Impact of Dimensionality Reduction and Classification in Breast Cancer

Durgalakshmi B, Vijayakumar V

Abstract: ***Breast cancer is the main reason for the female casualty across the world and researchers are aiming to provide a best solution to early diagnosis so that the mortality rate can be reduced. Inorder to understand the problem, the Wisconsin prognostic Breast cancer (WPBC) data set has been obtained from the UCI is utilized for medical research by selecting the best features by correlation matrix. Dimensionality reduction and memory optimization is achieved by suing the feature selection algorithm. Followed with, the classifiers such as support vector classification, logistic regression and random forest is deployed to provide high detection accuracy and reduced error rate. The classifier model thus compares the accuracy with the existing methods and the best classifier model is built. Thus, the efficient model is subjected to the breast cancer cells detection and the improved results provides a major contribution to the early diagnosis of the cancer cells.***

***Index Terms*: Breast cancer, Principle compound analysis,WDBC dataset**

I. INTRODUCTION

Cancer is the deadliest disease and the global cancer burden has been increased to 18.5 million and 9.5 million people were dead in 2018. The increase in the mortality rate is due to various reasons such as population, ageing process, lifestyle associated with socio-economic development. The disease is prevalent in the low economy countries because they are not diagnosed at the early stages of the disease occurrence. The death rate in Asia due to global cancer is 57.3% because they lack in prognosis. Female breast cancer is at the 5th position that leads to 6.6% death rate including the developed countries. Breast cancer contributes around 24.2% of the death rate worldwide and it is prevalent in 154 countries that includes both the developed and developing countries. The statistics of death rate in women due to breast cancer is about 15%, lung cancer is about 13.8% and cervical about 6.6% [1]. Breast cancer affects 2.1 million each year it is the major reason of the deaths among women. It is being understood from the studies that 627,000 women fault Styledeaths were due to breast cancer. The higher rates are more alarming, and it is reported that the death rates have increased in all geographical regions [2]. Diagnosis of breast cancer cells are really challenging in the medical industry because the datasets are unrelated and outmoded. This scenario will reduce the accuracy of the diagnosis report while dealing with massive data. Early diagnosis of the breast cancer cells is important in the field of medical research.

Revised Manuscript Received on June 09, 2019.

Vijayakumar V, currently Associate Dean of the School of Computing Science and Engineering at VIT University;

B. Durgalakshmi, currently pursuing PhD in the School of Computing Science and Engineering at VIT University,chennai, India.

The decision system enhances the detection accuracy of the malignant cells. Several pattern recognition techniques are used for this purpose. In the field of data mining, there is problem called the "Curse of Dimensionality" due to the massive dataset. Due to the presence of less significant and least representative features in the dataset the detection efficiency is reduced at the greater extent. Breast cancer is the most alarming cancer amongst the women in the developing countries. The rapid advancements in the field of genomics has made the dimensionality reduction inevitable for real time analysis of immense data. The medical field relies on the model that has good precision and high classification accuracy. However, most of the recent fields such as bioinformatics and forensic analysis rely on the dataset within higher dimensional space. There is a need for a medical system utilizes the available best machine learning to identify the cancer cells by image processing techniques. Classification techniques are effectively used in the early diagnosis of any misorientation in the structure of the human cells. Dimensionality reduction methods are most popular and extensively used method, because of its inherent capability to put the features with higher dimension into a lower space to achieve dimensionality reduction within less time. Machine learning is branched from artificial intelligence and it comprises of various methods such as statistics, probability and optimization. The system learns from the training dataset and generates pattern from larger datasets for diagnosing the disease based on the test reports. With this thought in mind, this paper is arranged as follows. The following section depicts the related work in the research problem Section 3 illustrates the correlation matrix and classification approaches Section 4 portrays the obtained results and conclusion in section 5

II. RELATED WORK

Support Vector Machine (SVM) is used for selecting the features of the Wisconsin Breast Cancer Dataset (WBCD) for pre-diagnosis of cancer cells. The decision-making model provides greater accuracy of 98.2% where 80% of the records are training data and 20% of the records are testing data. MATLAB is used for deep understanding of the data and threshold levels for the benign and cancer cells are derived. The system is tested for various training scenarios until the maximum prediction accuracy is reached[3]. Inorder to improve the lower performance indices of Random Projection (RP), and enhanced method is proposed that includes other reduction algorithms such as Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Feature Selection (FS).

The accuracy and the run time of the classifier model is studied with various combination of the techniques. The classification efficiency is increased by 14.77% with FS when compared to the existing methods using BC-TCGA dataset. LDA gives 13.65% accuracy and provides extensively high discriminative space [4]. Several machine learning algorithms such as naïve bayes, multi-layer perceptron, radial basis network, nearest neighbour model and conjunctive rule has been used to increase the accuracy. The metrics that are used to verify the creditability of the model are precision, recall, F-measure, ROC, classifier rate and correlation coefficient. The experimental results show that naïve bayes classifier outperforms the existing methods with high detection accuracy and reduced error [5]. The feature selection algorithms extensively utilized are PCA, correlation selection, t-test selection and Random Feature Selection (RFS). RFS provides high prediction accuracy with the high cross validation by selecting the most important features. Logistic regression provides 97.77% accuracy whereas linear SVM and cubic SVM yields 97.87% and 97.98% respectively when coupled with RFS technique. An ensemble model is derived by combining all the models to yield 98.56% detection accuracy[6].Features in the breast cancer dataset are optimized automatically using the Teaching Learning Based Optimization (TLBO) and fitness evaluation is carried out by naive bayes and multi-layer perceptron. Decision tree, random forest, logistic regression is the well-suited classifier models deployed by the proposed system. The experimental results reveals that the suggested system gives high accuracy on WDBC dataset for segregating the malignant and benign tumors. The proposed scheme is effective in optimizing the features and sustainable during the process of decision making of tumour cells [7].The proposed system uses the PCA for feature selection and Artificial Neural Network (ANN) for classification for improving the detection accuracy. Scree Test and cumulative variance are the rule utilized in the PCA. After feature selection, the reduced number of data is passed to back propagation ANN to distinguish the benign and cancer data [8]. The proposed system for feature selection is genetic algorithm .It makes us to understand the most significant parameters for cancer detection. Artificial neural network(ANN), particle swarm optimization and genetic algorithms are utilized to determine the detection accuracy of the classifier models on WDBC and WPBC datasets. Particle swarm optimization outperforms the other classifiers in WDBC dataset. Artificial neural network provides good detection accuracy in both WDBC and WPBC datasets. Hence, feature selection increases the detection accuracy before passing onto the classifier model[9]. Hybrid systems are constructed using the independent component analysis(ICA) with discrete wavelet transform for dimensionality reduction for WDBC dataset. Probabilistic neural network (PNN) is operated to analyse between the benign and malignant cells. The system provides detection efficiency of 96.31% and 98.88% sensitivity. The computational overhead is reduced because the dataset features are reduced before passing to the PNN classifier [10]. The independent component analysis (ICA) is further explored for its adaptability as the decision system for WDBC dataset. The classifier used to verify the classification results are k-nearest neighbor, ANN, RBFNN and SVM. The metrics evaluated are ROC, specificity, sensitivity, detection efficiency and F-measure. The confidence value is about 95% and the system has low computation overhead [11].A hybrid system is proposed with SVM and two-level clustering methods for enhancing the accuracy and reduced error on WBC dataset. By doing so, the patterns for benign and malicious are derived and it differentiates the breast cancer cells from the normal with high detection accuracy of 99.1%. ranking on UCI breast cancer dataset. The results show that efficiency is improved with low computation power because of the usage of dimensionality reduction techniques[13].

III. PROPOSED SYSTEM

The proposed system utilizes the Wisconsin breast cancer datasets and classifies the data with various classifiers as in Figure 1. Before the model is built, the feature selection is made to increase the accuracy of the classifier. When the samples are classified after suitable feature selection algorithms, the efficiency is increased with the reduced error rate. The dimensionality reduction is achieved by the feature selection algorithm. It also provides reduce memory utilization and latency for critical applications. Most of the existing methods have proven appreciable results when the classification is carried out after feature selection. Hence, the input dataset is passed through the correlation based feature selection algorithm and the reduced features are outcome of this phase. The validation of the heat map is successful and then the attributes are passed through the classifier models. The algorithms used are logistic regression, linear SVC and random forest. The comparison of the various classifiers based on the high detection accuracy and the reduced error yields the best classifier and the model is then trained and tested with the breast cancer dataset. The performance of the system is evaluated using various metrics such as accuracy, true positive, false positive, precison, recall and F-measure. In this segment, In this research, the Wisconsin breast cancer datasets is utilized, and it is available from UCI Repository. It is done by Dr. William H. Wolberg at the University of Wisconsin–Madison Hospital. WBC dataset consists of 699 records and 9 attributes with the identification number and class. The dataset has 65.5% malignant and 34.5% benign records. WDBC consists of 569 rows and 30 attributes with the identification number and class.Wisconsin Breast Cancer Dataset (WDBC) that comprises of several attributes such as

- Radius
- Texture
- Area
- Perimeter
- Compactness
- Concavity
- Concave points
- Smoothness
- Symmetry

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A. Feature Selection

Correlation Matrix: The correlation matrix provides the heat-map for the input features. The greater the correlation value between two features, then it is observed as the related and inclusion of one feature is more sufficient. The correlation matrix aggregates of the most closely related features such as radius, perimeter and area and in the same fashion the other features such as concavity and its points could be aggregated. Hence, the feature selection technique provides the reduced attributes such as radius_mean, perimeter_mean, concavity_mean ,area_mean and concave points_mean. The selected features from the heat map can be used for the classification and regression analysis

IV. METHODOLOGY

A. Logistic Regression Logistic regression is a model that is used for predicting the probability of event occurring by fitting the input into logistic curve. The prediction variables are based on either number or category and it is a general linear kind of modelling utilized for binomial regression analysis. For an example, the probability of an individual getting affected by heart disease could be predicted based on the knowledge that comprises of age, sex and BMI. The regression is utilized widely in the field of medical research and social studies. It is also used as the predictive model in marketing to understand the persons interest on purchasing a product or subscription. Logistic regression provides the relationship between the dependency variable and independent variable. It is applied in the cases where the variables are dichotomous. It provides the relationship between the two variables such as predictive variable and the output variable.

$$x = \pi(y) + \varepsilon \quad (1)$$

Where,

y - Vector that comprises $y_i = 1, 2, \ldots, n$ and they are independent predictive variables.

$\pi(y)$ - Conditional probability of occurrence of the event X

ε - Random error term.

$\pi(y)$ and β are given by equation 2 and equation 3.

$$\beta = ln\left(\frac{\pi}{1-\pi}\right) = (\beta_0 + \beta_1 y_1 + \ldots . \beta_n y_n) \quad (2)$$

β- parameter of the regression model

$$\pi(y) = p(x = \frac{1}{y)=} \frac{e^{y^T}\beta}{1+e^{y^T}\beta} \quad (3)$$

The odd ratio provides the effects of x with variation in the odd ratio. The remaining parameters are constant. Larger values of x will aid in ranking the indicators.By measuring the Pearson and deviation metrics, the value of β<0.07 has high degree of confidence. The statistic measure is given by equation

$$X2 = \sum i \ (Obi - Exi)2 / Exi \quad (4)$$

Where,

Obi - observed count

Exi - expected count.

B. Linear SVC

Support vector machine (SVM) finds its application in various fields that involves classification regression analysis of the input samples. It is of two categories such as linear and non-linear. The input samples are epitomized in the space with multi-dimensional characteristics and the data is segregated in a hyperplane with maxima margin. The higher the margin, the lesser is the error rate. Since, SVM provides high efficiency with reduced error it is mostly applied in the real time applications. Support vector classification (SVC) are of two kinds such as linear and n-SVC. Linear model is effective when the data samples can be segregated in the linear mode. The discriminant function is given by equation 6.

$g(x) = w^T x + b \quad (6)$

$g(x)$ – concludes the class

w-coefficient vector

x- data point

b- offset

For a linear model, $g(x) \geq 0$ for a class 1 and $g(x) < 0$ for class 2. The quadratic problems are applied in determining the classifier in a normalized form. The area (2/||w||) must have the maximum value. The independent/dependent variables are influenced by the determinant function along including the noise. It is given by equation 7

Min Φ (w) = ½ || w || 2 = ½ (w, w), Such that: g_i (w. xi + b) >= 1 (7)

The n-SVM, the error is made minimum by the constraints by equation 8

$$½ \ w^T w - \nu\rho + 1/N \sum_{i=1}^{N} \varepsilon_i \quad (8)$$

Where,

ε_i – operates on non-seperable data

ν- observed value

C. Random Forest

Decision tree is a kind of supervised algorithm utilized as the best model for classification and regression. The rule induction technique is used for building the tree and to the target can be labelled. Set are rules are designed to create, train and test the model with the help of rule induction. It is represented as the ensemble of the classifier and regression model. The algorithm is utilized on the larger sets of decision tree and gives the target one with the classification/regression with each set of trees. The bagging is that operates on the randomly selected set of features. Extra tree is also much like the random forest since it is the ensemble of each tree. The cut is made randomly inspite of using local optima and the sub trees are trained with the whole set of samples. Information gain is the deciding parameter to determine the local cut and it is random in nature.

The higher value is chosen that makes the split very good and the default value is being chosen for the root nodes. Thus various trees are created from the sub trees given from the sample input and it provides an average and reduces the variance and overfitting issues. The gini impurity metric is utilized to measure the split quality and is given by the equation 9.

$$GI = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (9)$$

p_i – probability of selecting the datapoint from the class

$_k$- total class*k* denotes the labels which is either 0 or 1 and p_i denotes the likelihood of labelling as *k*. This impurity metric provides the measure when the labels are chosen randomly based on the subsets distribution. Whenever a new sample enters the model, it would execute deep down the trees and arrive at the leaf node of every individual tree. The outcome is determined by the majority vote of the leafs.

V. RESULTS AND DISCUSSIONS

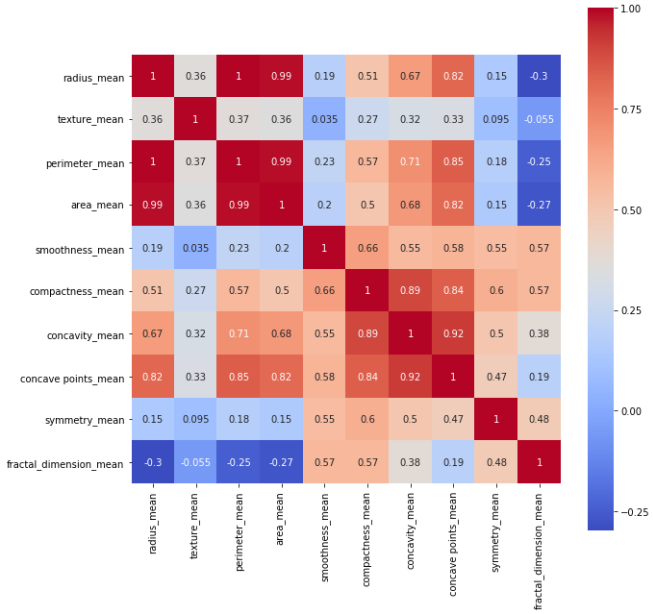

Fig 1. Correlation Heat Map

The efficiency of the system is justified by the confusion matrix and the model is more suitable for classifying the benign and cancer cells. the correlation between the attributes are studied to understand whether the cells are malignant or not. the proposed is utilized for arriving at the conclusion of the predictors that provides good classification result.Precision provides the part of the extracted information from the dataset based on the searching criteria. Precision shows the number of instances that has classified correctly in confusion matrix. It is the ratio between the true positive (TP) divided by the sum of true positive and false positive (FP). Recall provides the part of the extracted information from the dataset based on the searching criteria where it is successful. It is the ratio between the true positive and sum of true positive and false positive. F-measure provides the mean of both the precision and recall. Accuracy is the ration of the true positive and true negative and the results obtained from the data.

Precision= TP/TP+TN
Recall= TP/TP+TN
F-measure=2*Precision*Recall/Precision+Recall
Accuracy=TP+TN/TP+FP+FN+TN

	Accuracy	precision	Recall	F-Measure
LOG-REGRSSION	0.73	0.72	0.61	0.60
LINEAR-SVC	0.71	0.65	0.55	0.44
RANDOM FOREST	0.92	0.93	0.90	0.91

Table 1.performance analysis

Hence the high accuracy is achieves by random-forest method .it proves that it will seperate the benign classes and malignant cases. So that after identifying the cancer ,these results helps us to gor the treatment.

VI CONCLUSION

This work efficiently removes the insignificant features from the massive scale of dataset using the correlation matrix method. It indirectly improves the classification accuracy. after identifying the relevant features, the classification model has been built .The proposed method improves the accuracy. so that computation cost is also compared to other methods. In near future we can further improve by identifying dependency factors between the featues and observe the changes in the classification model. since big data plays a vital role in these days, parallelization can be achieved in these model .so that accuracy and computational cost may be further improved.

REFERENCES

1. RLatest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018, online: https://www.who.int/cancer/PRGlobocanFinal.pdf
2. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
3. Vikhyat Srivastava,Ritika Choudhury, L.Shanmukhi Jyotirmayee Dash IA Classifier Model For Diagnosis of Benign/Malignant Breast Cancer Tumors International Journal of Pure and Applied Mathematics ,Volume 118 No. 20 2018, 3497-3500
4. KHaozhe Xi,Jie Lia,Qiaosheng Zhanga,Yadong W,Comparison among dimensionality reduction techniques based on Random Projection for cancer classification,Preprint submitted to Computational Biology and Chemistry
5. B.Tamilvanan,V.Murali baskaran,An efficient classifications model for breast cancer prediction based on dimensionality reduction techniques. International Journal of Advanced Research in Computer Science,Volume 9, No. 1, January-February 2018
6. Todor K. Avramov,Dong si, Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis,ICCDA '17, May 19-23, 2017,
7. Mohan Allam, M. Nandhini,Feature Optimization using Teaching Learning Based Optimization for Breast Disease,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4, November 2018
8. Hasmarina Hasan, Nooritawati Md Tahir,Feature Selection of Breast Cancer Based on Principal Component Analysis,2010 6th international colloqium on signal processing and its applications.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

9. OShokoufeh Aalaei, Hadi Shahraki , Alireza Rowhanimanesh, Saeid Eslami,Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets,Iran J Basic Med Sci, Vol. 19, No. 5, May 2016.
10. Ahmet Mert,Niyazi KJlJç,Erdem Bilgili, and Aydin Akan,Breast Cancer Detection with Reduced Feature Set,Computational and Mathematical Methods in Medicine Volume 2015, Article ID 265138, 11 pages
11. Ahmet Mert, Niyazi K l ç and Ayd n Akan,An Improved Hybrid Feature Reduction for Increased Breast Cancer Diagnostic Performance,Biomed Eng Lett (2014) 4:285-291
12. Ahmed Hamza Osman,An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique,(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 8, No. 4, 2017
13. Nitika Sharma, Kriti Saroha,A Novel Dimensionality Reduction Method for Cancer Dataset using PCA and Feature Ranking,978-1-4799-8792-4/15/$31.00 c 2015 IEEE

AUTHORS PROFILE

B.Durgalakshmi is currently pursuing PhD in the School of Computing Science and Engineering at VIT University,chennai, India. She received M.Tech at SRM University Chennai in 2013 in Database systems and B.Tech in Information Technology from AVC College of Eng ineering in 2011. Her current research interests include smart care health systems, Machine learning.

Vijayakumar V is currently Associate Dean of the School of Computing Science and Engineering at VIT University; He has more than 15 years of experience which includes 9 years in teaching and 6 years in Industry. He is also a Division Chair of Cloud Computing Research Group; His area of research includes Grid Computing, Cloud computing, Big Data, Web semantics and also involved in the domain like Bio-medical applications-Mammogram, Autism, Immune system, and Vehicular Adhoc Networks .