ORIGINAL RESEARCH PAPER

Computer Science

KEY WORDS: Breast Cancer, Support Vector Machine (SVM), Relevance Vector Machine (RVM), Big Data and Machine Learning.

BREAST CANCER CLASSIFICATION USING BIG DATA APPROACH

Dr. Savita Kumari Sheoran

Department of Computer Science & Engineering Indira Gandhi University Meerpur, Rewari (Haryana) – INDIA

ABSTRACT

Cancer has been identified as most dreadful disease worldwide. According to available statistical estimates 8.8 million people died every year due to cancer worldwide despite of costing US$ 1.16 trillions of money on treatment. In India alone around 2.5 million people are living with cancer and every year 0.7 million new cases are added. There are various types of cancers, many of which are gender susceptible as well as survival rate depends upon type of cancer. The breast cancer is a specific type of cancer in women only, which cause highest death among the women. In India 27% of women diagnosed with cancer are suffering with breast cancer, 50% of whom could not be survived despite of huge economic expenses. The breast cancer is curable but its survival rates are considerable if diagnosed and prevented early. For diagnosis, prevention and prediction of breast cancer it is significant that the available patient data be analysed properly to extract the proper cause. Also, Machine Learning approaches like Support Vector Machine (SVM) and Relevance Vector Machine (RVM) have been identified as best way to classify the Breast Cancer dataset. The clinical data set of cancer comprises of huge amount of data ranging from peta byte to exa byte which could not effectively be handled with ordinary data base management techniques. The Big Data is term recently coined to manage such enormous size data. However, only a few researches have been done worldwide to explore the breast cancer dataset through Big Data approach but they guided the meaning insight to cure this disease. This research paper investigates the Wisconsin Breast Cancer Data through Big Data approach using Hybrid SVM-RVM Model as classifier. The results show that this Hybrid SVM-RVM Model outperforms the nad've SVM and RVM.

1. INTRODUCTION

Cancer is a group of disease caused due to uncontrolled growth of cells in human body. Medically it is a malignant neoplasm which accumulate a heavy mass around it, known as tumor and may cause death if not cured. The tumors are generally of two types viz. benign and malignant. The benign tumor does not invade the nearby part tissues of body and hence present as a non cancerous cist only while malignant tumor invade the nearby parts of body and cause cancer. Cancer has various stages and survival of patient depend upon the stage of treatment. Owing to its low survival rate, the WHO has listed the cancer as one of the most dreadful disease which causes about 13% of world mortality. According to estimate drafted by National Cancer Institute 8.8 million people died every year due to cancer worldwide despite of costing US$ 1.16 trillions of money on treatment, which is roughly greater than the GDP of India. In India alone around 2.5 million people are living with cancer and every year more than 7 lakhs new cases of different types of cancers are reported at varied stages. These human and economic costs designate cancer as one of central topic research for medical, paramedical and allied fraternity. The breast cancer is a specific type of cancer in women only, which cause highest death among the women. In India 27% of women diagnosed with cancer are suffering with breast cancer, 50% of whom could not be survived despite of huge economic expenses.

The SEER (Surveillance Epidemiology and End Results) results of National Cancer Institute of USA reveals that cancer susceptibility in human is gender specific. There are more than two hundred types of cancer but breast cancer is one of most fatal form of cancer found only in the female only which cause about one-fourth of cancer among the women. In India 27% of women diagnosed with cancer are suffering with breast cancer, 50% of whom could not be survived despite of huge economic expenses.

The treatment of cancer includes diagnosis through various tests such as surgical biopsy and prognosis with medicines, therapy and radiations. The survival rate of patient depends upon the stage at which cancer tumor diagnosed and prognosis started. The Surgical Biopsy, most commonly used diagnosis test to confirm cancer cause high cost and have adverse impact on patients' psychology and family status. The cancer can also be diagnosed based on various symptoms and cell behavour, if proper data base for same is developed and analysed to find the exact pattern. Recently the Machine Learning various techniques have been developed to classify the data and find a meaningful pattern among them. Researches envisage that Machine Learning techniques such as Support Vector Machine (SVM) and Relevance Vector Machine (RVM) have useful impact on pattern classifications and useful for diagnosis and prognosis of cancer. The cancer dataset comprises of huge size because even a cell data may require peta bytes to store. Therefore, such datasets may size from peta bytes to exa bytes, which fall in the category of Big Data. The analysis of such data could not effectively be carried out with ordinary database management techniques and require Big Data analytics. Map Reduce is one of the most popular pattern extraction techniques. This paper primarily aims to develop a hybrid model based on SVM and RVM and analyse the performance of proposed model through Big Data analytics in comparison to nad've strategies. The rest part of this section will present such issues and proper database to test the proposed model.

1.1. Breast Cancer

The breast is milk production gland in female having lobules and nipple connected through ducts. The breast cancer is most common form of cancer accounting for approximately one-fourth of cancerous deaths and late detection put women at higher risk of death. About 70-80% of breast cancers developed in lobules while ducts cancer comprised only of about 20% of breast cancer cases. The beast cancer can be of three types as shown in figure 1, below.

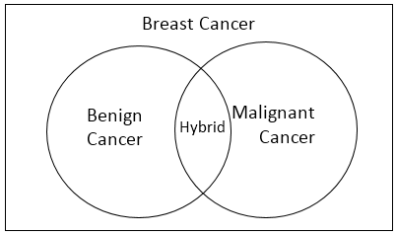

Figure 1: Different types of breast cancers

However the breast cancer is a complex disease and could not be attributed to single cause rather there are various risk factors, which contribute to possibility of ailment. These risks factors may be classified in to two categories as shown in figure 2, below.

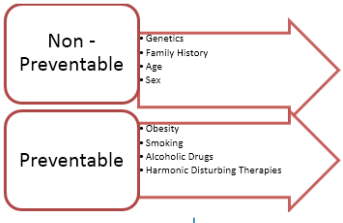

Figure 2: Risk factors of breast cancers

The breast cancer can primarily be identified with physical symptoms. Such identification paves a way for confirmation tests to ensure timely prognosis. Figure 3, enlists the main symptoms of breast cancer.

	Breast Lump	
	Symptoms of Breast Cancer	
Skin Dimpling		Breast Swelling
Nipple change		Blood Stain Discharge
	Breast Pain	

Figure 3: Symptoms of Breast Cancer

1.2. Machine Learning Approaches

Recently Machine Learning approaches have been found much useful to detect breast cancer. The most common Machine Learning approaches are elucidated below:

1.2.1. Support Vector Machine

The Support Vector Machine (SVM) is a non probabilistic binary and non linear statistical tool based on supervised learning. It analyse the data, recognise the pattern and classify the data based on common attributes by using kernel tricks. The SVM work to minimize the structural risk instead of objective function. It minimizes the bound on generalization error for all the data which were not used during the training. The SVM function is represented by equation

$$f_{SVM}(x_i) = W^T \emptyset(x_i) + b$$

1.2.2. Relevance Vector Machine

The Relevance Vector Machine (RVM) is a statistical tool for data classification based upon Bayesian estimation. It has been widely used on various types of cancers other than breast cancer and known to deliver optimal results using only few training samples. The classification function for Relevance Vector is

$$f_{RVM}(x_i) = \sum_{i=1}^{N} \propto_i K_{x,x_i}$$

Where K = Kernel Function, and
x_i= training sample

1.1. Big Data Analytics

Recent years have witnessed accumulation of huge amounts of unstructured, semi-structured and structured data. By collecting, storing, analyzing, and mining these data, an enterprise can obtain large amounts of individual users' sensitive data. This data is commonly known as "Big Data" due to its volume, the velocity with which it arrives, and the variety of forms it takes. These data not only meet up the demands of the enterprise itself, but also give services to other businesses if the data are stored on a big data platform. There are three characteristics of Big Data called "3V":

- **Volume** (the data volumes are huge which cannot be processed by traditional methods),
- **Velocity** (the data is created with great velocity and must be captured and processed quickly) and
- **Variety** (variety of data types: structured, semi-structured, and unstructured).

Since the Cancer Data meet all the characteristics requirements of Big Data and hence could be analysed in more better way using Big Data analytics. There are various platform like HADOOP MapReduce, 'R' and Hive tool to deal with Big Data. MapReduce is one of the popular platforms for analysing data for pattern extraction and classification.

1.2. Breast Cancer Experimental Data Set

Wisconsin Breast Cancer dataset archived by UCI Machine Learning Repository is freely available dataset to study breast cancer. The dataset is collection of multivariate data of fine needle aspirate (FNA) of a breast mass of donors obtained at University of Wisconsin Hospitals. The Data consists of following datasets:

- Wisconsin (Original) Breast Cancer Dataset
- Wisconsin Diagnostic Breast Cancer Dataset
- Wisconsin Prognostic Breast Cancer Dataset
- Breast Tissue Dataset

This paper used only the original Wisconsin Breast Cancer (WBC) Dataset for analysis purpose. The WDBC dataset contains 699 instances based on 11 attributes out of which 458 are benign and 241 are malignant. In this dataset 16 datasets have missing and 583 are complete. Figure 4 presents the composition of WBC dataset used for experimental purpose.

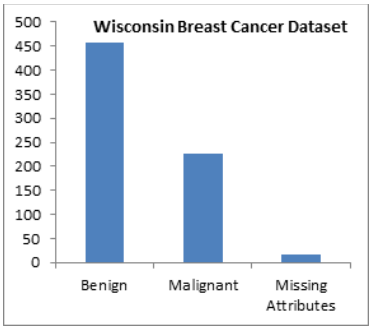

Figure 4: Composition of Wisconsin Breast Cancer Dataset

The rest of this paper is composed as follow: Section 2 review the available literature. Section 3 proposed the Hybrid SVM-RVM Model and analyse presents the result obtained through MapReduce. Section 4 concludes the paper with future possibility of extension of the results obtained in this research.

2. LITERATURE REVIEW

The Wisconsin Breast Cancer Dataset was developed by Dr. Wolberg in 1995. Thereafter, a lot of work is done by various medical and allied practitioners to simplify the breast cancers detection and prognosis. In 1999 Xin Yao et al. has implemented artificial neural network for breast cancer diagnosis using Negative correlation training algorithm using two approaches viz. evolutionary and ensemble approach. In 2004 Tuba Kiyan et al. has applied Neural Network on WBCD to estimate the diagnosis accuracy of various techniques. In 2007 Dr. Sumathi et a. have used genetic algorithms approach to WBDC and found that genetic algorithm not only improve the accuracy but also reduce the time taken to train the network. In 2010 Val´erie have comparatively studied various statistical models to obtain a desirable result from WBCD.

In 2009, Y. Iraneus Anna Rejani and Dr. S. Thamarai Selvi used SVM for early detection of breast cancer. In 2012, Muhammad Rafi et al. used SVM and RVM techniques for document classification without using minimum accuracy limit and find that predicting accuracy of RVM is much higher than SVM. In 2012, Z Qinli et al. apply SVM approach and its application to breast cancer diagnosis. And obtain results for both artificial and real data. It is observed that it is competent to reduce the generalization error and computational cost. In 2013, D. Kishore used Big Data approaches to Medical field and paved a way for its implementation to detect Breast Cancer.

This literature survey reveals that a lot of work is accomplished on detection, diagnosis and prognosis of breast cancer. The SVM and RVM has been explored and found effective machine learning tools to accurate and valuable techniques for breast cancer detection. But still the hybrid model of SVM and RVM has not explored. This research intends to explore this unaddressed segment through WBCD using MapReduce.

3. EXPERIMENTAL RESULTS

This section presents a Hybrid model which implement both SVM and RVM techniques together, known as Hybrid SVM-RVM Model. The experimental results obtained over MapReduce platform are

presented in figure 5 and figure 6. For performance analysis two parameters viz. classification accuracy and implementation times were considered. The experiential results reveals that the Hybrid SVM-RVM model outperform the nad've SVM and RVM model.

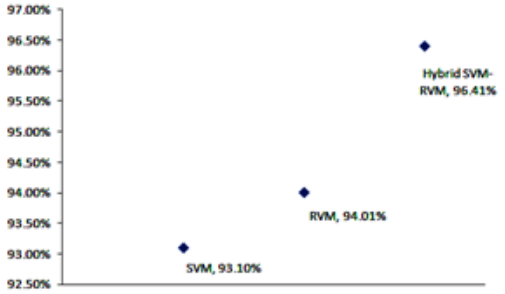

Figure 5: Classification Accuracy of Different Models

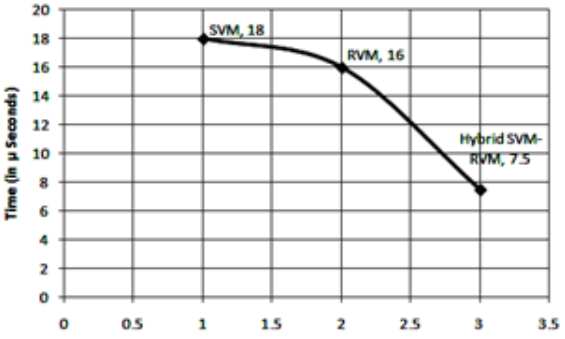

Figure 6: Implementation Time (in μ seconds) of Different Models

4. CONCLUSION

Since breast cancer is one the dreadful form of cancer. SVM and RVM techniques of Machine Learning are frequently been used for improving the classification efficiency of breast cancer datasets. This research has explored a Hybrid SVM-RVM model through an innovative approach of Big Data. The experimental results were obtained using WBCD. The experimental results presented in section 3 reveals that the proposed hybrid model analysed with Big Data technique outperform the nad've model for parameters of classification accuracy and implementation time. The application of this model with Big Data to real time data may yield better and disease specific results. Also this model may be explored for other type of cancer as well.

REFERENCES

[1]. Zaslavsky, A., Perera, C., and Georgakopoulos, D., 2012. Sensing as a Service and Big Data. Proceedings of the International Conference on Advances in Cloud Computing (ACC), pp. 21-29.
[2]. William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive.ics. uci. edu/ml/] .
[3]. Tuba kiyan, Tulay Yildirim, 2004 "Breast cancer diagnosis using statistical neural networks", Journal of Electrical and Electronic Engineering.
[4]. C.P. Sumathi, T. Santhanam and A. Punitha 2007, "Combination of genetic algorithm and ART neural network for breast cancer diagnosis", Asian Journal of information technology, Medwell journals.
[5]. Val´erie Bourd`es, St´ephane Bonnevay, Paolo Lisboa, R´emy Defrance, David P´erol, Sylvie Chabaud, Thomas Bachelot, Th´er`ese Gargi, and Sylvie N´egrier 2010, "Comparison of Artificial Neural International Journal of Distributed and Parallel Systems (IJDPS) 4(3), Network with Logistic regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient outcomes".
[6]. Y.Ireaneus Anna Rejani, Dr.S.Thamarai Selvi Noorul 'Early Detection Of Breast Cancer Using SVM Classifier Technique', International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-130.
[7]. M. Rafi and M.S. Shaikh, 2013, "A Comparison of SVM and RVM for Document Classification" Proceedings of Computer Science, pp 3-8.
[8]. D. Kishor, 2013, "Big Data: The New Challenges in Data Mining". International Journal of Innovative Research in Computer Science & Technology, 1(2), pp. 39-42.