includes valuable information and knowledge which is often hidden. Processing these huge data and retrieving meaningful information from it is a difficult task. Data Mining is a powerful tool for handling this task. Data mining in breast cancer research has been one of the important research topics in medical science during the recent years [1]. 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. This paper empirically compares performance of three classical decision tree classifiers that are suitable for direct interpretability of their results.

A. *Breast Cancer (An overview)*

Cancer begins in cells, the building blocks that make up all tissues and organs of the body, including the breast. Normal cells in the breast and other parts of the body grow and divide to form new cells as they are needed. When normal cells grow old or get damaged, they die, and new cells take their place. Sometimes, this process goes wrong. New cells form when the body doesn't need them, and old or damaged cells don't die as they should. The buildup of extra cells often forms a mass of tissue called a lump, growth, or tumor.

Tumors in the breast can be benign (not cancer) or malignant (cancer):

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Benign tumors:
Are usually not harmful
Rarely invade the tissues around them
Don't spread to other parts of the body
Can be removed and usually don't grow back
Malignant tumors:
May be a threat to life
Can invade nearby organs and tissues (such as the chest wall)
Can spread to other parts of the body
Often can be removed but sometimes grow back

B. *Risk Factors*

Although risk factors don't tell everything. Many risk factors may increase chance of having breast cancer; it is not yet known just how some of these risk factors cause cells to become cancer (American cancer society, 2002).

- **Gender:** Breast cancer is about 100 times more common in women than in men.
- **Age:** The chance of getting breast cancer goes up as a woman gets older.
- **Genetic risk factors:** Inherited changes (mutations) in certain genes like *BRCA1* and *BRCA2* can increase the risk.
- **Family history:** Breast cancer risk is higher among women whose close blood relatives have this disease.
- **Personal history of breast cancer:** A woman with cancer in one breast has a greater chance of getting a new cancer in the other breast or in another part of the same breast.
- **Race:** Overall, white women are slightly more likely to get breast cancer than African-American women. Asian, Hispanic, and Native-American women have a lower risk of getting and dying from breast cancer.
- **Dense breast tissue:** Dense breast tissue means there is more gland tissue and less fatty tissue. Women with denser breast tissue have a higher risk of breast cancer.
- **Certain benign (not cancer) breast problems:** Women who have certain benign breast changes may have an increased risk of breast cancer. Some of these are more closely linked to breast cancer risk than others.
- **Lobular carcinoma in situ:** In this condition, cells that look like cancer cells are in the milk-making glands (lobules), but do not grow through the wall of the lobules and cannot spread to other parts of the body. It is not a true cancer or pre-cancer, but having LCIS increases a woman's risk of getting cancer in either breast later.
- **Menstrual periods:** Women who began having periods early (before age 12) or who went through the change of life (menopause) after the age of 55 have a slightly increased risk of breast cancer.
- **Breast radiation early in life:** Women who have had radiation treatment to the chest area (as treatment for another cancer) as a child or young adult have a greatly increased risk of breast cancer. The risk from chest radiation is highest if the radiation were given during the teens, when the breasts were still developing.
- **Treatment with DES:** Women who were given the drug DES (diethylstilbestrol) during pregnancy have a slightly increased risk of getting breast cancer
- **Not having children or having them later in life:** Women who have not had children, or who had their first child after age 30, have a slightly higher risk of breast cancer. Being pregnant many times or pregnant when younger reduces breast cancer risk.
- **Certain kinds of birth control:** Studies have found that women who are using birth control pills or an injectable form of birth control have a slightly greater risk of breast cancer than women who have never used them.
- **Using hormone therapy after menopause:** Taking estrogen and progesterone after menopause increases the risk of getting breast cancer.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

- **Not breastfeeding:** Some studies have shown that breastfeeding slightly lowers breast cancer risk, especially if breastfeeding lasts 1½ to 2 years.
- **Alcohol:** The use of alcohol is clearly linked to an increased risk of getting breast cancer. Even as little as one drink a day can increase risk (Ranstam & Olsson, 1995).
- **Being overweight or obese:** Being overweight or obese after menopause is linked to a higher risk of breast cancer (Pujol et. al. 1997).

The remainder of this paper is organized as follows: The background section investigates provides the reader with the background information on breast cancer research, survivability analysis, commonly used prognosis factors and previously published relevant literature., the method section explains the proposed classification techniques for enhancing applied methods accuracy in diagnosing breast cancer patients, and the results section is followed by a conclusion section.

II. BACKGROUND

There is large number of papers about applying machine learning techniques for survivability analysis. Several studies have been reported that they have focused on the importance of technique in the field of medical diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies. Here are some examples:

Bittern et al. [2] used artificial neural network to predict the survivability for breast cancer patients. They tested their approach on a limited data set, but their results show a good agreement with actual survival.

Vikas Chaurasia et al. [3] used RepTree, RBF Network and Simple Logistic to predict the survivability for breast cancer patients.

Djebbari et al. [4] consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results.

Liu Ya-Qin's [5] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.

Tan AC's [6] used C4.5 decision tree, bagged decision tree on seven publicly available.

Bellaachi et al. [7] used naive bayes, decision tree and back-propagation neural network to predict the survivability in breast cancer patients. Although they reached good results (about 90% accuracy), their results were not significant due to the fact that they divided the data set to two groups; one for the patients who survived more than 5 years and the other for those patients who died before 5 years.

Jinyan LiHuiqing Liu's [8] experimented on ovarian tumor data to diagnose cancer using C4.5 with and without bagging.

Vikas Chaurasia et al. [9] used Naive Bayes, J48 Decision Tree and Bagging algorithm to predict the survivability for Heart Diseases patients.

Vikas Chaurasia et al. [10] used CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) to predict the survivability for Heart Diseases patients.

Pan wen [11] conducted experiments on ECG data to identify abnormal high frequency electrocardiograph using decision tree algorithm C4.5 with bagging.

My Chau Tu's [12] proposed the use of bagging with C4.5 algorithm, bagging with Naïve bayes algorithm to diagnose the heart disease of a patient.

Dong-Sheng Cao's [13] proposed a new decision tree based ensemble method combined with feature selection method backward elimination strategy with bagging to find the structure activity relationships in the area of chemometrics related to pharmaceutical industry.

Dr. S.Vijayarani et al., [14] analyses the performance of different classification function techniques in data mining for predicting the heart disease from the heart disease dataset. The classification function algorithms is used and tested in this work. The performance factors used for analyzing the efficiency of algorithms are clustering accuracy and error rate. The result illustrates shows LOGISTICS classification function efficiency is better than multilayer perception and sequential minimal optimization.

Tsirogiannis's [15] applied bagging algorithm on medical databases using the classifiers neural networks, SVM'S and decision trees. Results exhibits improved accuracy of bagging than without bagging.

My Chau Tu's [16] used bagging algorithm to identify the warning signs of heart disease in patients and compared the results of decision tree induction with and without bagging.