A Novel Optimized Model to Anticipate Breast Cancer in Women

Mallojjala. Raja Shree
JNTUH: Department of CSE
Keshav Memorial Institute of Technology
Hyderabad, India

Midde. Venkateswarlu Naik
JNTUH: Department of CSE
Keshav Memorial Institute of Technology
Hyderabad, India

Abstract— This article proposed a novel optimized model to anticipate breast cancer in hospital domain. Usually breast cancer is the most treacherous disease through which enormous women are suffering for last decades worldwide. In this situation building optimal model to anticipate breast cancer is essential to prevent in women. Our model has been enhanced accuracy level than existing models and techniques. Earlier authors have been achieved performance of machine learning techniques up to 97% to predict cancer. Our novel optimal approach achieved 99% of the accuracy to anticipate breast cancer using machine learning algorithms.

Keywords—Prediction of Cancer;Optimized Model; Breast Cancer dataset;Rapid Miner.

I. INTRODUCTION

A constant growth related to cancer research has been performed [1] since decades. Earlier authors employed various techniques to identify different types of cancer before they cause. They have built innovative tactics for the forecast of cancer treatment. Huge amount of cancer information have been gathered and are available to the therapeutic research area. The challenging and the most interesting tasks for medical doctor is the precise guess of the malignancy infection.

The second major cause of women's death is breast cancer (after lung cancer) [4]. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated [5]. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women [6]. Information and Communication Technologies (ICT) can play potential roles in cancer care. In fact, Big data has advanced not only the size of data but also creating value from it. Big data, that becomes a synonymous of data mining, business analytics and business intelligence, has made a big change in BI from reporting and decision to prediction results [7]. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives. Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis and prognosis [8-11]

The rest of the article is ordered as follows. Section II describes about literature survey. Section III deals about machine learning techniques employed in this article. Section IV describes regarding over all methodology applied during experiments. Section V deals about results and discussions. Section VI describes about conclusion and future scope of this research work.

II. LITERATURE SURVEY

Classification is essential activity in data mining domain. An earlier author has been carried out lot of research towards prediction of cancer in hospital domain. These experiments have been utilized machine learning techniques such as Naïve Bayes, RBF NN (Neural Networks), SVM-RBF kernel, J48 in breast cancer datasets. Among all above specified techniques, SVM-RBF has outperformed 96.84% of accuracy. The author [2] has applied C4.5, SVM ,NB, k-NN and gained an accuracy of 97.13% using SVM and rest of the techniques has got lesser than SVM accuracy. The author [3] has employed the cancer susceptibility prediction for breast cancer using ANN has gained an accuracy of 96.5% and those of the rest techniques achieved lesser than ANN. There are various methods used for breast cancer recurrence prediction in this paper namely, BN, SVM, Graph-based SSL algorithm of which BN has achieved 96% of accuracy.

In this article we employed machine learning techniques such as Optimized decision tree algorithm, Random Forest algorithm for anticipating breast cancer. Among these algorithms decision tree with optimization process achieved 99% of accuracy to predict cancer.

III. Overview of Techniques Employed

A. Optimized Decision Tree

It generates a decision tree for classification of both nominal and numerical data. A decision tree is a like a graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a target attribute based on several input attributes of the exampleset. Decision trees are generated by recursive partitioning. Recursive partitioning means repeatedly splitting on the values of attributes. In every recursion the algorithm follows the following steps:

- An attribute A is selected to split on. Making a good choice of the attributes to split on each stage is crucial to generation of a useful tree. The attribute is selected depending upon a selection criterion which can selected by the criterion parameter.
- Exanples in the exampleset are sorted into subsets, one for each value of the attribute A in case of a nominal attribute. In case of numerical attributes. Subsets are formed for disjoint ranges of attribute values.
- A tree is returned with one edge or branch for each subset. Each branch has a descendant subtree or a label value produced by applying the same algorithm recursively.
- There are less than a certain number of instances or examples in the current subtree. This can be adjusted by using the minimal size for split parameter.
- No attribute reaches a certain threshold. This can be adjusted by using the minimum gain parameter.
- The maximal depth is reached. This can be adjusted by using the maximal depth parameter.

Apart from Optimization (Grid) process will help to choose best parameters to show outperformed results. Subsequently, The Optimize Parameters (Grid) operator has a sub process in it. It executes the sub process for all combinations of selected values of the parameters and then delivers the optimal parameter values through the *parameter* port. The performance vector for optimal values of parameters is delivered through the *performance* port. Any additional results of the sub process are delivered through the *result* ports.

For example, if you select 3 parameters and 25 steps for each parameter then the total number of combinations would be above 390625 (i.e. 25 x 25 x 25). The sub process is executed for all possible combinations. This operator returns optimal parameter set which can also be written to a file with the Write Parameters operator. This parameter set can be read in another process using the Read Parameters operator.

B. Random Forest

The Random Forest operator generates a set of random trees. The random trees are generated in exactly the same way as the Random Tree operator generates a tree. The resulting forest model contains a specified number of random tree models. The *number of trees* parameter specifies the required number of trees. The resulting model is a voting model of all the random trees. For more information about random trees please study the Random Tree operator.

The representation of the data in form of a tree has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a *target attribute* (often called *class* or *label*) based on several input attributes of the ExampleSet. Each interior node of the tree corresponds to one of the input attributes. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labeled with disjoint ranges. Each leaf node represents a value of the *label* attribute given the values of the input attributes represented by the path from the root to the leaf. For better understanding of the structure of a tree please study the Example Process of the Decision Tree operator.

Pruning is a technique in which leaf nodes that do not add to the discriminative power of the tree are removed. This is done to convert an over-specific or over-fitted tree to a more general form in order to enhance its predictive power on unseen datasets. Pre-pruning is a type of pruning performed parallel to the tree creation process. Post-pruning, on the other hand, is done after the tree creation process is complete.

IV. Methodology

In Fig 1, overall optimized model has been designed to anticipate breast cancer using machine learning techniques. This section deals about methodology how an experiment takes place for prediction process. The overall process describes into three steps.

A. Step1

During this step, load the desired dataset which contains attribute information with respective labels to implement supervised learning.

B. Step2

During this step, preparation of data using tasks such as Extraction, Transformation, learn and also mark the target label attribute.

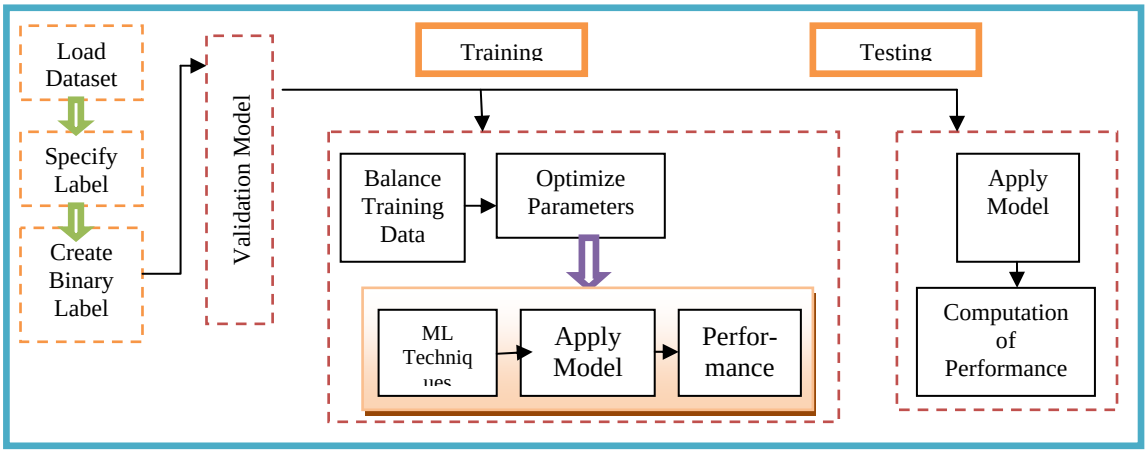

Fig. 1 Optimized Machine Learning model for Prediction breast cancer

C. *Step3*

During this step, cross validation splits the dataset for training and then independent testing. This splitting is done numerous times to get a better performance estimate. Subsequently, training data would be balanced in order to learn our model according to the expected behaviour and rebalance the data to focus on the case we are interested in.

Instead of just adding a model configured by hand, it should be optimize. Based on the "wisdom of the crowed" we applied decision tree and optimize the maximal depth in the range from 20 to 29. The tree configuration with an optimal accuracy for the training set will be chosen and passed on to the testing phase.

The model trained on the training data is applied to the independent test data set and the model performance is calculated. The performance values obtained on the different folds of the cross validation are finally averaged to produce an average performance measure of its dispersion, which gives an estimate of the model stability when applied to different data samples.

V. Results and Discussions

In this article, author has employed breast cancer dataset with sample of 280 records. We employed decision tree with optimization grid process and random forest machine learning algorithms for predicting breast cancer. We compared with existing algorithms such as C4.5, SVM, NB, and KNN. Among all algorithms SVM has outperformed around 97.13 as accuracy.

We employed decision tree with optimization process and random forest algorithm, among these random forest algorithm has accuracy of 97.90% which has best accuracy than earlier employed algorithm on same dataset. Subsequently, decision tree with optimization process has achieved 99.30%

	Machine Learning Algorithms					
Evaluation Parameters	C4.5	SVM	NB	K-NN	Decision Tree	Random Forest
Time taken to do experiment	0.06	0.07	0.05	0.01	0.07	0.11
Correctly Classified Instances	665	678	671	666	284	280
Incorrectly Classified Instances	34	21	28	33	2	6
Accuracy	95.13	97.13	95.99	95.27	**99.30**	**97.90**

VI. Conclusion & Future enhancements

This article proposed a novel optimization model for prediction of breast cancer with substantial progress in terms of accuracy with 99.30% using Decision Tree Optimization process. Subsequently, we employed Random Forest technique which outperformed 97.90% of accuracy than earlier authors applied techniques. K-NN technique has 0.01 seconds of time to test the entire model. Therefore, K-NN technique plays important role when time boundary is taken into consideration.

The scope of this research article is to enhance accuracy of existing techniques when breast cancer data would be enormous. So under certain circumstances, consider more attributes or paramaters for predicting root causes of breast cancer.

ACKNOWLEDGMENT

We would like to thank our "Keshav Memorial Institute of Technology" management for

encouraging this research work morally and financially.

REFERENCES

[1] Hanahan D, Weinberg RA, "Hallmarks of cancer: the nest generation.Cell 2011; 144:646-74.

[2] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noell "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), pp.1064-1069..

[3] Konstantina Kourou, Themis P. Exarchos, Konstantinos P.Exarchos, Michalis V. Karamouzis, Dimitrios I.Fotiadis "Machine Learning Applications in Cancer Prognosis and Prediction, " Coputational and Structural Biotechnology Journal, vol.13, pp. 8-17 (2015).

[4] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012..

[5] Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332.

[6] "Globocan 2012 - Home." [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015].

[7] Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud-Technol-Appl.2015:pp.1-7.

[8] O. Fotunato, M.Boeri, C. Verri, D.Conte, M.Mensah,P. Suatoni, " Assessment of Circulating microRNAs in plasmaa of lung cancer patients Molecules", 19(2014), pp. 3038-3054.

[9] H.M.Heneghan, N.Miller, M.J.Kerin, " MiRNAs as biomarkers and Therapeutic targets in cancer currOin Pharmacol", 10 (2010), pp. 543-550.

[10] D.Madhavan, K.Cuk, B. Burwinkel, R.Yang, " Cancer diagnosis and prognosis decoded by blood-based circulating MicroRNA signatures,, 4(2013).

[11] K.Zen, C.Y.Zhang, "Circulting micoRNAs: a novel class of biomarkers to diagnose and moniter human cancers", 32(2012),pp.326-348.