International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification

Mücahid Mustafa Saritas[1] , Ali Yasar[2,*]

Submitted: 12.12.2018 Accepted : 02/05/2019

Abstract: Classification is an important data mining technique with a wide range of applications to classify the various types of data existing in almost all areas of our lives. The purpose of this discovery study can be used to estimate the potential of having breast cancer by taking advantage of anthropometric data and collected routine blood analysis parameters. The study was performed using data from patients who were admitted to the clinic with the suspicion of breast cancer. The values of Age (years), BMI (kg/m2), Glucose (mg/dL), Insulin (µU/mL), HOMA, Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), MCP-1(pg/dL) were used. In our study, classification algorithms were applied to the data and they were asked to estimate the disease diagnosis. The classification performance of Artificial neural networks and Naïve Bayes classifiers which were applied to data with 9 inputs and one output were calculated and theperformance results were compared. This article sheds light on the performance evaluation based on correct and incorrect data classification examples using ANN and Naïve Bayes classification algorithm. When we look at the performances obtained, it is predicted that using the anthropometric data and the collected routine blood analysis parameters, the potential for diagnosing breast cancer is high using these data.

Keywords: *ANN, Breast Cancer, Classification, Artificial Neural Network, Machine Learning Database, Naïve Bayes*

1. Introduction

Data mining is widely used in a wide variety of applications, such as medical diagnosis, targeted marketing, estimating the shares of television audiences, financial forecasting product design, automated abstraction, analysis of organic compounds, credit card fraud detection [1]. Breast cancer is cancer that forms in the cells of the breasts[2]. Breast cancer is the most common cancer among women in the Western world[3]. In developed countries, breast cancer is the most common cancer among women and it is accepted that one out of every 9 women will catch this cancer. One out of every 4 women with cancer is suffering from breast cancer. Breast cancer screening is an important strategy that should be allowed to diagnose cancer.

It is of great importance to have a good result in early diagnosis and treatment. In order to make a good conclusion, routinely collecting blood data analysis and further screening tools will make a significant contribution. There is a comprehensive literature review of the relationship between body weight and risk of breast cancer, but the role of adipocytes still remains a matter of doubts about the role of insulin and glucose and the interference between these players, regardless of obesity [4]. Cancer risk increases with age[5]. In this study, the data that can be collected by using the blood analyses which are routinely collected can be firstly Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI). We used these data to predict the presence of breast cancer. There are 9 data, all of which are quantitative and binary dependent variables, which indicate the presence or absence of breast cancer and allow us to estimate breast cancer. Prediction is anthropometric data and parameters that can be collected in routine blood analysis. Predictive models based on these determinative parameters can potentially be used as biomarkers of breast cancer if considered correct. In our study, artificial intelligence and machine learning techniques have been applied to publicly available data in UCI Machine Learning Database. In this study, the classification was realized using the Breast Cancer Coimbra Data Set obtained from (http://archive.ics.uci.edu/ml/datasets.html).

2. Materials And Methods

We have included women who were diagnosed with breast cancer (BC) between 2009-2013 at the Department of Obstetrics and Gynecology of the University of Coimbra (CHEA)[6]. Using the Ages, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI) data from the UCI Machine Learning Repository (Center for Machine Learning and Intelligent Systems), the patients are classified with artificial neural network whether they have cancer or not. The diagnosis for each patient came from a positive mammography and was confirmed histologically. All samples were naïve, ie collected before surgery and treatment. All patients were excluded from treatment before consultation. Female healthy volunteers were selected and included as controls. All patients had no previous cancer treatment, and all participants did not have any infections or other acute diseases during registration [2].

In a study realized by Bebis at al. the quantitative characteristics of the patients and the healthy controls were defined in terms of the median and the interquartile intervals in Table 1. Interquartile intervals were given as median. The p-values in the table were obtained by Mann-Whitney U tests after evaluating normality assumptions with a Shapiro-Wilk test for each variable. BMI body mass index and MCP-1 monocyte chemoattractant protein-1 are the HOMA homeostasis model evaluation for insulin resistance [7]. Shu et al. confirmed the previously reported inverse

[1] *Department of Biomedical Eng., Selcuk University, Konya, Turkey, ORCID ID: 0000-0001-5451-9092*
[2] *Computer Programming, Guneysinir Vocational School of Higher Education Selcuk University Guneysinir, Konya, 42190,Turkey, ORCID ID: 0000-0001-9012-7950*
**Corresponding Author Email: aliyasar@selcuk.edu.tr*

association of genetically predicted BMI with breast cancer risk, and showed a positive association of genetically predicted fasting insulin and 2-h glucose and an inverse association of WHR adj BMI with breast cancer risk[8]. Age, sex, and family history, risk of developing breast cancer is largely linked[9]. Adiponectin, a peptide hormone secreted by the adipose tissue, has been inversely related to BC risk both in observational studies and in a phase II chemoprevention trial in premenopausal women at increased risk[10].

Table 1. Descriptive statistics of the clinical features

Inputs	Patients	Controls	p-value
Age (years)	53.0 (23.0)	65.0 (33.2)	0.479
BMI (kg/m2)	27.0 (4.6)	28.3 (5.4)	0.202
Glucose (mg/dL)	105.6 (26.6)	88.2 (10.2)	<0.001
Insulin (µU/mL)	12.5 (12.3)	6.9 (4.9)	0.027
HOMA	3.6 (4.6)	1.6 (1.2)	0.003
Leptin (ng/mL)	26.6 (19.2)	26.6 (19.3)	0.949
Adiponectin (µg/mL)	10.1 (6.2)	10.3 (7.6)	0.767
Resistin (ng/mL)	17.3 (12.6)	11.6 (11.4)	0.002
MCP-1(pg/dL)	563.0 (384.0)	499.7 (292.2)	0.504

In the study, the Neaural Network Toolbox and Naïve Bayes of the Matlab R2017b program was used.

2.1. Artificial Neural Network:

Artificial neural networks (ANN)[11] is a popular machine learning technique inspired by the biological neural network in the human brain[12]. Feed forward neural networks[7] are a common type of ANN which sends the weight values of each artificial neuron as output to the next layer after processing with inputs from neurons in the previous layer. An important class of feed forward neural network is Multilayer Perceptron (MLP) [13]. The back propagation algorithm is the most widely used MLP training technique. This changes the weights between neurons to minimize the error. This model is quite good in learning patterns. It can easily adapt to new values in the data, but the system can show a slow convergence and has the risk of a local optimum [14]. The determination of the number of layers and the number of neurons in the hidden layer and the connection between them is an important problem. These parameters and problems are of great importance in the performance of the artificial neural network. The results may vary greatly in any of these parameters. Different ANN architectures will give different results for different problems. However, it is important to come to an optimal ANN architecture by trial and error. The artificial neural network model forming our system is shown in Figure 1.

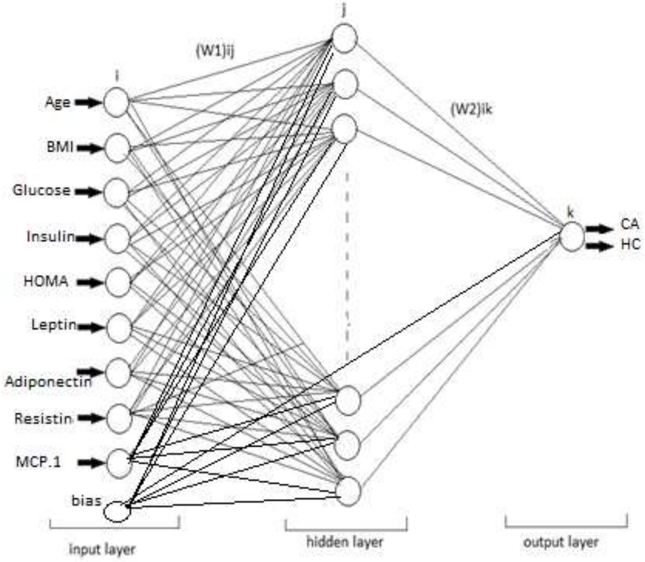

Figure 1. The structure of ANN

The training data set was used to determine ANN neuron and bias weight values. Training was repeated to obtain the lowest level of error by changing the number of neurons and the epoch number. Then, the trained algorithm was applied on the test data set.
As can be seen from Figure 1, our neural network consists of 10 inputs, a 10-layer hidden layer and 1 outputs.

2.2. Application of ANN for A Data Set

In the study, the following procedures were performed in order to determine breast cancer detection or health checks by using the data of descriptive statistics of 64 patients with breast cancer and 52 healthy controls (especially age, BMI and inflammatory and metabolic parameters).

1) Randomly, 29 (25%) data were selected as test data.
2) Randomly, 12 (10%) data were selected as validation data.
3) The remaining 75 (65%) were selected as training data.

The data were trained and classification procedures were provided in ANN. At the end of these procedures, the network structure that yielded the best classification is given in Table 2.

Table 2. The parameters and properties used in ANN

Parameters	Properties
Number of neurons in the input layer	10
Number of the hidden layers	1
Number of neurons in the hidden layer	10
Number of neurons in the output layer	1
Learning rate (α)	0,2
Coefficient of momentum (β)	0,3
Learning algorithm	Levenberg-Marquardt (trainlm)
Transfer function	Logarithmic sigmoid (logsig)

As a result of the training: Regression graph of the training set is given in Figure 2.

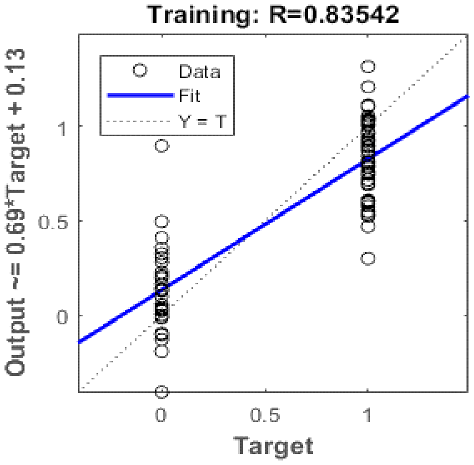

Figure 2. Regression of Training Data Set

The Regression Graph of the test set is shown in Figure 3.

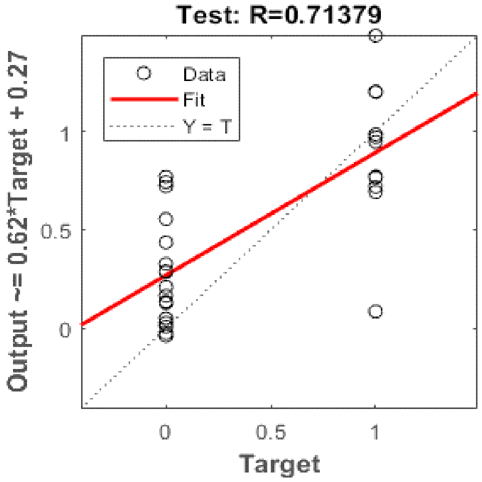

Figure 3. Regression of Test Data Set

The Regression Graph of the validation set is given in Figure 4.

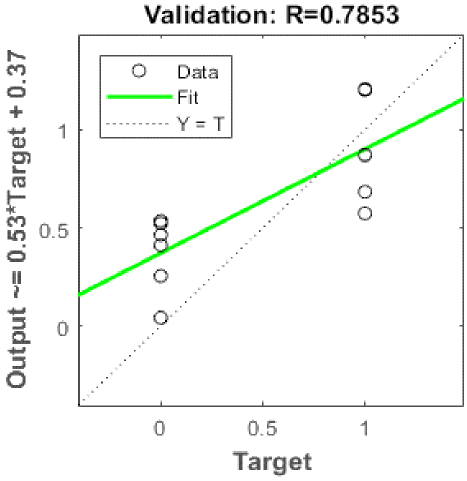

Figure 4. Regression of Validation Data Set

The general regression graph of the ANN is given in Figure 5.

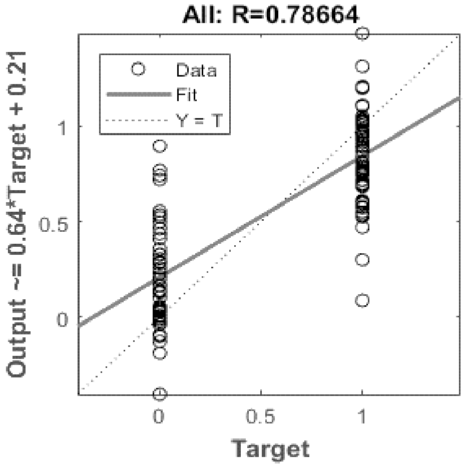

Figure 5. Regression of All Data Set

Figure 6 shows the histogram graph of the error of all of our data set.

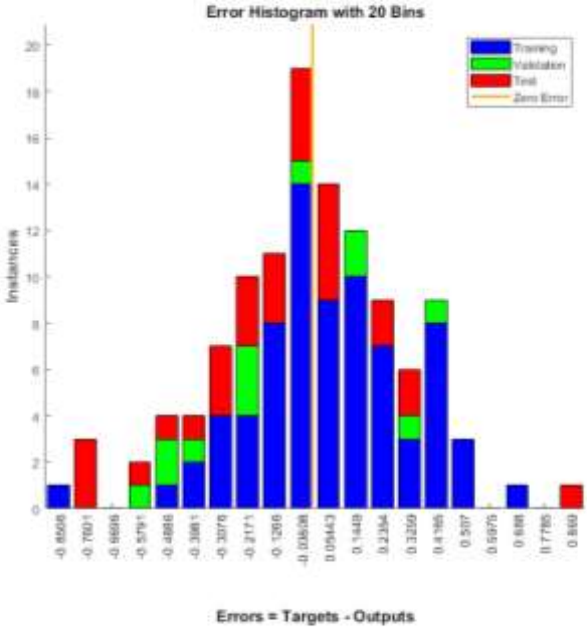

Figure 6. Error Histogram

2.3. Naive Bayes

The Naive Bayes algorithm is a simple probability classifier that calculates a set of probability by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes's theorem and assumes that all variables are independent considering the value of the class variable. This conditional independence assumption is rarely valid in real-world applications, so it is characterized as Naive, but the algorithm tends to learn quickly in a variety of controlled classification problems. [15].

Bayes' theorem is a mathematical formula used to determine conditional probability (Eq1), which is named after 18th century British mathematician Thomas Bayes.

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)} \tag{1}$$

Here;
P(A|B) is the probability of the occurrence of event A when event B occurs,
P(A) is the probability of the occurrence of A,
P(B|A) is the probability of the occurrence of event B when event A occurs,
P(B) is the probability of the occurrence of B.

Naive Bayes is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. Healthy controls or Patients.

Confusion Matrix:

```
 1    2 ← classified as
 9    1    | 1 = Healthy controls
 2   11    | 2 = Patients
```

For above confusion matrix, true positives for class 1=' Healthy controls' is 9 while false positives are 1 whereas, for class b=' Patients', true positives are 11 and false positives is 2 i.e. diagonal elements of matrix 9+11 =20 represents the correct instances classified and other elements 1+2 = 3 represents the incorrect instances. As a result of the classification process, an accuracy of 86.95% was obtained. Elapsed time is 0.204262 seconds.

3. Result and Discussion

The performance of classification algorithm is generally examined by measuring the accuracy of the classification. In this study, it is seen that artificial neural networks and Naïve Bayes algorithms can be used and good results can be obtained from classification algorithms. In our study, it was observed that breast cancer, which is an especially important type of cancer, was classified with an accuracy of 86.95 with ANN and 83.54 with Naïve Bayes algorithms using routinely acquired and controlled parameters from the patients, thus showing that the algorithms could be used for early breast cancer detection.

4. Conclusions

In this study, we suggest a model for the detection of breast cancer using ANN ve Naïve Bayes based on biomarkers. Glucose, Resistin, Age, BMI, HOMA, Leptin, Insulin, Adiponectin, MCP-1 were the default biomarkers for evaluating the ANN study. According to these markers, it is thought that they can be classified by ANN or Naïve Bayes.

References

[1]. Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

[2]. URL1, https://www.mayoclinic.org/diseases-conditions/breast -cancer/symptoms-causes/syc-20352470. Last Access(12.12.2018).

[3]. Nyante, S.J., et al., The association between mammographic calcifications and breast cancer prognostic factors in a population-based registry cohort. Cancer, 2017. 123(2): p. 219-227.

[4]. Crisóstomo, J., et al., Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. Endocrine, 2016. 53(2): p. 433-442.

[5]. Saritas, I., Prediction of breast cancer using artificial neural networks. Journal of Medical Systems, 2012. 36(5): p. 2901-2907.

[6]. URL2, http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#. Last Access (05.12.2018).

[7]. Bebis, G. and M. Georgiopoulos, Feed-forward neural networks. IEEE Potentials, 1994. 13(4): p. 27-31.

[8]. Shu, X., et al., Associations of obesity and circulating insulin and glucose with breast cancer risk: a Mendelian randomization analysis. International journal of epidemiology, 2018.

[9]. Anderson, K.N., R.B. Schwab, and M.E. Martinez, Reproductive risk factors and breast cancer subtypes: a review of the literature. Breast cancer research and treatment, 2014. 144(1): p. 1-10.

[10]. Guerrieri-Gonzaga, A., et al., Abstract P4-11-16: Low serum adiponectin level is an independent risk factor of DCIS in postmenopausal women at increased risk of breast cancer. 2015, AACR.

[11]. Hopfield, J.J., Artificial neural networks. IEEE Circuits and Devices Magazine, 1988. 4(5): p. 3-10.

[12]. Bhardwaj, A. and A. Tiwari, Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 2015. 42(10): p. 4611-4620.

[13]. Haykin, S.S., et al., Neural networks and learning machines. Vol. 3. 2009: Pearson Upper Saddle River.

[14]. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, Learning representations by back-propagating errors. nature, 1986. 323(6088): p. 533.

[15]. Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:1206.1121.