International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 3, Issue 3 (May-June 2015), PP. 130-133

PERFORMANCE EVALUATION OF DIFFERENT CLASSIFIER ON BREAST CANCER

Rashmi Nagpal[1], Rashmi Shrivas[2], Mridu Sahu[3], Saransh Shirke[4]
[1]Department of Computer Science and Engineering, MATS, University Raipur.
[2]Departmet of Electronics and Telecommunication, NIT Raipur, C.G., India.
[3]Departmet of Electrical Engineering, NIT Raipur, C.G., India
rashmi.nagpal006@gmail.com, mrisahu.it@nitrr.ac.in

Abstract. In this work article founds the best classifier for breast cancer diseases classification, the survey shows it is one of the major diseases in India. For classification of Breast cancer is a multivariate data analysis it is big challenge among researcher to classify multivariate data because more than one decision variables present .The proposed work founds which classifier is best among 48 different categories of classifier and it founds logistics function is best among all the classifiers, it shows 80.9524 accuracy. Logistics function is also called logistic regression; it is a statistical method to analysis breast cancer data set.

Keywords: Breast Cancer, Classification, Multivariate Data, Logistic Function.

I. INTRODUCTION

Breast Cancer is developed from breast tissues [1]. There are several types of breast cancer. There are three ways by which cancer can spreads in body firstly by Tissues, by Lymph System or by Blood[2].There are many stages of breast cancer it is shown in literature[3][4][5].Stages description is shown in Table[1].The proposed article is finding best classifier for different types of breast cancer ,the data set description shown in table [2].The first type is Carcinoma is a type of cancer that develop from epithelial cells[6],second type is Fibro-adenoma of non cancerous tumors[7],third is Mastopathy is cover all types of breast changes[8].Forth type is Glandular epithelial cells are specialized epithelial[9] .Fifth type is Connective tissues that supports connects or spreads different types of tissues and organ in the body[10].Sixth type is Adipose tissues or fat its role is to store energy in the form of fat[11] .The proposed article finds dataset in to six different classes it processed on 48 different classifiers.

The 48 classifier comes from different categories like tree based, naïve based, function based, meta (Combination of more than two classifier) etc, the function based classifier performing well for present breast cancer dataset. In recent survey by breast cancer.org, shown in figure [1], says that in India many persons having this problem. Maximum patient found in stage I and stage IIA.

Table1: Stages of Breast Cancer

Stages	Definition
Stage 0	It is a condition in which abnormal cells are found in breast.
Stage I	In this stage tumor is 2 centimeters or smaller. In this stage the smaller clusters of breast cancer cells found in lymph nodes
Stage II	It is also divided into two parts 2A and 2B, 2A describes invasive breast cancer, 2B the tumor is larger than 2 centimeter but no larger than 5centimeter.
Stage III	It is also divided into three parts 3A, 3B and 3C.In 3A tumor is larger than 5 centimeter, 3B tumor may be any size and 3C there may be no sign of cancer in the breast.
Stage IV	In this cancer has spread to other organs of the body, most often the bones, lungs, liver, or brain.

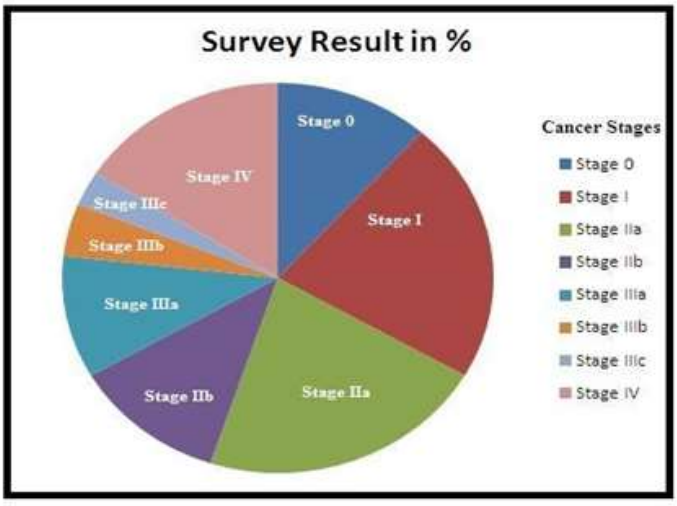


Figure 1: Survey Report For Breast Cancer

II. CORPUS (BREAST CANCER DATASET)

In this paper microarray dataset used it is gene expression dataset. The gene expression data set contains numeric values of genes, these expressions are either positive or negative. The proposed article uses UCI dataset and it is a one of the machine learning repository .The dataset having nine features and one class attribute and the description of data set is given below. The dataset is multivariate and attribute characteristics is integer and number of instances are 106[12] [13].

	Instances	Count
Car	Carcinoma	21
Fad	Fibro-adenoma	15
Mas	Mastopathy	18
Gla	Glandular	16
Con	Connective	14
Adi	Adipose	22
	Total	**106**

Table 2: Class Labels of Breast Cancer Dataset

Feature	Description
I0	Impedivity (ohm) at zero frequency
PA500	Phase angle at 500 KHz
HFS	High-frequency slope of phase angle
DA	Mpedance distance between spectral ends
AREA	Area under spectrum
A/DA	Area normalized by DA
MAX IP	Maximum of the spectrum
DR	Distance between I0 and real part of the Maximum frequency point
P	Length of the spectral curve

Table 3: Attributes Of Breast Cancer Dataset

III. EXTREME VALUE REMOVAL

The extreme value removal is a part of data cleaning step for data mining. The procedure for applying the extreme value theorem is to first establish that the function is continuous on the closed interval [14]. The next step is to determine the critical points in the given interval and evaluate the function at these critical points and at the end points of the interval. If the function f(x) is continuous on closed interval [a, b] then f(x) has both a maximum and a minimum on [a, b] [15]. In proposed method inter-quartile range [IQR] is used for extreme value calculations. IQR is major of variability based on dividing the dataset into quartiles [16].

Proposed article found two instances after removal of this breast cancer data set contain 104 instances for classification task.

IV. CLASSIFICATION

Classification is task performed with data-mining tools; it is process of creating groups for similar data. Instances are present in similar groups are highly correlated with each other. Machine learning algorithms like tree based,naïve based ,function based etc performing groping of instances that are similar to each other placed in same class and instances that are dissimilar with each other placed in another class. Inter Class dependency is low and intra class dependency is high the instances placed using classification algorithms [17].Proposed article found, function based classifier like logistic regression is a good classifier the performance of this is better than other Categories of classifier.

V. FUNCTION BASED LOGISTIC (FBS)CLASSIFICATION

FBS is using a multinomial logistic regression model with a ridge estimator.The Ridge value in the log-likelihood.If there are k classes for n instances with m attributes, the parameter matrix B to be calculated will be an m*(k-1) matrix[18].

The probability for class j with the exception of the last class is:

$$Pj(Xi) = \exp(XiBj)/((\text{sum}[j=1..(k-1)]\exp(Xi*Bj))+1) \quad \text{Eq(1)}$$

The last class has probability

$$1-(\text{sum}[j=1..(k-1)]Pj(Xi)) = 1/((\text{sum}[j=1..(k-1)]\exp(Xi*Bj))+1 \quad \text{Eq(2)}$$

The (negative) multinomial log-likelihood is thus

$$L = -\text{sum}[i=1..n]\{ \text{sum}[j=1..(k-1)](Yij * \ln(Pj(Xi))) +(1 - (\text{sum}[j=1..(k-1)]Yij)) * \ln(1 - \text{sum}[j=1..(k-1)]Pj(Xi)) \} + \text{ridge} * (B^2)$$

In order to find the matrix B for which L is minimized, a Quasi-Newton Method is used to search for the optimized values of the m*(k-1) variables. Note that before we use the optimization procedure, we 'squeeze' the matrix B into a m*(k-1) vector.

Although original Logistic Regression does not deal with instance weights, we modify the algorithm a little bit to handle the instance weights [19].

VI. PROPOSED METHOD

Proposed methodology is shown in figure [2].For finding the best classifier for breast cancer dataset is a challenge because the nature of data is multivariate .Missing value handling and extreme value removal from multivariate is a big task it is shown by literatures. Step by Step methodology is described below:

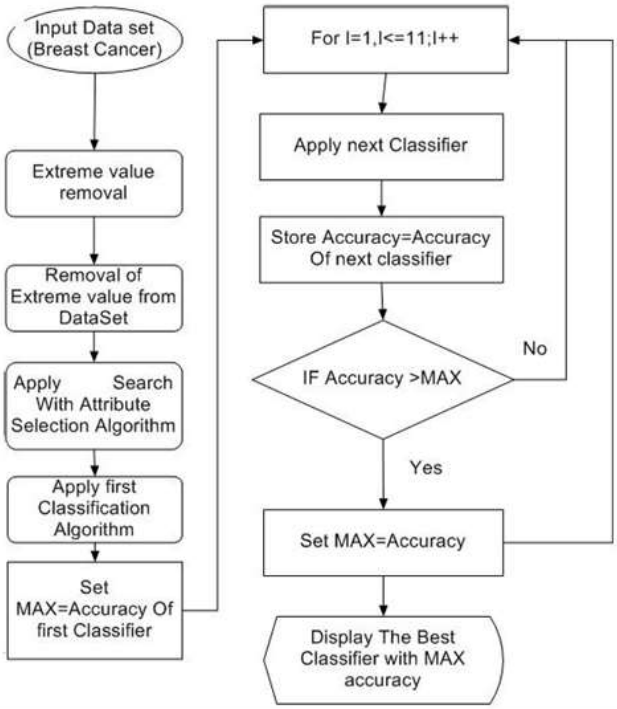


Figure [2]: Flowchart of Proposed article

Step1. Take input as Breast Cancer Data sets it is having 10 attributes. 9 features+1class attribute 106 instances.

Step2. Create training, testing and validation set.

Step3. Apply Attribute searches

Step4. Set Max value equal to first classifier accuracy.

Step5. Then apply for loop (for i=1,i<=48, i++).This loop is test one by one classification accuracy.

Step6. Create Performance evaluation matrix

Step7. After testing all classifier which classifier classification accuracy is high this classifier is best for another classifier.

Step8. The result shows the best classifier with the highest accuracy in breast cancer classification.

VII. RESULT ANALYSIS

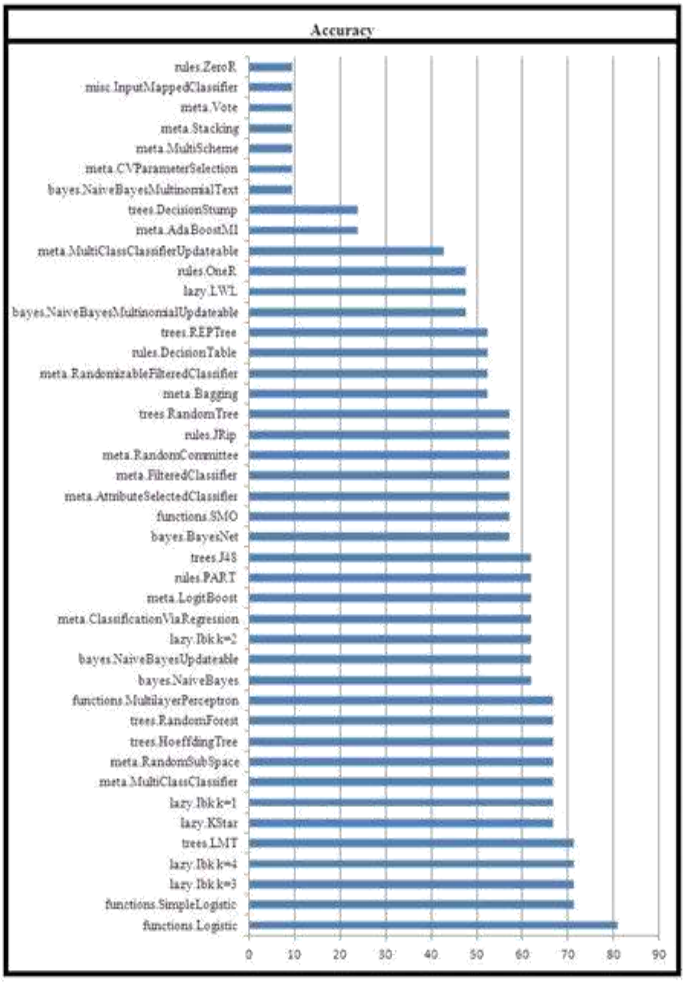
Experiment is performed with Java and Machine learning tool (Weka), for multivariate data set, 48 classifier accuracy result is shown in figure[3],for measuring the performance of classifier various matrices are present and these are mapped in table[4].In this only top 10 classifier performance is mapped and it is showing Function based logistic gives better performance as compare to others. Statistical Measures like TP(True Positive), FP(False Positive), Recall etc also mapped in table[4].

VIII. CONCLUSION

The proposed article gives how to evaluate classifier performance for multivariate cancer dataset. It is implemented with java and weka tools for execution, the classifier performance is depends upon the correlation found between two instances, based on this article found logistic function is good choice for multiple decision variables for class labels. It is showing 80.9524.

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Accuracy	Error
functions.Logistic	**0.81**	**0.033**	**0.857**	**0.81**	**0.811**	**0.784**	**0.899**	**0.746**	**80.9524**	**19.0476**
functions.SimpleLogistic	0.714	0.039	0.595	0.714	0.635	0.627	0.92	0.762	71.4286	28.5714
lazy.Ibk k=3	0.714	0.03	0.871	0.714	0.722	0.725	0.885	0.731	71.4286	28.5714
lazy.Ibk k=4	0.714	0.033	0.81	0.714	0.717	0.702	0.866	0.679	71.4286	28.5714
trees.LMT	0.714	0.039	0.595	0.714	0.635	0.627	0.92	0.762	71.4286	28.5714
lazy.KStar	0.667	0.058	0.782	0.667	0.671	0.637	0.903	0.753	66.6667	33.3333
lazy.Ibk k=1	0.667	0.075	0.746	0.667	0.669	0.617	0.796	0.591	66.6667	33.3333
meta.MultiClassClassifier	0.667	0.061	0.679	0.667	0.656	0.597	0.835	0.66	66.6667	33.3333
meta.RandomSubSpace	0.667	0.072	0.658	0.667	0.644	0.584	0.914	0.73	66.6667	33.3333
trees.HoeffdingTree	0.667	0.058	0.659	0.667	0.647	0.593	0.908	0.699	66.6667	33.3333

Table 4: Top 10 Accuracy Classifier Result



REFERENCES

[1] Van't Veer, Laura J., Hongyue Dai, Marc J. Van De Vijver, Yudong D. He, Augustinus AM Hart, Mao Mao, Hans L. Peterse et al. "Gene expression profiling predicts clinical outcome of breast cancer." nature 415, no. 6871 (2002): 530-536.

[2] Wooster, Richard, Graham Bignell, Jonathan Lancaster, Sally Swift, Sheila Seal, Jonathan Mangion, Nadine Collins et al. "Identification of the breast cancer susceptibility gene BRCA2." Nature 378, no. 6559 (1995): 789-792.

[3] Sotiriou, Christos, Soek-Ying Neo, Lisa M. McShane, Edward L. Korn, Philip M. Long, Amir Jazaeri, Philippe Martiat, Steve B. Fox, Adrian L. Harris, and Edison T. Liu. "Breast cancer classification and prognosis based on gene expression profiles from a population-based study." Proceedings of the National Academy of Sciences 100, no. 18 (2003): 10393-10398.

[4] Mangasarian, Olvi L., W. Nick Street, and William H. Wolberg. "Breast cancer diagnosis and prognosis via linear programming." Operations Research 43, no. 4 (1995): 570-577.

[5] Van Dongen, J. A., H. Bartelink, I. S. Fentiman, T. Lerut, F. Mignolet, G. Olthuis, E. Van der Schueren, R. Sylvester, J. Winter, and K. Van Zijl. "Randomized clinical trial to assess the value of breast-conserving therapy in stage I and II breast cancer, EORTC 10801 trial." Journal of the National Cancer Institute. Monographs 11 (1991): 15-18.

[6] Bokhman, Jan V. "Two pathogenetic types of endometrial carcinoma." Gynecologic oncology 15, no. 1 (1983): 10-17.

[7] Dupont, William D., David L. Page, Fritz F. Parl, Cindy L. Vnencak-Jones, Walton D. Plummer Jr, Margaret S. Rados, and Peggy A. Schuyler. "Long-term risk of breast cancer in women with fibroadenoma." New England Journal of Medicine 331, no. 1 (1994): 10-15.

[8] London, Stephanie J., James L. Connolly, Stuart J. Schnitt, and Graham A. Colditz. "A prospective study of benign breast disease and the risk of breast cancer." Jama 267, no. 7 (1992): 941-944.

[9] Kämäräinen, M., M. Seppälä, I. Virtanen, and L. C. Andersson. "Expression of glycodelin in MCF-7 breast cancer cells induces differentiation into organized acinar epithelium." Laboratory investigation; a journal of technical methods and pathology 77, no. 6 (1997): 565-573.

[10] Hishikawa, Keiichi, Barry S. Oemar, Felix C. Tanner, Toshio Nakaki, Thomas F. Lüscher, and Tomoko Fujii. "Connective tissue growth factor induces apoptosis in human breast cancer cell line MCF-7." Journal of Biological Chemistry 274, no. 52 (1999): 37461-37466.

[11] Boyd, N. F., J. W. Byng, R. A. Jong, E. K. Fishell, L. E. Little, A. B. Miller, G. A. Lockwood, D. L. Tritchler, and Martin J. Yaffe. "Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study." Journal of the National Cancer Institute 87, no. 9 (1995): 670-675.

[12] http://archive.ics.uci.edu/ml/

[13] Asuncion, Arthur, and David Newman. "UCI machine learning repository." (2007).

[14] Castillo, Enrique. Extreme value theory in engineering. Elsevier, 2012.

[15] Gumbel, Emil Julius. Statistics of extremes. Courier Corporation, 2012.

[16] Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. "Advances in knowledge discovery and data mining." (1996).

[17] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[18] Witten, Ian H., Eibe Frank, Leonard E. Trigg, Mark A. Hall, Geoffrey Holmes, and Sally Jo Cunningham. "Weka: Practical machine learning tools and techniques with Java implementations." (1999).

[19] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.