International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Comparative Study of Machine Learning Algorithms applied to Predictive Breast Cancer Data

Kedar Potdar[1], Rishab Kinnerkar[2]

[1]Student, Watumull Institute of Electronics Engineering and Computer Technology, (Mumbai University), R.G. Thadani Marg, Worli, Mumbai, India 400 018

[2]Student, Department of Electrical and Computer Engineering, Iowa State University, Ames, IA USA

Abstract: ***Diagnostic errors are the most frequent non-operative medical errors. Diagnosis should be more data-driven than trial-and-error. Machine Learning provides techniques for classification and regression purposes which can be used for solving diagnostic problems in different medical domains. Predictive analysis of fatal ailments like cancer using existing data can serve as a diagnosis tool for doctors. The paper aims at a comparative study of Machine Learning algorithms on a predictive breast cancer dataset. The algorithms used for comparison - Artificial Neural Networks (ANN), k-Nearest Neighbors (kNN) and Bayesian Network Classifiers – are supervised learning algorithms used widely for classification purposes and are chosen for their diversity. Based on analysis of this data, Artificial Neural Networks are better at classification with 97.4% accuracy than kNN and Bayesian Classifiers.***

Keywords: machine learning, medical diagnosis, breast cancer, neural networks, k nearest neighbors, Bayesian classifiers

1. Introduction

Routine breast cancer screening allows the disease to be diagnosed and treated prior to it causing noticeable symptoms. The goal of screening tests for breast cancer is to find it before it causes symptoms (like a lump that can be felt). Screening refers to tests and exams used to find a disease in people who don't have any symptoms [5]. Early detection means finding and diagnosing a disease earlier than waiting for symptoms to start [7].

The process of early detection involves examining the breast tissue for abnormal lumps or masses. If a lump is found, a fine-needle aspiration biopsy is performed, which uses a hollow needle to extract a small sample of cells from the mass. A clinician then examines the cells under a microscope to determine whether the mass is likely to be malignant or benign.

To Err Is Human,[6] the report from the National Academy of Sciences, Institute of Medicine (2000), estimated that as many as 98,000 people die every year in the US because of mistakes committed by medical professionals in hospitals. Diagnostic errors were the most frequent non-operative errors.

If machine learning could assist and eventually automate the identification of cancerous cells from the data obtained in biopsies, it would provide considerable benefit to the health system. An automated screening system might provide greater detection accuracy by removing the inherently subjective human component from the process.

2. Literature Review

Machine Learning has been documented as a classification mechanism for medical and other applications. In this section we furnish you with some of the relevant work which has already made its mark in this field of study.

In [1], the authors have implemented the Support Vector Machine (SVM) and kNN machine learning algorithms for diagnosis of respiratory pathologies using pulmonary acoustic signals on the RALE lung sound database and showed that the generalization capability of the kNN classifier is higher compared with that of SVM. The accuracy of kNN was found out to be 98.26% as compared to 92.19% of SVM. In [2], the authors carried out a comparative study on diabetes disease diagnosis using neural networks on a pima-diabetes dataset and show that multilayer neural networks with Levenberg–Marquardt (LM) algorithm outperformed other neural network based classifiers.

In [3], the authors performed a comparative study on Parkinson's Disease diagnosis and compared the results of DMNeural, Neural Network, Regression, and Decision tree classification models on a data of 197 Parkinson's disease patients. A total of 22 factors were compared and neural networks outperformed rest of the classification techniques with an accuracy of 92.9%.

In [4], the authors used Na.ve Bayes , Support vector machines Radial Basis Function (RBF) kernel, Radial basis neural networks, Decision trees J48 and simple CART for disease detection and found that SVM RBF kernel method outperformed other classifier techniques.

3. Data Acquisition and Preprocessing

3.1 Breast Cancer Dataset

The breast cancer data used for this paper was obtained from the Wisconsin Breast Cancer Diagnostic dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). This data was donated by researchers of the University of Wisconsin and includes the measurements from digitized images of fine-needle aspirate of a breast mass.

The breast cancer data includes 569 instances of cancer biopsies. Each record has 32 attributes. One attribute is an identification number, another is the cancer diagnosis, and 30 are numeric-valued laboratory measurements. The diagnosis is coded as "M" to indicate malignant or "B" to indicate benign. The values represent the characteristics of the cell nuclei present in the digital image [8].

The 30 numeric measurements comprise the mean, standard error, and worst (that is, largest) value for 10 different characteristics of the digitized cell nuclei. The 10 real valued features calculated for each cell nucleus are:

1) Radius (mean of distances from center to points on the perimeter)
2) Texture (standard deviation of gray-scale values)
3) Perimeter
4) Area
5) Smoothness (local variation in radius lengths)
6) Compactness (perimeter^2 / area - 1.0)
7) Concavity (severity of concave portions of the contour)
8) Concave points (number of concave portions of the contour)
9) Symmetry
10) Fractal dimension ("coastline approximation" - 1)

All feature values are recoded with four significant digits. Furthermore, there are no missing values.

Statistics of the classes in data:

Table 1: Statistics of dataset

Class	Instances	% distribution
Benign	357	62.74
Malignant	212	37.26
Total	569	100

3.2 Preprocessing and experimental setup

The first attribute, ID, was excluded as it is simply a unique identifier for each patient and not useful data. The numeric values in the breast cancer database are not normalized. The smoothness ranges from 0.05 to 0.16 and area ranges from 143.5 to 2501.0. Since the range of data is not normalized, it can negatively affect performance of the classifiers. Hence, numeric features are first normalized using the formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The values of diagnosis feature – "B" and "M" are converted to numeric values 0 and 1 respectively using dummy coding.

4. Methods

In this work, three different classifiers were used, namely Naïve Bayes classifier, Artificial Neural Networks (ANN) and k-nearest neighbour (kNN). A detailed description of the classifiers used can be found in this section.

4.1 Naïve Bayes (NB) Classifier

Naïve Bayes is a relatively simple machine learning technique based on probability models - Bayesian theorem. It belongs to the family of probabilistic classifiers in machine learning based on Bayes' theorem with a strong statistic independence assumed between the features.

$$P(h_k|x_j) = \frac{P(x_j|h_k)P(h_k)}{\Sigma_{i=0}^{n} x_j} ; 0 < k < n+1 ; \{i, j, k \epsilon Z\} \quad (2)$$

This classification technique analyses the relationship between each feature and the class for each instance to derive a conditional probability for the relationships between the feature values and the class [10].

The conceptual framework for Naïve Bayes is based on joint probabilities of features and classes to estimate the probabilities of a given document belonging to a given class.

Given a training set, the naïve Bayes algorithm first estimates the prior probability $P(C_j)$ for each class by counting how often each class occurs in the training data. For each attribute value xi can be counted to determine $P(x_i)$. Similarly, the probability $P(x_i|C_j)$ can be estimated by counting how often each value occurs in the class in the training data. When classifying a target tuple, the conditional and prior probabilities generated from the training set are used to make the prediction. Then estimate $P(t_i|C_j)$ by the following:

$$P(t_i|C_j) = \prod_{k=1}^{p}(x_{ij} \mid C_j) \quad (3)$$

To calculate $P(t_i)$ we estimate the likelihood that t_i is in each class. The probability that t_i is in a class is the product of the conditional probabilities for each attribute value. The class with the highest probability is the one chosen for the tuple [9].

4.2 k-Nearest Neighbors (kNN)

kNN is a lazy learning algorithm for instance based learning used to classify objects based on their closest training examples in the feature space [5]. An object is classified in a class to which its k-nearest neighbors belong. In the kNN algorithm, the classification of a new test feature vector is determined by the classes of its k-nearest neighbors [11].

Here, the kNN algorithm was implemented using Euclidean distance metrics to locate the nearest neighbor. The Euclidean distance metrics d(x,y) between two points x and y is calculated using the equation below. Where N is the number of features such that $x = \{x_1, x_2, x_3, \ldots, x_n\}$ and $y = \{y_1, y_2, y_3, \ldots, y_n\}$ [12].

$$d(x, y) = \Sigma_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (4)$$

In our case, the number of features for each instance is 30, i.e N = 30.
A rule of thumb is that k equals the square root of the number of points in the training data set [5]. Here, the value of k required for training 500 instances of biopsy data chosen is 22, as 222=484 which is the closest perfect square to 500.

4.3 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is an interconnected group of nodes that uses a computational model for information processing. It changes its structure based on external or internal information that flows through the

network. ANN can be used to model a complex relationship between inputs and outputs and find patterns in data. The output of ANN is determined by characteristics of the features and the weights associated with the interconnections among them. The connections between nodes are modified in the training process to adapt the network to desired outputs [14]. The neural network gains the experience initially by training the system to correctly identify pre-selected examples of the problem. The response of the neural network is reviewed and the configuration of the system is refined until the neural network's analysis of the training data reaches a satisfactory level. In addition to the initial training period, the neural network also gains experience over time as it conducts analyses on data related to the problem [15].

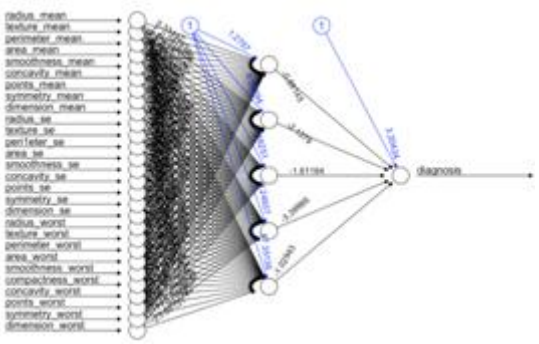

Figure 1: Artificial Neural Network trained on the data

5. Classification Accuracy

In this study, two methods of result verification have been used: the conventional method with one training set and one testing set and k fold cross-validation [17].

For the conventional method, 500 instances were used as training data set and remaining 69 instances were used as test set [18].

In k-fold cross validation, whole data is divided into k mutually exclusive and approximately equal size sets. The classification algorithm is trained and tested k times. In each test, one set is used for testing purposes, the remaining (k-1) sets are used for training. The average accuracy of all tests gives the test accuracy of the algorithm [19].
Here, the value of k chosen for cross validation is 3.

5.1 Comparison of results

We conducted the simulation of kNN, ANN and NB algorithms on predictive breast cancer data. Table 2 demonstrates classification accuracy of the algorithms for the 3-fold cross-validation test:

Table 2: Accuracy of 3-fold Cross Validation

Method	1st Fold % accuracy	2nd Fold % accuracy	3rd Fold % accuracy	Averages % accuracy
k-Nearest Neighbors	96.3	96.8	94.8	96.0
Artificial Neural Networks	98.4	97.4	96.3	97.4
Naïve Bayes	93.4	92.4	95.8	93.9

Table 3 demonstrates the accuracy of the algorithms for the Conventional validation technique with 500 training samples and 69 test samples:

Table 3: Accuracy of the Conventional Method

Method	*Accuracy (%)*
k-Nearest Neighbors	95.6
Artificial Neural Networks	100
Naïve Bayes	93.8

This comparative study shows that the classification accuracy of ANN is higher than that of kNN and NB classification algorithms in the diagnosis of predictive breast cancer data from the UCI Wisconsin Breast Cancer datase. We can see that ANN is the most accurate algorithm having classified the samples with 97.4% accuracy in 3-fold cross validation and with a 100% accuracy in Conventional validation. kNN comes second and the classification accuracy of the NB classifier is least in both the cases.

The limitation of this study is the size of data used. The number of samples used for training and testing is low. The analysis of data with respect to clinical setting should be carried out with a larger dataset.

6. Conclusion

In this paper, we compared the classification accuracy of 3 Machine Learning algorithms – kNN, ANN and NB on UCI Wisconsin Breast Cancer dataset. The aim of this comparative study was to find the most accurate machine learning tool that can act as a tool for diagnosis of breast cancer. According to the prediction results, ANN has highest accuracy for the given dataset. This shows that ANN can be used for prediction of breast cancer as compared to kNN and NB.

References

[1] R.Palaniappan, K.Sunderaj, S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals", BMC Bioinformatics, 15.1, pp. 1-8, 2014.
[2] H. Temurtas, N. Yumusak, and E. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks." Expert Systems with applications Vol. 36 No. 4, pp. 8610-8615, 2009.
[3] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, Vol. 37 No .2, pp. 1568-1572, 2010.
[4] S. Aruna, Dr S.P. Rajagopalan, L.V. Nandakishore, "An empirical comparison of supervised learning algorithms in Disease Detection", International Journal of Information Technology Convergence and Services (IJITCS), Vol .1 No. 4, August 2011.
[5] Brett Lantz, Machine Learning with R, Packt Publishing Limited, 2013. ISBN - 978-1782162148.
[6] L.T. Kohn, J.M. Corrigan, M.S. Donaldson, "To err is human: Building a safer health system.", National Academy Press , Washington D.C., 2000.
[7] American Cancer Society, "Detailed Guide: Breast Cancer", cancer.org, 2014 [Online]. Available: www.cancer.org/Cancer/BreastCancer/DetailedGuide/index. [Accessed: Sept. 10, 2016].

[8] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming", SIAM News, Vol. 23 No. 5, September 1990, pp 1 & 18.
[9] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers." Machine learning Vol. 29 No. 2-3, 1997,pp. 131-163.
[10] I. Rish, "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 2001.
[11] S.A. Medjahed, T. A. Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules." International Journal of Computer Applications Vol. 62 No. 1 , 2013.
[12] Sarkar, Manish, and Tze-Yun Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem." Proceedings of the AMIA Symposium. American Medical Informatics Association, pp. 759-763, 2000.
[13] Liu, Juanmei, et al, "Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population." BMC medical informatics and decision making Vol. 13 No. 1, pp. 1, 2013.
[14] Y. Bar-Yam, Dynamics of Complex Systems, Westview Press, 1997.
[15] P. Stavroulakis, M. Stamp, Handbook of Information and Communication Security, Springer Publishing Company, 2010.
[16] A. B. Watkins,“AIRS: A resource limited artificial immune classifier”,. Diss. Mississippi State University, 2001.
[17] T. Hastie, R. Tibshirani, and J. Friedman, Springer Series in Statistics Springer New York Inc., New York, NY, USA, 2001.
[18] Er O, Yumusak N, Temurtas F,“Chest diseases diagnosis using artificial neural networks”, Expert Systems Appl., Vol. 37, No. 12, pp.7648–55. 2010.
[19] A. Chaudhary, S. Kolhe, and Raj Kamal, "Machine Learning Classification Techniques: A Comparative Study.", International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol. 2 No. 4, pp. 2319-2526, 2013.
[20] N.K. Bose, P. Liang, Neural network fundamentals with graphs, algorithms, and applications, McGraw Hill, New York, NY, USA, 1996.

Author Profile

Kedar Potdar is pursuing B.E. Computer Engineering from Watumull Institute of Electronics Engineering & Computer Technology, Mumbai University. He works as a freelance business strategy consultant and is interested in Data Analytics, Machine Learning and Artificial Intelligence.

Rishab Kinnerkar is pursuing B.S. in Computer Engineering from Iowa State University. He studied at Birla Institute of Technology (BITS) Pilani (Dubai) till 2016, post which he took a transfer to Iowa State University. He too works as a freelance business strategy consultant and is interested in Finance, Machine Learning and Artificial Intelligence.