Performance Analysis of Regression Data Mining Techniques Implemented on Breast Cancer Dataset

Ritu Tayal
Student M Tech (CSE)
Amity University, Sector-125, Noida, India

Anshul Tickoo
Assistant Professor,
ASE, Amity University, Noida, India

Abstract— Regression is a data mining function that helps in predicting a number. Regression analysis determines the values of parameters for a function that causes the function to best fit a set of data observation. In this paper, we aim to compare performance of Regression Data Mining Techniques implemented on Breast Cancer data Set. Breast Cancer is the third most reason of death from cancer in women of all ages of death from cancer. The diagnosis and treatment of breast cancer has become a serious women's health issue in whole Asian-American region. The development of breast cancer is associated with various factors like age, ethnicity and family history of cancer in woman and many more. In this paper, seven regression techniques are analysed and implemented on Breast Cancer dataset to know best regression technique among those compared.
Keywords- Data Mining, Knowledge Discovery, Regression, Weka, Breast cancer, Regression performance.

I. INTRODUCTION

Regression analysis establishes a relationship between a dependent and a set of predictors. Regression, as a data mining technique, is a form of supervised learning. Supervised learning divides the database into training and validation data.

Regression analysis determines the values of parameters for a function that causes the function to best fit a set of data observations that are provided by you. The following equation expresses these relationships in symbols and shows that regression estimates the value of a continuous target (y) as a function (F) of one or more predictors (x1, x2... xn), a set of parameters (θ1, θ2... θn), and a measure of error (e).

$$y = F(x,\theta) + e$$

The predictors in regression are independent variables and the target is a dependent variable. Error in this equation, also called as residual, is difference between expected and predicted value of the dependent variable. These regression parameters are also known as regression coefficients.

A regression algorithm helps in estimating value of target as a function of the predictors for each case in the build data. The relationships between predictors and target are summarized in a model, which can be applied to a different data set in which these target values are unknown. [1, 2]

In this paper, we have carried out performance analysis of various regression techniques on Breast Cancer dataset. Analysis in this paper will help in determining best regression techniques based on its correctly classified instances and relative absolute error. For this we have taken Breast Cancer data set from UCI repository. The analysed regression techniques are: Linear Regression, Pace Regression, Simple Linear Regression, and Regression by Discretization, Isotonic Regression, Logistic Regression and Simple Logistic Regression.

Breast Cancer

Breast cancer is one of the most dangerous and common cancer in women. Breast cancer occurs in situation when there is an uncontrolled growth of cancer cells that form a benign/malignant tumour. Breast cancer leading cause of death among women ages 40-60[3,4]. Though predominantly in women, breast cancer can also lead to death in men [5]. In the United States, of the 40,600 death from breast cancer in 2001,400 men were suffering from breast cancer[6].

Currently there are three methods used for the diagnosis of breast cancer, i.e. mammography, FNA (fine needle aspirate) and surgical biopsy. The diagnosis accuracy of mammography is from 68% to 79%, whereas the accuracy of FNA is inconsistent and varies from 65% to 98%, with accuracy of surgical biopsy as 100%. The process of surgical biopsy, however, is both tedious and costly [7]. Breast cancer diagnosis is categorized into

classes Benign and Malignant. Benign lumps are abnormal lumps but not cancerous while Malignant lumps are cancerous lumps in female body.

Although cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common compliment. [11] In the medical domain where data and statistics driven research is successfully applied, new and innovative research directions are identified for further clinical and biological research [8].Predicting outcome of a disease is one of the most innovative and challenging task to develop data mining applications. With the increased use of systems powered with automated tools, storage and retrieval of large volume of medical data are being collected and is being made available to the medical research community interested in predicting models for survivability. As a result, new research areas such as knowledge discovery in databases(KDD),whose data mining techniques, has become popular research tool for medical researchers who are looking for identifying and exploiting patterns & relationship among large number of variables, and predict the outcome of a disease using the historical cases stored within datasets[5,9,10].

II. KNOWLEDGE DISCOVERY AND DATA MINING

A. Data Mining

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably but in real world, data mining plays vital role in the KDD process. Data mining (also known as knowledge discovery in database) can be defined as the non trivial extraction of implicit, previously unknown and useful information from datasets. Data mining involves the use of sophisticated data analysis tools to discover and analyse previously unknown, valid pattern and relationships in big datasets that can further help in decision making process.[12] Data mining uses machine learning, various statistical and visualization techniques to discover and present knowledge in a form which can be easily understood. These tools can include statistical models, mathematical algorithms and machine learning methods and various other methods [1] (algorithms that improve their performance automatically through experience such as KNN, neural network or decision tree) consequently; data mining is plays vital role in management and collection of data.

Data mining can be performed on any data that can be represented in quantities, textual or multimedia form over a period of time. Data mining applications can use n number of parameters with different attributes to examine the data, which can include association (patterns which are interconnected) sequence or path analysis (patterns where one event lead to another event), classification (identifications of new patterns), clustering (finding and visually documenting group of previous unknown facts) and forecasting (discovering patterns from which one can make predictions regarding future activities). Data mining can be successfully applied in a wide range of applications, including finance, telecommunication, banking, biomedicine, stock exchange, medical diagnostics and many more.

Regression technique is an attempt to find a function which provides the data with the least error. Regression techniques help in examining quality of data and gives a mathematical formula using coefficient to predict the result from given data set. There are various families of regression functions and different ways of measuring the error.

1) Linear Regression

A linear regression technique can be used if the relationship between the predictors and the target can be approximated with a straight line. [1] Regression with a single predictor is the easiest way to visualize data. Simple linear regression with a single predictor is shown in Figure 1.

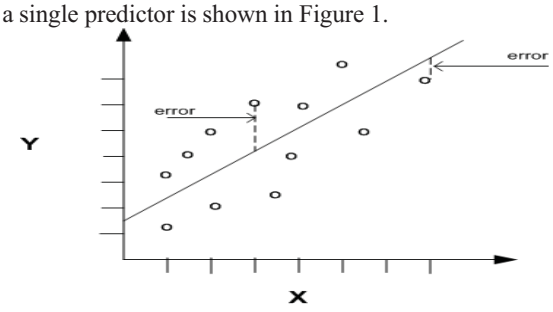

Fig 1 Linear Regression with a Single Predictor

Linear regression with a single predictor can be expressed with the following equation.

$$y = \theta 2x + \theta 1 + e$$

The regression parameters in simple linear regression are: 1) Slope of the line (θ2) — the angle between a data point and the regression line. 2) The y intercept (θ1) — the point where x crosses the y axis (x = 0)

2) Multiple Linear Regression

In multiple linear regressions, there are p explanatory variables, and one dependent variable whose values has to be calculated. The relationship between the dependent variable and the explanatory variables is represented by the following equation, where:

β0 is the constant term and

β1 to βp are the coefficients relating the p explanatory variables to the variables of interest.[13]

Hence, multiple linear regression is an extension of simple linear regression, where there are p explanatory variables, or simple linear regression can be thought of as a specific case of multiple linear regression, where p=1. The term 'linear' is used because in multiple linear regressions we assume that y (i.e. dependent variable) is directly related to a linear combination of the explanatory variables.

Correlation Linear Regression Theory

The correlation coefficient, r, is calculated using:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Where,

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Is the variance of x from the sample, which is of size n

$$Var(Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Is the variance of y, and,

$$Var(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

is the covariance of x and y.

3) Non Linear Regression:

Non linear regression is a type of regression analysis where observational data are modelled by a function which is a nonlinear combination of the model parameters and dependent on one or more independent variables. The data can be fitted by a method of successive approximations. The data will consist of error-free independent variables (explanatory variables), x, with their associated observed dependent variables (response variables), y. Each dependent variable y is modelled as a random variable with a mean given by a nonlinear function f(x, β). Systematic error can be present but can be treated outside the scope of regression analysis. If the independent variables are posses some error, then there will be an errors-in-variables model. For example, the Michaelis–Menten model for enzyme kinetics

$$v = \frac{V_{max}[S]}{K_m + [S]}$$

can be written as

$$f(x,\beta) = \frac{\beta_1 x}{\beta_2 + x}$$

where β_1 is the parameter V_{max}, β_2 is the parameter K_m and [S] is the independent variable, x. This function is nonlinear because it cannot be expressed as a linear combination of the βs

4) Logistic Regression: Logistic regression is a generalized mechanism of liner regression [10], which is currently being used primarily for predicting or multi-class dependent variables. As the response variables are discrete, logistic regression cannot be modelled directly by liner regression. Therefore, rather than predicting

estimate of the event itself, it builds the model to predict the odds of its occurrence. Though logistic regression is a very powerful modelling tool, it assumes that the response variables (the log odds, not the event the itself) are linear in the coefficients of the predictor variables.

III. LITERATURE REVIEW

This section gives brief description of various review and technical articles on data mining regression techniques.

Delen et al.[15] compared ANN, decision tree and logistic regression techniques for breast cancer survival analysis . They used the SEER(Surveillance Epidemiology and End Results) data's twenty variables in the prediction models. The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy. C4.5 is a well known decision tree induction learning technique which has been used by Abdelghani

Bellaachia et al.[16] along with two other techniques i.e. Naïve Bayes and Back-Propagated Neural Network. They also presented an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and used the new version of the SEER Breast Cancer Data. The pre-processed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They have adopted a different approach in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and used the Weka toolkit to experiment with these three data mining algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, they found

out that model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques. The results obtained in their work differed from the study of

Delen et al because they used a newer version of same dataset, a different pre-classification and different toolkit. Their experimental results showed that their approach outperformed the approach used by Delen et al. They also proposed that after including the missing data in EOD attribute of used dataset can also increase the performance more.

A. Soltani Sarvestani et al.[17] provided a comparison of the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which classify WBC and NHBCD data. RBF and PNN were proved as the best classifiers when used for training set. But PNN gave the best classification accuracy when the test dataset is considered.

A. Soltani Sarvestani et al.[18] provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. RBF and PNN were proved as the best classifiers in the training set. But the PNN gave the best classification accuracy when the test set is considered. This work showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well.

Dr. Medhat Mohamed Ahmed Abdelaal et al.[19] investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest.

K. Rajiv Gandhi et al.[20] constructed classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In this study to cope with heavy computational efforts, the problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce a smaller fuzzy rule bases system with higher accuracy. The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining the underlying attributes effectively.

Manaswini Pradhan et al.[21] suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients. The neural network, used in back propagation algorithm, is m-n-1 type network. The GA is used for optimally finding out the number of neurons in the single hidden layered model. For training and testing 10-fold cross validation method was adopted for Pima Indian Diabetes.

For Pima dataset the ANN gives the best accuracy with 5 neurons in the hidden layer. Best accuracy being 72% with average accuracy of 72.2%. The designed model was compared with the Functional Link ANN (FLANN) and several classification systems like NN (nearest neighbour), kNN(k-nearest neighbour), BSS(nearest neighbour with backward sequential selection of feature, MFS1(multiple feature subset) , MFS2(multiple feature subset) for Dataclassification accuracies. It was revealed from the simulation that their suggested model performed better than compared to all of the participating techniques for comparison.

J. Padmavati[22] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011 158 logistic regression. When comparing RBF and MLP neural network models , it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP.
Humar Kahramanli et al. [23] presented an algorithm for extracting comprehensible classification rules for diagnosis of liver disorders. This algorithm considered all input attributes and extracts rules from the trained neural network with adaptive activation function efficiently. They observed that the neural network trained with adaptive activation function achieved high classification accuracy. Therefore, Neural network was trained by adaptive activation function in hidden layer and fixed sigmoid activation function in output layer. OptaiNET that is an Artificial Immune Algorithm (AIS) used in extracting rules from the trained neural networks. This approach was applied to BUPA Liver Disorders classification problems. The results of comparison experiments showed that the developed approach generated more accurate rules.

Conclusion from literature Review

The application of more regression techniques, new approaches and different tools over different datasets can improve the decision making. This can help researchers from healthcare to do better decision making and value formulation. From above discussion it is clear that more efficient work can be done over the healthcare problems by using new approaches and algorithms in data mining.

IV. METHODOLOGY

KDD structure constructed a framework of data mining, which has five steps: problem definition, data set selection, cleaning and pre-processing, data analysis, and evaluation and application, shown as fig 2.
The iterative process consists of the following steps which are described as follows:

1. Problem Definition

The diagnosis and treatment of breast cancer has become a serious and important health issue among women in the whole world. This disease is most common cause of death from cancer in women above the age of 45.

2. Dataset Selection

The dataset is taken from the UCI machine learning repository [23].

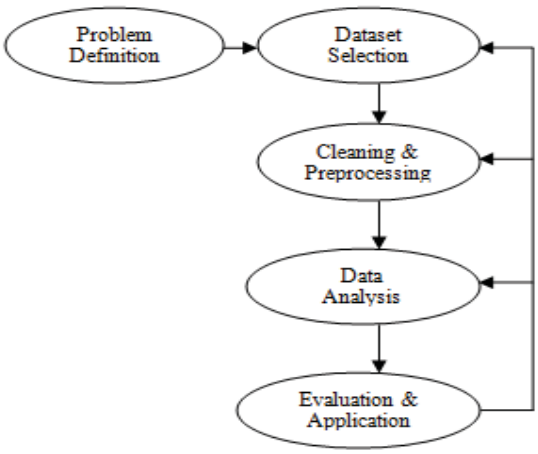

Fig. 2 Research Structure

3. Cleaning and Pre-processing

In this step, the noisy data is removed from complete dataset and final correct target data set is prepared that can be used for further analysis and decision making.

4. Data Analysis

In this step, different data mining regression algorithms are used to investigate the purposeful rules from the data of breast cancer.

5. Evaluation and Application

In this step after the application of the classification techniques on the dataset of breast cancer, some set of rules are extracted which are important for the diagnosis and treatment of breast cancer.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results and analysis done for this study. The research methodology for experiments has been explained in section 4. For the experiments, various different regression techniques have been applied on the Breast Cancer dataset taken from UCI repository.

Linear Regression Model

Diagnosis = -0.0709 * Radius -0.0047 * Perimeter + 0.0008 * Area + 3.8735 * Compactness -1.1601 * Concavity -1.9936 * ConcavePoints -0.969 * RadiusSE + 0.0363 * PerimeterSE + 0.0029 * AreaSE -19.4375 * SmoothnessSE +3.0168 * ConcavitySE -0.0105 * WorstTexture -0.0058 * WorstPerimeter + 0.0002 * WorstArea -0.3631 * WorstConcavity -1.8387 * WorstConcavePoints -0.811 * WorstSymmetry -3.909 * WorstFractalDimension + 3.133

Pace Regression Model

Diagnosis = 3.2045 + 3.6423 * Compactness -4.627 * ConcavePoints -0.2914 * RadiusSE -19.9938 * SmoothnessSE + 3.3815 * ConcavitySE -6.8627 * ConcavePointsSE -0.1377 * WorstRadius -0.0107 * WorstTexture + 0.0009 * WorstArea -0.7108 * WorstConcavity -0.607 * WorstSymmetry -4.7815 * WorstFractalDimension

Simple Linear regression on WorstConcavePoints

Diagnosis =-5.84 * WorstConcavePoints + 1.3

Regression by Discretization

```
WorstArea <= 880.8
|  WorstConcavePoints <= 0.1357
|  |  AreaSE <= 36.46: '(0.9-inf)' (319.0/3.0)
|  |  AreaSE > 36.46
|  |  |  Radius <= 14.97
|  |  |  |  TextureSE <= 1.978: '(0.9-inf)' (11.0)
|  |  |  |  TextureSE > 1.978
|  |  |  |  |  TextureSE <= 2.239: '(-inf-0.1]' (2.0)
|  |  |  |  |  TextureSE > 2.239: '(0.9-inf)' (3.0)
|  |  |  Radius > 14.97: '(-inf-0.1]' (2.0)
|  WorstConcavePoints > 0.1357
|  |  WorstTexture <= 27.37
|  |  |  WorstConcavePoints <= 0.1789
|  |  |  |  AreaSE <= 21.91: '(0.9-inf)' (12.0)
|  |  |  |  AreaSE > 21.91
|  |  |  |  |  PerimeterSE <= 2.615: '(-inf-0.1]' (6.0/1.0)
|  |  |  |  |  PerimeterSE > 2.615: '(0.9-inf)' (6.0)
|  |  |  WorstConcavePoints > 0.1789: '(-inf-0.1]' (4.0)
|  |  WorstTexture > 27.37: '(-inf-0.1]' (21.0)
WorstArea > 880.8
|  Concavity <= 0.0716
|  |  Texture <= 19.54: '(0.9-inf)' (9.0/1.0)
|  |  Texture > 19.54: '(-inf-0.1]' (10.0)
|  Concavity > 0.0716: '(-inf-0.1]' (164.0)

Number of Leaves :  13
```

Isotonic regression

Based on attribute: WorstPerimeter

prediction:	1	cut point:	85.1
prediction:	0.97	cut point:	91.7
prediction:	0.93	cut point:	101.65

prediction:	0.75	cut point:	102.05
prediction:	0.68	cut point:	105.15
prediction:	0.67	cut point:	105.95
prediction:	0.5	cut point:	114.45
prediction:	0.33	cut point:	117.45
prediction:	0.1	cut point:	120.35
prediction:	0.05	cut point:	127.2
prediction:	0		

Logistic Regression

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.939	0.042	0.93	0.939	0.934	0.978	Malignant
	0.958	0.061	0.963	0.958	0.961	0.977	Benign
Weighted Avg.	0.951	0.054	0.951	0.951	0.951	0.977	

Simple Logistic Regression

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.948	0.008	0.985	0.948	0.966	0.995	Malignant
	0.992	0.052	0.97	0.992	0.981	0.9995	Benign
Weighted Avg.	0.975	0.036	0.976	0.975	0.975	0.995	

Performance Comparison

	Relative absolute error	Root relative squared error
Linear Regression	41.07	50.49
Pace Regression	40.97	50.39
Simple Linear Regression	50.33	60.72
Regression By Discretization	15.53	52.44
Isotonic Regression	23.99	50.65
Logistic Regression	10.59	45.84
Simple Logistic regression	9.27	29.67

VI. CONCLUSION

This paper explores that Logistic Regression gives best results to diagnose Breast Cancer with given attributes among seven regression techniques with least Relative Absolute Error (10.59%) and Least Root Relative Squared Error. (45.84%) using Weka.

By knowing the best regression technique over a dataset a set of rules can be generated for that particular dataset and these rules will complement the healthcare researchers' study for intelligent value formulation. At last for future work it is suggested that more experiments can also be done on datasets from different domain using different parameters and techniques.

REFERENCES

[1] Oracle® Data Mining Concepts 11g Part No B28129. Available: http://www.comp.dit.ie/btierney/Oracle11gDoc/datamine.111/b28129/regress.htm.

[2] Avinash R. Pinglae, Aparna A.Junnarkar "RaajHans: A Data Mining Tool using Soft Computing Techniques" 2014

[3] Calle J. Breast cancer facts and figures 2003-2004.American Cancer Society 2004, pp1-27 2004.

[4] Breast cancer Q&A/facts and statistics [Online]. Advailable: http://www.komen.org/bci/bhealth/QA/q_and_a.asp.

[5] Delen Dursun,Walker Glenn,Kadam Amit, "Predicting breast cancer survivability: a comparison of three data mining methods" *Artif intell Med.* 2004.

[6] Jerez-Aragones JM, Gomez-ruiz JA,Ramos-Jimenez G,Munoz-perez J,Alba-conejo E. " A combinational neural network and decision tree model for prognosis of breast cancer relapse," in Proc Artif intell Med. 2003.pp. 45-63.

[7] Xiong Xiunngch , Kim Yangon, Baek Yuncheol, Rhee Wong Dae , Kim Hong Soo, "Statistical Analysis of BREAST Cancer Using Data Mining & Techniques", in Proc. of The Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks, 2005.

[8] Ritter M. Gene tied to manic-depression Newspaper article in Tulsa world June 16, 2003.

[9] Nada Lavrac, "Selected techniques for data mining in medicine," in *Proc*. Artificial Intelligence in Medicine*,* 1999, pp.3-23.

[10] G Richard, VJ Smith Rayward,PH Sonksen,S Carey, C. Weng, "Data mining for indicators of early mortality in a database of clinical records", in *Proc. Artif intell Med.* 2001,pp. 215-31.

[11] Shelly Gupta, Dharminder Kumar, Anand Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis" 2011

[12] Jeffrey W. Seifert "Data Mining and Homeland Security : An Overview" 2008

[13] Mark Tranmer, Mark Elliot," Multiple Linear Regression"

[14] Salford Systems, "Nonlinear Regression: Moderm Approaches and Applications"http://1.salford-systems.com/free-white-paper-non-regression

[15] Delen Dursun, Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-27 , June 2005.

[16] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[17] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang, "Rule discovery using hierarchical classification structure with rough sets," IFSA World Congress and 20th NAFIPS International Conference, 2001, vol.1 , pp. 447-452.

[18] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques," Software Technology and Engineering (ICSTE), 2nd International Conference, 2010, vol.2, pp.227-231.

[19] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cancer," in Proc. International multi conference on computer science and information Technology, 2010, pp. 11-17.

[20] Gandhi Rajiv K., Karnan Marcus and Kannan S., "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets," Signal Acquisition and Processing ICSAP, International Conference, 2010, pp. 233 – 237.

[21] Manaswini Pradhan and Dr. Ranjit Kumar Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004), pp.303 -311, vol. 2, iss. 2, 2011.

[22] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.

[23] Kahramanli Humar and Allahverdi Novruz, "Mining Classification Rules for Liver Disorders", International Journal of Mathematics and Computers in Simulation, vol. 3, 2009.

[24] Hornik k.,Stinchcombe M,white H., "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network," Neural networks pp.359-66,1990.

[25] Hastie T,Tibshirani R,Friedman J. The elements of statistical learning. New York, NY: Springer-verlag, 2001.

[26] Wu Hui-Chun,Fang Kwoting, Chen Cheng Ta, "Applying data mining for prostate cancer,"in Proc. International Conference on New Trends in Information and Service Science, pp. 1063-1065.

[27] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001.

[28] Ribeiro X. Marcela, M. J. Agma, jr.Caetano Traina, and Azevedo- Marques M. Paulo, "An Association Rule-Based Method to Support Medical Image Diagnosis with Efficiency ",in Proc. of IEEE Transactions on Multimedia, Vol. 10, No. 2, pp.277-285, 2008.

[29] Hassanien Ella Aboul and Ali H.M. Jafar, "Rough set approach for generation of classification rules of Breast cancer data," Journal Informatica, 2004, vol. 15, pp. 23–38.

[30] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose, "A Data Mining approach for detection of high-risk Breast Cancer groups," Advances in Soft Computing, vol. 74, pp. 43-51, 2010.

[31] Suneetha N., Hari V. M. K. and Kumar V.S., "Modified Gini Index Classification: A Case Study of Heart Disease Dataset", International Journal on Computer Science and Engineering, issue 6, vol. 2, pp. 1959-1965, 2010.

[32] Elsayad A. M., "Diagnosis of Erythemato-Squamous Diseases using Ensemble of Data Mining Methods", ICGST-BIME Journal, issue 1, vol. 10, 2010.

[33] Ubeyli E., "Multiclass support vector machines for diagnosis of erythemato-squamous diseases". Expert Systems with Applications, 35(4):1733–1740, 2008.

[34] http.//archive.ics.uci.edu/ml/breast+cancer+wiscosin+ (diagnosis)

[35] Thomas Grubinger, A. Zeileis, Karl-Peter Pfeiffer, evtree: Evolutionary learning of globally optimal and classification and regression trees in R 2011-20. EEECON, University of Innsbruck [Online]. Available: http://eeecon.uibk.ac.at/

[36] Hothorn T, Zeileis A (2011). "partykit: A Toolkit for Recursive Partytioning." R package version 0.1-1, URL http://CRAN.R-project.org/package=partykit

[37] William B.King, "R tutorials" http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html