Syed Seema Anjum et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Special Issue: Conscientious Computing Technologies, April 2018, pp. 179-185

Prediction of Breast Cancer Using a Supervised Learning Approach

Syed Seema Anjum[1], V. Sujatha[2]

[1]M. tech Scholar, Computer Science & Engineering, Vignan's Nirula Institute of technology & Science for Woman, Pedapalakaluru Guntur, Andhra Pradesh, India
[2]Assistant Professor, Computer Science & Engineering, Vignan's Nirula Institute of technology & Science for Woman, Pedapalakaluru Guntur, Andhra Pradesh, India

Abstract: Breast cancer is the second most driving disease happening in ladies contrasted with every single other growth. Watched rates of this malignancy increment with industrialization and urbanization and furthermore with offices for early recognition. In underdeveloped, developing and developed nations, the major cause of death among women is due to breast cancer. Breast cancer is fatal in under portion of all cases and is the main source of death from disease in ladies, representing 16% of all tumor passing around the world. The goal of this paper is to show a provide details regarding breast cancer where we exploited those accessible innovative progressions to create prediction models for breast cancer survivability We utilized Logistic Regression to build up the prediction models using an expansive dataset (699 breast cancer cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of this prediction model for getting better performance. The results (based on average accuracy Breast Cancer dataset) indicated that the Logistic Regression is the best predictor with 96.48% accuracy on the holdout sample.

Keywords: Breast cancer, Logistic regression, Prediction

1. Introduction

The number and the span of databases recording therapeutic information are expanding quickly. Medicinal information, created from estimations, examinations, solutions, and so on. , are put away in various databases consistently. This colossal measure of information surpasses the capacity of customary strategies to investigate and look for intriguing examples and data that is covered up in them. Accordingly, new systems and instruments for finding helpful data in these information storehouses are ending up more demanding. Analysing these information with new investigative techniques keeping in mind the end goal to discover fascinating examples and shrouded learning is the initial phase in expanding the conventional capacity of these information sources.

The organs and tissues of the body are comprised of little building squares called cells. Tumor is a malady of these cells. In spite of the fact that cells in each piece of the body may look and work in an unexpected way, most repair and imitate themselves similarly. Regularly, cells isolate in an efficient and controlled way. Be that as it may, if for reasons unknown the procedure gains out of power, the phones bear on partitioning and form into a knot called a tumor. Bosom tumors are typically caused by an excess of the phones coating the bosom pipes. They can be either generous or malignant [1]. In a generous tumor, the cells develop anomalous and shape a knot. Yet, they don't spread to different parts of the body as are not malignancies. The most widely recognized sort of considerate bosom tumor is known as a fibroadenoma. This may should be carefully evacuated to affirm the conclusion. No other treatment is fundamental. In a threatening tumor, the growth cells can spread past the bosom in the event that they are left untreated. For instance, if a harmful tumor in the bosom isn't dealt with, it might develop into the muscles that lie under the bosom. It can likewise develop into the skin covering the bosom. In some cases cells split far from the first (essential) malignancy and spread to different organs in the body. They can spread through the circulation system or lymphatic framework. At the point when these cells achieve another territory they may continue separating and shape another tumor. The new tumor is regularly called an auxiliary or metastasis. Bosom malignancy happens when cells inside the bosom conduits and lobules end up dangerous. On the off chance that got at a beginning period, bosom malignancy can frequently be cured. In the event that the tumor has spread to different regions of the body it can't more often than not be cured, however it can typically be viably controlled for quite a while.

A less obtrusive screening strategy is the Fine Needle Aspiration (FNA) cytology of bosom masses. It is viewed as sheltered and has a high affectability to recognize a benevolent tumor and a threatening one, being exceptionally helpful in diminishing the false positives. The FNA technique requires visual elucidation by a specialist to give analysis or guess, however exactness has a high variety [2].

Consequently, Machine Learning (ML) strategies have been recommended to help in the clinical choice process [2], [3]. At the point when computer-aided detection/diagnosis (CAD) frameworks are utilized, precision can be higher, with less difference and, above all, help medicinal specialists sparing lives. Along these lines, explore on such CAD frameworks is of incredible significance for wellbeing administrators.

In this work, we assess the grouping exactness on the Wisconsin Diagnostics Breast Cancer (WDBC) dataset. This double order dataset, arranged by William H. Wolberg at University of Wisconsin [2], contains perceptions of 569 patients among which 357 are considerate cases, and 212 are harmful cases, being portrayed by 30 numerical highlights. These highlights depict qualities of the cell cores in advanced pictures of FNA of bosom masses. Mangasarian et al. [2] utilized the Multi surface Method, Recurrence Surface Approximation, and an altered Back-spread Artificial Neural Network (ANN) to naturally group a patient's information into generous or threatening case accomplishing a satisfactory order precision. Such dataset has been examined by numerous analysts with numerous unmistakable strategies with fluctuated comes about [4].

For the most recent decades, Evolutionary Computation (EC) calculations, all the more particularly techniques, for example, the Genetic Programming (GP) [4]– [6] calculation that can be utilized as hyper-heuristics, have been utilized to create lead or tree based classifiers that could be deciphered by people, rather than the end result for ANNs. Another intriguing utilization of EC calculations is to scan for helpful changes of the dataset or to build new highlights that are mixes of the first highlights of the dataset. These new highlights can be either utilized alone (dimensionality diminishment) or to grow the first dataset.

Highlight development is the errand of consolidating a few highlights of the first dataset utilizing, for example, mathematical, geometric, and boolean tasks into a solitary new component. In the restorative space this might be comprehended as the making of a list. As a straightforward illustration, consider the Body Mass Index which is an esteem got from the mass and stature of a man: BMI =masskg/ (height2 m). Such errand should be possible physically by a scientist, however the time has come devouring and likely one-sided by the analyst's experience learning, abstaining from attempting files that don't "bode well". Then again, as EC calculations can attempt a less one-sided investigation of the hunt space, they may find records that permit incredible precision yet are not yet reasonable by the scientist. For a similar reason, they for the most part require the assessment of a substantial number of trial arrangements until the point when a decent one is found. Likewise, arrangements can get exceptionally mind boggling if no control strategy is utilized, making them totally uninterpretable. Numerous exploration has been done to decrease such issues accomplishing great outcomes, however it is as yet an open issue. By the by, EC calculations have been generally utilized and extremely effective for highlight development [5]. Each lady needs to comprehend what she can do to bring down her danger of bosom growth [7].

2. Related work

A few examinations have been accounted for that have concentrated on bosom disease survivals. These examinations have connected distinctive ways to deal with the given issue and accomplished high arrangement exactnesses. Points of interest of a portion of the past research works are given in the accompanying:

Decision table (DT)-based predictive models for breast cancer survivability, concluding that the survival rate of patients was 86.52%. They employed the under-sampling C5 technique and bagging algorithm [8] to deal with the imbalanced problem, thus improving the predictive performance on breast cancer.

The neural network classifier is used on Wiscosin Prognosis Breast Cancer (WPBC) by using MLP algorithm to get the accuracy of 70.725% [9]. Tan and Gilbert4 demonstrated the usefulness of employing ensemble methods in classifying microarray data and presented some theoretical explanations on the performance of ensemble methods. As a result, they suggest that ensemble machine learning should be considered for the task of classifying gene expression data for cancerous samples.

The execution rule of administered learning classifiers is looked at ,, for example, Naı¨ve Bayes, SVM-RBF portion, RBF neural systems, Decision Tree (Dt) (J48), and basic classification and regression tree (CART), to locate the best classifier in bosom tumor datasets. The test result demonstrates that SVM-RBF piece is more exact than different classifiers. What's more, three mainstream information mining calculations: CART, ID3 (iterative dichotomized 3), and DT for diagnosing heart sicknesses, and the outcomes displayed showed that CART acquired higher exactness inside less time are inferred [10-11].

An examination is directed to recognize the most widely recognized information mining calculations, actualized in current Medical Diagnosis, and assess their execution on a few medicinal datasets. Five algorithms were chosen: Naı¨ve Bayes, RBF Network, Simple Logistic, J48 and Decision Tree [12]. For the evaluation two Irvine Machine Learning Chaurasia and Pal 3 Repository (UCI-UC) databases were used: heart disease and breast cancer datasets. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, area under the curve (AUC), precision, recall, F-measure, and a set of errors.

Numerous broadened and huge tenets from high-dimensional profiling information and proposed collection of the segregating energy of these guidelines for solid expectations are found. The found guidelines are found to contain low-positioned highlights; these highlights are observed to be now and again fundamental for classifiers to accomplish culminate exactness [13].

A new classification algorithm tree bagging and weighted clustering (TBWC) combination of decision tree with bagging and clustering [14] This algorithm is experimented on two medical datasets: cardiocography1, cardiocography2 and other datasets not related to medical domain.

Another choice tree-based gathering technique joined with include determination technique in reverse disposal system with sacking is proposed [15] to discover the structure action connections in the territory of chemo metrics identified with pharmaceutical industry.

3. Methodology

This paper uses popular algorithm Logistic regression on breast cancer dataset, Logistic regression is a popular method to predict a binary response. It is a special case of Generalized Linear models that predicts the probability of the outcome, that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable .a logistic regression produces a logistic curve, which is limited to values between 0 and 1.The Logistic regression curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group

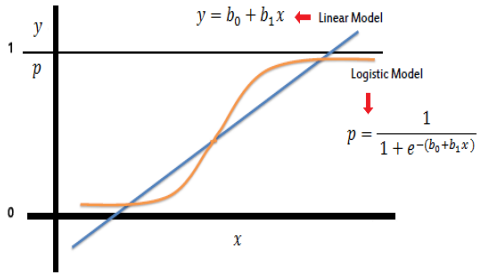

Figure:1

Performance of a Logistic Regression Model

To evaluate the performance of a logistic regression model, we must consider a few metrics irrespective of tool (SAS, R , Python).

1. AIC (Akaika information criteria): The analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore we always prefer model with minimum AIC value.
2. Null Deviance and Residual Deviance: Null Deviance indicates the response the response predicted by model with nothing but an intercept. Lower the value better the model. Residual Deviance indicates the response predicted by model by adding the independent variables. Lower the value, better the model.
3. Confusion matrix: It is nothing but the actual vs predicted value. This help to find the accuracy of the model and avoid overfitting. It is how it looks.

		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

The accuracy of a model is calculate by using the following formula

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

From confusion matrix specificity and sensitivity can be calculated as below

$$\left.\begin{aligned}\text{True Negative Rate (TNR), specificity} &= \frac{A}{A+B}\\ \text{False Positve Rate (FPR), } 1-\text{specificity} &= \frac{B}{A+B}\end{aligned}\right\}\text{sum to 1}$$

$$\left.\begin{aligned}\text{True Positive Rate (TPR), sensitivity} &= \frac{D}{C+D}\\ \text{False Negative Rate (FNR)} &= \frac{C}{C+D}\end{aligned}\right\}\text{sum to 1}$$

Specificity and Sensitivity plays a crucial role in deriving ROC curve.

4. ROC curve: Receiver Operating characteristics (ROC) summarizes the model's performance by evaluating the trade-offs between true positive (sensitivity) rates and False positive (1-Specificity). For plotting ROC, it is advisable to assume P > 0.5 due to success rates. ROC summarizes the predictive power of all positive values of P > 0.5. The area under curve (AUC), referred to as index of accuracy (A) or concordance index. It is a perfect performance matric of ROC

curve. The ROC of a perfect predictive model has TP as 1 and FP as 0. This curve will touch the top left corner as shown in figure2.

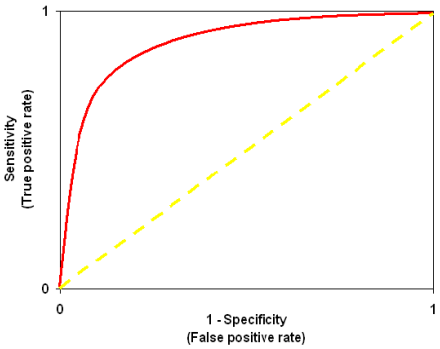

Figure2: ROC curve

The given algorithm is implemented on cancer dataset to predict the outcomes

Algorithm 1 Logistic Regression with L1 regularization

Procedure STOCHASTICGRADIENTDECENT (D, Labels, *I*ter)
Input: *Dataset D, Labels of Dataset, Iteration num*
Output: optimal weight of logistic regression
w ← [1, 1, …, 1]
Initialize q_i with zero for all i
For k = 1 → *I*ter **do**
chooseData = D
  for I = 1→ m **do**
    γ ← Learning Rate
    λ ← Regulation Lambda
    ***u* = *u* + γλ**
    Select a index of chooseData *idx* randomly
    x ← chooseData[*idx*]
    del chooseData[*idx*]
    for i € *featuresinsamplex* **do)**
      $w_i = w_i - \gamma \frac{\partial loss(w,xi)}{\partial w}$
      wh ← *w*
      if $w_i > 0$ **then**
        $w_i \leftarrow \max(0, w_i - (u + q_i))$
      else if $w_i < 0$ **then**
        $w_i \leftarrow \max(0, w_i + (u - q_i))$
      end if
      $q_i \leftarrow q_i + (w_i - wh)$
    end for
  end for
end for

4. Feature Extraction and Selection

An essential advance in breast cancer analysis model is Feature extraction. The Optimum list of capabilities ought to have powerful and separating highlights, while generally lessen the repetition of feature space to avoid "curse of dimensionality" issue. The "curse of dimensionality" proposes that the examining thickness of the preparation information is as well low to guarantee an important estimation of a high dimensional characterization work with the accessible limited number of preparing information.

The given table lists Breast cancer Dataset Attributes and its Values

Table1: Breast cancer Dataset Attributes

Attribute	Domain
1. Sample code number Id	
2. Clump thickness	1-10
3. Uniformity of cell size	1-10
4. Uniformity of cell shape	1-10
5. Marginal adhesion	1-10
6. Single epithelial cell size	1-10
7. Bare nuclei	1-10
8. Bland Chromatin	1-10
9. Normal nucleoli	1-10
10. Mitoses	1-10
11. Class	2 for benign 4 for malignant

In the Clump thickness generous cells have a tendency to be gathered in monolayers, while destructive cells are regularly assembled in multilayer. While in the Uniformity of cell size/shape the growth cells have a tendency to differ fit as a fiddle. That is the reason these parameters are important in deciding if the cells are carcinogenic or not. On account of Marginal adhesion the ordinary cells tend to stick together, where tumor cells have a tendency to lose this capacity. So loss of adhesion is a sign of malignancy. In the Single epithelial cell estimate the size is identified with the consistency specified previously. Epithelial cells that are essentially amplified might be a threatening cell. The Bare nuclei is a term utilized for cores that isn't encompassed by cytoplasm (whatever is left of the cell). Those are normally observed in generous tumors. The Bland Chromatin portrays a uniform "surface" of the core seen in considerate cells. In disease cells the chromatin has a tendency to be coarser. The Normal nucleoli are little structures found in the core. In typical cells the nucleolus is normally little if obvious. In tumor cells the nucleoli turn out to be more noticeable, and in some cases there are a greater amount of them. At long last, Mitoses is atomic division in addition to cytokines and deliver two indistinguishable little girl cells amid prophase. It is the procedure in which the cell partitions and repeats. Pathologists can decide the review of tumor by checking the quantity of mitoses.

5. Experimental Results

To evaluate the effectiveness of our method, experiments on Wisconsin Diagnostic Breast Cancer (WDBC) are conducted. These databases were obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg. This is freely accessible dataset in the Internet. Table 2 demonstrates the portrayal of WDBC database.

Table2:

No	Attributes	No of Attributes
1.	Number of instances	699
2.	Number of Attributes	10
3.	Attribute 2 through 10	instances
4.	Classes	1 for benign 2 for malignant
5.	Class distribution	1. Benign:444 2. Malignant:239

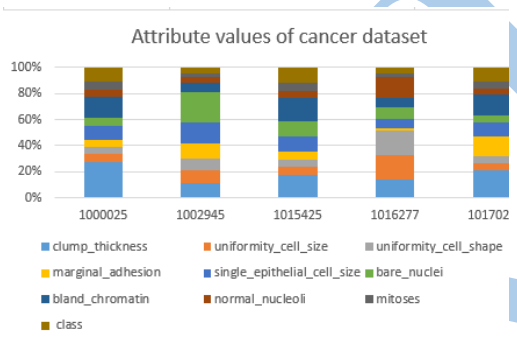

Figure3: Attribute values of cancer dataset

Figure3 shows the attributes used for classification of benign and malignant cases from breast cancer dataset. There are total 10 number of attributes.

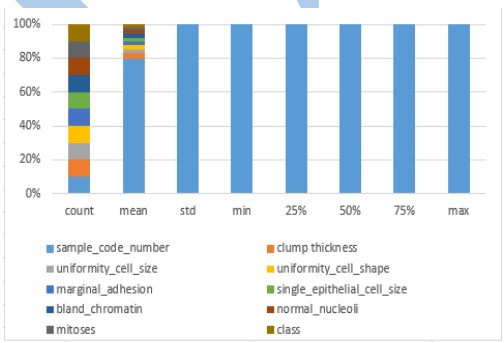

Figure4: Different cell types may expose to cancer.

Figure4 shows represents the defferent cell types may expose to cancer. The missing values from dataset are calculated by using the mean,std of current attributes.

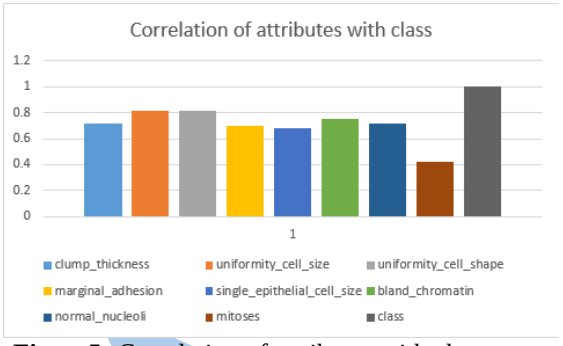

Figure5: Correlation of attributes with class

Figure5 represents the correlation of attributes with their class.Here we find correlation for each attribute.

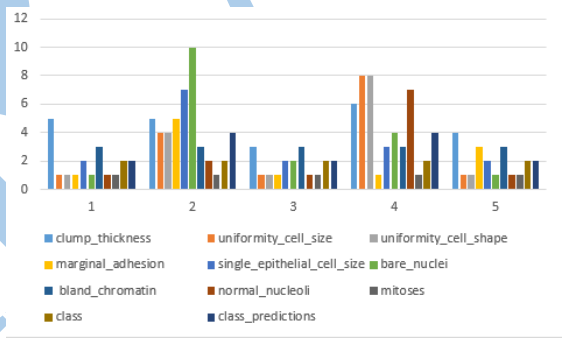

Figure6: Cell Classes

Figure6 represents the cell classes of each attribute. Here we predict the cell classes for each attribute.

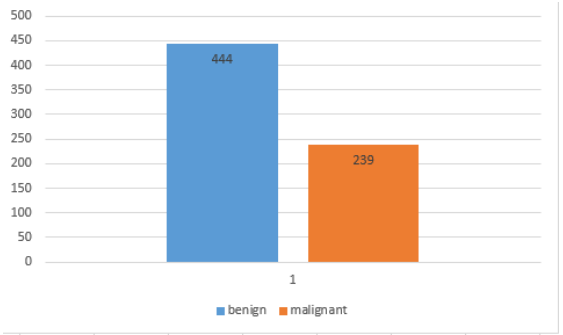

Figure7: Classification of number of cancer cases

Figure7 shows the classification of breast cancer. Out of 699 cases we found 444 cases as benign and 239 cases as malignant.

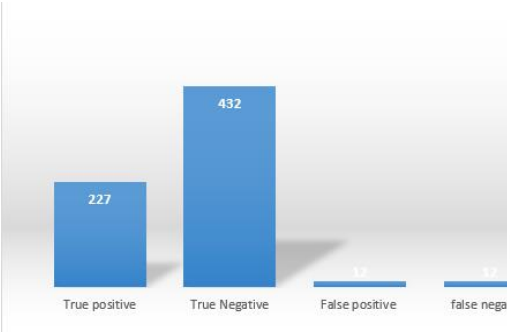

Figure8: Classification of cancer conformation based on examination data

Figure8 shows the classification of canser conformation based on examination data. Here we found 227 True positive cases, 432 True negative cases, 12 asFalse positive cases and 12 as False negative cases out of 699 cases.

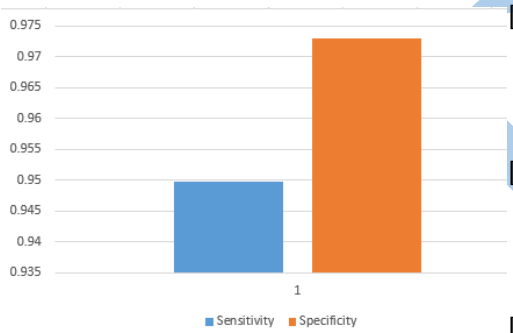

Figure9: Accuracy difference of cancer

Figure9 shows the accuracy differences of cancer . This paper results the sencitivity as 94.97% and Specificity as 97.29%.

6. Conclusion

In this paper, we connected a prediction display for breast cancer survivability on two parameters: benign and malignant disease patients by utilizing Logistic Regression. We gained a dataset (699 occasions) from the UCI Machine Learning storehouse. We applied data selection, pre-processing, and transformation to build up the prediction model. In this research, we utilized a binary categorical survival variable, which was ascertained from the variables in the crude dataset, to speak to the survivability where malignant is spoken to with an estimation of "1" and benign is spoken to with "0". With a specific end goal to gauge the impartial forecast precision of the strategy, we utilized a 10-fold cross-approval system, that is we separated the dataset into 10 totally unrelated parcels (k-folds) utilizing a stratified inspecting method. This gave us a less one-sided forecast execution measures. The obtained results demonstrated that the Logistic Regression played out the best with a classification precision of 96.48%. In addition to the prediction model, we also conducted sensitivity analysis and specificity analysis on Logistic Regression in order to gain insight into the relative contribution of the independent variables to predict survivability. The sensitivity comes about demonstrated that the visualization factor "Class" is by a wide margin the most critical indicator.

REFERENCES

[1]. R. Ariga, K. Bloom, V. B. Reddy, L. Kluskens, D. Francescatti, K. Dowlat, P. Siziopikou, In addition to the prediction model, we also conducted sensitivity analysis and specificity analysis on Naı¨ve Bayes, RBF and P. Gattuso, “Fine-needle aspiration of clinically suspicious palpable breast masses with histopathologic correlation,” The American journal of surgery, vol. 184, no. 5, pp. 410–413, 2002.

[2]. O. L. Mangasarian, W. N. Street, and W. H. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” Operations Research, vol. 43, no. 4, pp. 570–577, 1995.

[3]. W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, “Computerized breast cancer diagnosis and prognosis from fine-needle aspirates,” Archives of Surgery, vol. 130, no. 5, pp. 511–516, 1995.

[4]. H.-L. Wei, S. Billings et al., “Feature subset selection and ranking for data dimensionality reduction,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 1, pp. 162–166, 2007

[5]. A. Cano, S. Ventura, and K. J. Cios, “Multi-objective genetic programming for feature extraction and data visualization,” Soft Computing, pp. 1–21, 2015.

[6]. P. Wang, K. Tang, T. Weise, E. Tsang, and X. Yao, “Multiobjective genetic programming for maximizing roc performance,” Neurocomputing, vol. 125, pp. 102–118, 2014.

[7]. www.breastcancer.org/risk/factors (accessed 12 January 2018).

[8]. Liu Y-Q, Wang C and Zhang L. Decision tree based predictive models for breast cancer survivability on imbalanced data. In: 3rd international conference on bioinformatics and biomedical engineering, 11-13 June 2009, Beijing, China, 2009.

[9]. Vaibhav Narayan Chunekar, Hemant P. Ambulgekar (2009). Approach of Neural Network to Diagnose Breast Cancer on three different Data Set. 2009 International Conference on Advances in Recent Technologies in Communication and Computing.

[10]. Kaewchinporn C, Vongsuchoto N and Srisawat A. A combination of decision tree learning and clustering for data classification. In: 2011 eighth international joint conference on computer science and software engineering (JCSSE), MAY 11-13, 2011, Faculty of ICT, Mahidol University, Nakhon Pathom,THAILAND .

[11]. Vikas C and Pal S. Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease. Rev Res 2014; 3: 1–13.
[12]. Li J, Liu H, Ng S-K, et al. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 2003; 19: ii93–ii102.
[13]. Chaurasia V and Pal S. Data mining techniques: to predict and resolve breast cancer survivability. Int J Comput Sci Mobile Comput 2014; 3: 10–22.
[14]. Chaurasia V and Pal S. A novel approach for breast cancer detection using data mining techniques. Int J Innovative Res Comput Commun Eng 2014; 2: 2456–2465
[15]. Cao D-S, Xu Q-S ,Liang Y-Z, et al. Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. Chemometr Intell Lab Syst 2010; 103: 129–136.

