International Journal of
Analytical, Pharmaceutical and Biomedical Sciences

Volume: 4: Issue-2: February-2015 *ISSN:2278-0246*

www.ijapbs.com

APPLICATION AND COMPARISON OF DIFFERENT LINEAR CLASSIFICATION METHODS FOR BREAST CANCER DIAGNOSIS

Shaho Zarei[1], Mina Aminghafari[2], Hakimeh Zali*[3]

[1, 2]Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
[3]Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
*Corresponding author: Email: h.zali@sbmu.ac.ir.

ABSTRACT: Aim: Breast cancer is one of the most common cancers among women compared to all other ones. Proper and early detection of tumor type in a patient is a critical subject that can increase the survival chance. Statistical methods, such as classification can be useful instrument for physicians as an auxiliary tool to detect the type of tumor, increase the speed, reduce the cost and time of diagnosis as well as reduce the error of physician in the diagnosis of the tumor. Methods: In this paper, we use multivariate linear regression, logistic regression, the K-nearest neighbor (KNN) method and discriminant analysis to determine tumor type in a patient using Wisconsin Breast Cancer (WBC) database. In addition to linear methods, quadratic discriminant analysis will be used. Stepwise method for variable selection is used in regression method. Results: The accurate percentage of classification on testing data, depends on selected method and the number of independent variables and at least equal to 90.8% for logistic regression and a maximum 99.6% for KKN with K=9. Our results show about 3.1% of doctors' diagnoses in the breast cancer may be incorrect. Conclusion: In General, if the assumptions used in each method is established, it can be said that the difference is not significant in choosing model type, and all of these methods are good tool for increasing precision of breast diagnosis.
Keywords: Linear Classification, Breast Cancer Diagnosis, Regression, Discriminant Analysis, KNN.

INTRODUCTION

Although breast cancer is the most common type of cancer among women but men also are at risk. Breast tumors can be divided into malignant and benign. Todays by using screening, physicians assess detection of early breast cancer. Detection and treatment of breast cancer in the early stages before metastatic stages can increase patient's treatment chance. However, diagnosis the type of tumor is a difficult task and depended on the physician experience. So inaccurate predictions may be occurred since the experiments are prone to human and visual error and may be affected by blurred mammogram visuals, Bagui et al. [1].

Advanced statistical methods and Data Mining, as auxiliary tool, are powerful instruments for solving this problem. Statistical methods can increase detection speed, reduce the cost and time of diagnosis as well as reducing the error of physician in the diagnosis of tumor type.

The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors, Joshi et al. [2]. There are many techniques and methods to predict and classification breast cancer pattern and large number of articles about application of data mining techniques to diagnosis breast cancer.

You and Rumbe [3], provided comparative analysis of Support Vector Machine, Bayesian classifier and other artificial neural network classifiers. Gouda et al. [4], perform a comparison among the different classifiers: decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-nearest neighbor (KNN) on three different databases of breast cancer:

Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) using classification accuracy and confusion matrix based on 10-fold cross validation method. Also, Bagui et al. [1], compared the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART, to find the best classifier in breast cancer datasets (WBC and Breast tissue). Their experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in WBC and 99.00% in Breast tissue. Chaurasia and Saurabh [5], compared three classification techniques in Weka software and their comparison results show that Sequential Minimal Optimization (SMO) has higher prediction accuracy i.e. 96.2% than IBK and Bloom filter trees (BF-Trees).

Statistical classification is one of the most important statistical methods in multivariate techniques and has widely used in practice. In next section we give a brief overview on this method and especially on linear classification method.

A BRIEF OVERVIEW ON CLASSIFICATION

Suppose we have known the class label, the aim of classification technique is to establish a rule whereby we can classify a new observation into one of the existing classes. The main idea of this method is finding a linear combination of the variables x_1,…,x_k, such that this combination separated two groups of observations as possible. These variables are called explanatory, predictor or independent variables or attributes. Several methods are suggested for classification. These method are based on tow general categories, one; calculating maximum posterior probability of designation or finding an appropriate experimental rule for classification.

If the boundaries of the decision or the equivalent predictive function is linear based on of predictors variables, classification method is called the linear classification. Due to the simplicity of the interpretation and appropriate accuracy, these methods are in a special category of classification. Multivariate linear regression of an indicator matrix, logistic regression, K-nearest neighbor (KNN) and linear discriminant analysis are popular linear classification methods. In all these methods, the response variable can has two or more classes or categories. In this research, response variable has two classes, and in addition to those linear methods, quadratic discriminant analysis will be used. Here, we give a brief introduce of these methods. For further study one can refer to Hastie et al. [6].

Linear Regression for Classification

In general, linear regression is an approach for modeling the relationship between a dependent variable y and one or more explanatory variables (or independent variable) denoted by x. In classification with using linear regression, the response categories are coded via an indicator variable and take values 0 or 1, Hastie et al. [6]. Given a data set {y_i,x_i1,…,x_ik }_(i=1)^n of n observations, the regression vector form is

$$y=X\beta+\varepsilon,$$

Here y is vector of 0 or 1 values, X is matrix of independent variables, β is regression coefficients vector and ε is error vector. From training data, we fit linear regression model and the fit is given by,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

Where β^=(β^_0,…,β^_k) and have $\hat{\beta} = (X'X)^{-1}X'y$. For new vector of independent variables, we calculate y^, if this value is equal or less than 0.5, this observation belong to first class and the second class otherwise.

The advantage of this method is that it doesn't have restrictive assumptions. Using the variable selection methods such as a stepwise procedure, one can determine effective variables, variables with significant relationship with response variable, and find easier model with lesser independent variables for doing classification. The serious problem with the regression approach occurs when the number of classes bigger than 2. Because of the rigid nature of the regression model, classes can be masked by others, Hastie et al. [6].

Logistic Regression

Whenever, in the regression approach, the response variable is binary, it is more appropriate to calculate the conditional probability values that the response variable belongs to class one with 0 value or class two with 1

value rather than predicted response variable directly. If π_1 is the probability of one new observation belonging to the second class, the logistic regression model is given by

$$\mathrm{Ln}\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

The coefficients β_0,…,β_k, usually by maximum likelihood method and numerical methods are estimated. If calculated probability for new independent variable, is equal or less than 0.5, this observation belong to first class elsewhere the second class. In this method, choosing the effective independent variables is possible, but results may not be the same with multivariate regression method.

K-Nearest Neighbor Method

K-Nearest Neighbor (KNN) classification Christobel and Sivaprakasam [7], classifies observation based on their similarity. In this method, to determine the class or category of an observation, we find K nearest neighbor, based on some criteria such as the Euclidean distance. After that, the average of the response variable for selected observations is calculated. This average is compared with a critical value, here 0.5, such as the previous two methods. The value of K is to determine and its value is usually a positive integer number between 1 to 15, Haowen and Rumbe [3]. A data-driven method for determining K, is cross validation technique.

Discriminant Analysis

In this method, independent variables are assumed to be normal distributed in each class. In comparing two class 0 and 1, it sufficient to look at the log–ratio of posterior probability, Hastie et al. [6]. Linear Discriminant Analysis (LDA) arises in the special case when we have a common covariance matrix in each class and otherwise discriminant analysis is Quadratic Discriminant Analysis (QDA). Discriminant function δ_k (x) is defined for LDA and QDA, respectively, as

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_x{}^T \Sigma^{-1} \mu_k + \ln(\pi_k)$$

And

$$\delta_k(x) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k{}^{-1}(x-\mu_k) + \ln(\pi_k)$$

Where, π_k is the prior probability of class k, μ_k is mean vector in each class, Σ is common covariance matrix and Σ_k is covariance matrix of class k for k=0 and 1. In practice the mean vectors and covariance matrixes are unknown and we estimated them by maximum likelihood method and the estimated values of the instance and the corresponding values to fall into place. In both cases, LDA and QDA, the maximum amount of the above discriminant functions for all values of k is determined and, if it has maximum at k = 0 then x belong to the class one and otherwise to the class two.

DATA DESCRIPTION

We use The Wisconsin Breast Cancer (WBC) datasets to distinguish malignant tumors from benign. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg , Frank and Asuncion [8]. The data set has 699 patients with a total of 10 independent variables and 2 classes (malignant or benign). These variables are used to fit models which can predict the diagnosis (malignant or benign) for a special patient. These variables have listed in Table 1. For simplicity, we denote these attribute with x_0,…,x_9 and y. It should be noted that the sample code number is nominal variable and we won't use it. Without loss of generality, since the class variable is a categorical response variable, we assume it 0 for benign and 1 for malignant.

Table1. Description of the variables Wisconsin Breast Cancer dataset.

	Attribute	Notation	Domain
1	Sample code number	x_0	Id number
2	Clump Thickness	x_1	1-10
3	Uniformity of Cell Size	x_2	1-10

4	Uniformity of Cell Shape	x_3	1-10
5	Marginal Adhesion	x_4	1-10
6	Single Epithelial Cell Size	x_5	1-10
7	Bare Nuclei	x_6	1-10
8	Bland Chromatin	x_7	1-10
9	Normal Nucleoli	x_8	1-10
10	Mitoses	x_9	1-10
11	Class	y	2 for benign, 4 for malignant

There are some obvious criteria to discriminant benign and cancer tumors. So that monolayer cells in clump thickness benign in contrast to multilayer cancerous cells or vary are seen in size and shape of cancer cells unlike normal cells. In addition marginal adhesion is preserved in normal cells while loss of adhesion is a sign of malignancy. For instance, single epithelial cell size is mentioned to the uniformity of normal cell, because significantly enlarged epithelial cells may be a malignant cell.

Other parameter is nucleus and its content. Nuclei that are not surrounded by cytoplasm related to the term Bare nuclei which typically seen in benign tumors. On the other hand, the nucleus of benign tumors has uniform texture of chromatin that named as the Bland Chromatin while in cancer cells are coarser. Furthermore, the nucleolus of normal cells unlike cancer cells is usually very small.

Mitoses are coupled to cell proliferation. The process in cell division by which, the nucleus and cytoplasm are divided. It has different patterns in neoplastic and non-neoplastic cells. However, with all these significant difference between normal and cancer cells that is important why these parameters are valuable in determining whether the cells are cancerous or not.

CASE STUDY AND EVALUOTION CRITERIA

First of all, we check the existence of the missing data and outliers in the WBC database. The results of this survey show that the Bare Nuclei has 16 missing data and there is no outlier in data. We use statistical methods (i.e. the local average imputation) to estimate these data and since this variable is integer, we round result to the nearest integer. It should be noted with respect to the large number of data one can neglect these missing data without loos any precision.

Using statistical classification methods for diagnosis breast cancer is the aim of this research. In fact, our goal is to firstly determine which variables have a significant relationship with tumor type, and secondly on the values of the selected variables diagnosis tumor type or calculate the probability of the individual person having certain tumors as uncertainty measure.

We divide data into two parts; training (with 450 instances) and testing data (with 249 instances). We choose these figures to avoid overfiting problem. From the training data we estimated the parameters of each model and perform classification, then with respect to obtained results, tumor type for each patient is determined in test data and compare our result with true class of tumor type. Also, the percentage of correct classification is accuracy measure. It should be noted, there is significance difference between corresponding means in levels of class variable.

EXPERIMENT RESULTS

In the regression method and in the completed model, we have 9 independent variables. The Coefficient of variation, Montgomery et al. [9], for this model is 0.867, so we can say that these variables express about 86.7% of variation in response variable and still there are variables that may be affected on tumor type that we should determine them. For reducing dimension of vector of independent variables, we use stepwise method as selection variables method. A corroding to this method, 6 predictive variables i.e., Clump Thickness, Uniformity of Cell Size Uniformity of Cell Shape, Bare Nuclei, Bland Chromatin and Normal Nucleoli, Marginal Adhesion, Single Epithelial Cell Size and Bland Chromatin are selected. The fitted model is given by

y=0.03 * ClumpThickness+0.021 * UniformityofCellSize+0.014*Uniformity of Cell Shape+0.057 * Bare Nuclei+0.013 * Bland Chromatin+0.023 * Normal Nucleoli-0.23 (1)

The Coefficient of variation for this model is 0.866 and accuracy of model is clear.

In logistic regression, three variables, i.e. Clump Thickness, Bare Nuclei and Normal Nucleoli are selected by

stepwise methods and fitted method is given by

$$\text{Ln}\left(\frac{\pi_1}{1-\pi_1}\right) = 0.722 * \text{ClumpThickness} + 0.727 * \text{Bare Nuclei} + 0.477 * \text{Normal Nucleoli} - 9.354,$$

Where π_1 is the probability which a patient has malignant tumor. With respect to table 1, we can write this model as below;

$$\pi_1 = exp\{-9.354 + 0.722x_1 + 0.727x_6 + 0.477x_8\}/(1 + exp\{-9.354 + 0.722x_1 + 0.727x_6 + 0.477x_8\}).$$

The Nagelkerke R Square, Nagelkerke [10] for this model is 0.952 as a result model has goodness of fit. Table 2 shows the classification results for this model on the train data.

Table 2: Classification table in logistic regression classifier

Observed		Predicted		
		response		Percentage Correct
		malignant	benign	
response	malignant	272	7	97.5
	benign	3	168	98.2
Overall Accuracy Percentage				97.8

Sometimes, this table is called confusion matrix, Han and Kamber [11]. From Table 2, For example π_1 for patient with amounts 5, 7 and 6 for Clump Thickness, Bare Nuclei and Normal Nucleoli, respectively, is 0.92 then this person with probability 92% has malignant tumor and this diagnosis has 97.8% of precision. In other hand, about 2.2% of doctors' diagnoses may be false. Whit similar calculation for WBC dataset, we can say about 3.1% of diagnosing tumor type by doctors may be incorrect.

We performed KNN classifier in three states with complete model based on all 9 explanatory variables and with 6 and 3 explanatory variables are obtained by Multi linear regression and logistic regression, respectively. Furthermore, we use the cross validation technique to determine the number of neighbors which is obtained K=9. Also, for each method, we calculate misclassification and denote them with Miss1, Miss2 and Miss3, corresponding. Table 3 shows our results. These results emphasize the importance of choosing the accurate K.

Table3: The number of misclassifications in KNN methods.

K	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Miss1	4	4	4	4	3	2	1	1	1	1	2	1	2	1	2
Miss2	4	4	5	4	3	3	2	2	2	2	2	2	3	2	2
Miss3	8	8	4	6	4	5	4	5	3	5	4	5	5	5	5

In discriminant analysis, similar to KNN, we use 9, 6 and 3 explanatory variables. For testing homogeneity of covariance matrix in multi normal distribution between two classes, we use Box's M test. P-value for this test is 0.000 so, homogeneity of covariance matrix is rejected at 0.05 level of significant, and thus we should use QDA. However, we do classification with LDA method as misspecification and we see reliable results as well. Table 4 shows our results.

Table 4: The number of misclassifications in LDA and QDA methods

The number of explanatory variables	9	6	3
LDA	4	3	9
QDA	6	4	8

In these methods the best result occurs in model with 6 explanatory variables.
Table 5 shows a comparison among classification accuracies for different proposed methods for WBC dataset.

Table 5: Caparison Accuracy of Classification methods.

Classification method	Number of Selected variables	Maximum Percentage of accuracy
Multivariate regression	6	92.00
Logistic regression	3	90.80
KNN1	9	99.60
KNN2	6	99.20
KNN3	3	98.79
LDA1	9	98.39
LDA2	6	98.79
LDA3	3	96.39
QDA1	9	97.59
QDA2	6	98.39
QDA3	3	96.78

A cording to our result, KNN classifier in complete model (KNN1) has higher prediction accuracy i.e. 99.6% than others.

CONCLUTION

With respect to coefficient of variation, although model 1(obtained from stepwise model), has three independent variables less than complete model, there is no significance difference between these two models. Despite this, the use of new model reduces the cost and time of diagnosis of tumor type. Generally, increasing the number of explanatory variables reduce misclassification numbers. KNN classifier has no restricted assumption and has a very good precision, so we suggest this method as auxiliary method for breast cancer diagnosis. The number of observation is important task and large number of observation cause to robustness in our result in using LDA or QDA classifiers. In practice, we suggest that one used at least two methods for diagnosing tumor type, if the results are be different, then we should do more experience. Also, before any classification we must test collinearity among independent variables

REFERENCES

[1] Bagui S and Pal, N. Breast cancer detection using rank nearest neighbor classifica-tion rules. Pattern recognition, 2003; pp.25-34.
[2] Joshi MJ, Doshi R and Patel J. Diagnosis and prognosis breast cancer using classifi-cation rules. International Journal of Engineering Research and General Science. 2014; 2(6).
[3] Haowen Y and Rumbe G. Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data. International Journal of Artificial Intelligence and Interactive Multimedia. 2010; 1(3).
[4] Salama GI, Abdelhalim MB and Zeid MAE. Breast cancer diagnosis on three different datasets using multi-classifiers. International Journal of Computer and In-formation Technology. 2012; 1(1).
[5] Chaurasia V and Saurabh P. A novel approach for breast cancer detection using data mining techniques. International Journal of Innovative Research in Computer and Communication Engineering. 2014; 2(1).
[6] Hastie T, Tibshirani R and Frideman J. The elements of statistical learning Data mining, inference, and prediction . 2001; Springer.
[7] Christobel A and Sivaprakasam Y. An empirical comparison of data mining classification methods. International Journal of Computer Information Systems. 2011; 3(2).
[8] Frank A and Asuncion A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of In-formation and Computer Science. 2010.
[9] Montgomery DC, Peck EA, Vinning GG. An introduction to linear regression anal-ysis. Fifth edition. Wiley.

[10] Nagelkerke, N.J.D. A note on a general definition of the coefficient of determina-tion.1991; Biometrika 78, 691-692.
[11] Han J and M Kamber. Data Mining Concepts and Techniques. 2000; Morgan Kauffman Publishers.