RESEARCH REVIEW JOURNALS
Volume-04
Issue-07
July-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[Peer Reviewed Journal]

Comparative Analysis of Breast Cancer Data Using C4.5 Classifier and Naive Bayes Algorithm

Jagannathan D

Teacher, Sakthi Vinayakar Hindu Vidyalaya, Thoothukudi (India)

ARTICLE DETAILS

Article History
Published Online: 15 July 2019

Keywords
WBCD,UCI,C4.5,NaïveBayes,Bayesian Classifier.

***Corresponding Author**
Email: jagan_dj1992[at]rediffmail.com

ABSTRACT

This study is aimed to identify the breast cancer using data mining classification methods. The dataset named Wisconsin Breast Cancer Database (WBCD) are obtained from university of California Irvine (UCI) respiratory and The Wisconsin Madison University. By using this dataset a comparison of two different classifiers that can be used machine learning algorithms, namely the Naïve Bayes algorithm and C4.5 Classification algorithm. Bayesian classifiers are the statistical classifiers. Bayesian classifiers predicts class membership probabilities such as the probability that a given tuple belongs to a particular class. The decision trees that are generated by using C4.5 classifier can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. In order to measure the performance, the holdout method is the simplest kind of cross validation. The data set can be separated into two sets, called the training set and the testing set. The function approximator allows to fit a function using the training set. The prediction of the output values for the data in the testing set is done by the function approximator (it has never seen these output values before).

1. Introduction

Today, Breast cancer is one of the major health problems for women across the world. Although the risk factors vary for different communities, it remains one of the biggest health concerns for women across the globe. Nowadays, the field of computer science and medicine are nested in order to provide a proper prognosis or diagnosis of the human diseases. Many methods are used for the identification of such health problems. Data mining is a critical procedure for registering applications in the space region of medicine. In data mining, research on breast cancer has been one of the important research topics in the field of medicine during the recent years. The classification of Breast Cancer data can be useful to identify the result of some diseases or even discover the genetic behavior of tumors. Many techniques are available to predict and classify breast cancer pattern. This work empirically compares performance of different classification rules that are suitable for direct interpretation of their results. For this reason the use of classifier systems in medical diagnosis has increased gradually.

2. Methodology

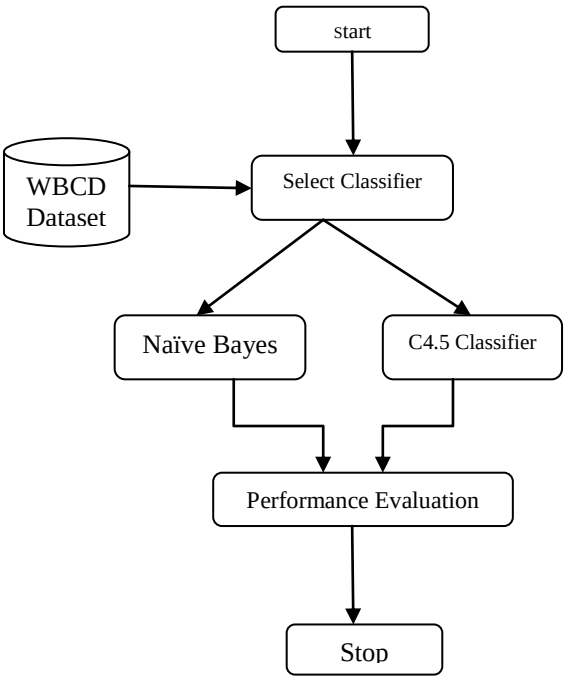

Fig.1 Block Diagram for WBCD dataset classification using Naïve bayes and C4.5 Classifier

A. Naïve Bayes classifier

A Naive Bayes classifier is a probabilistic classifier which applies Bayes' theorem with strong independence assumptions. The performance goal of this classifier is to predict the class of test instances as accurately as possible. This classifier is termed naive because it is based on two simplifying common assumptions: firstly, it usually assumes that the predictive attributes are conditionally independent of the given the class and secondly, the values of numeric attributes are normally distributed within each class. Naive Bayes' treats discrete and continuous attributes differently.

For each discrete attribute, the probability that the attribute X will take on the particular x when the class is c is modeled by a single real number between 0 and 1. In contrast, each continuous attribute is modeled by some continuous probability distribution over a range of that attribute's values. Let us consider C to be the random variable denoting the class of an instance and X to be a vector of random variables denoting the observed attribute values. Let c be considered as a particular class label and x representing a particular observed attribute value. If there is a test case x to classify, then the probability of each class given the vector of observed values for the predictive attributes may be obtained using the Bayes' theorem:

$$\mathbf{p(C = cj \mid X = x) = \frac{p(C = c)p(X = xj \mid C = c)}{p(X = x)}}$$

and then predicting the most probable class. Because the event is a conjunction of attribute values assignments, and because of the attributes conditional independence assumption, the following equation may be written:

$$\mathbf{p(X = xj \mid C = c) = \pi\, p(Xi = xij \mid C = c)}$$

which is quite simple to calculate for training and test data [1].

B. Decision Tree Classification

Decision tree classification approach is one of the most useful approach in classification problems. [3]. A decision tree is like a a flow chart structure where each node denotes a test on an attribute value, each branch representing an outcome of the test and tree leaves representing classes or class distribution. A decision tree is a predictive model which is most often used for classification. Decision trees would partition the input space into cells where each cell belongs to one class only. The partitioning is always represented as a sequence of tests. Each interior node in the decision tree will test the value of some input variable, and the branches from the node in particular are labelled with the possible results of the test. The leaf nodes of the tree represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is finally reached. Decision tree is represented in figure 2.

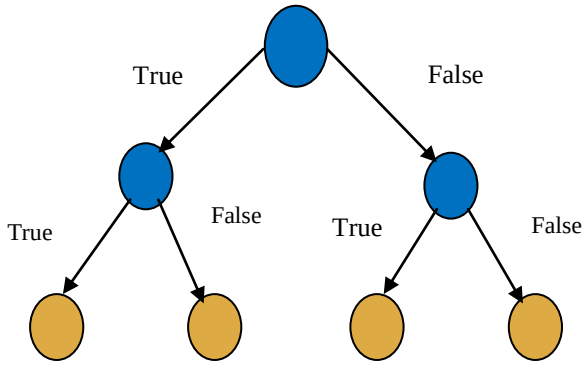

Fig.2 Decision Tree

In this classification method used in different type algorithm to classify the data sets, the algorithms are: [4]

- ID3 (Iterative Dichotomise)
- C4.5 (a Successor of ID3)
- Classification and Regression Trees (CART)

The algorithm usually follows a top-down approach, which would start with a training set of tuples and their associated class labels.

Advantages: Rules can be generated which are quite easy to interpret and understand. It is normally scalable for large database because the tree size is independent of the database size. Each tuple in the database must be filtered through the tree, and time is actually proportional to the height of the tree.

Disadvantages: It does not handle continuous data. Handling missing data is difficult because correct branches in tree could not be taken as the labels.

C. C4.5 Classifier

C4.5 algorithm is used to generate a decision tree which was developed by Ross Quinlan. C4.5 algorithm is just an extension of Quinlan's earlier ID3 algorithm. The decision trees that are generated by C4.5 can be used for classification, and for this reason, C4.5 is referred to as a statistical classifier. One limitation of ID3 algorithm is that it is sensitive to features with large numbers of values. This should be overcome if we are going to use ID3 as an Internet search agent. This is addressed by borrowing from the C4.5 algorithm, an ID3 extension.ID3's sensitivity to the features with large number of values is illustrated by Social Security numbers. Since Social Security numbers are unique for each and every individual, testing on its value will always yield us low conditional entropy values. However, this may not be a useful test. To overcome this particular problem, C4.5 uses a metric called "information gain," which is defined by subtracting the conditional entropy from the base entropy; that is, Gain (P|X) =E (P)-E (P|X). This computation does not produce anything new. Indeed it allows us to measure a gain ratio. Gain ratio, defined as Gain Ratio (P|X) =Gain (P|X)/E(X), where(X) is the entropy of the examples that are relative only to the attribute. This enhanced method of tree pruning helps to reduces misclassification errors due noise or too much detail in the training data set. Like ID3 the data is been sorted at every node of the tree so as to determine the best splitting attribute. It uses gain ratio impurity method in order

to evaluate the splitting attribute. Decision trees are being built in C4.5 by using a set of training data or data sets as in ID3. At each node of the tree, C4.5 would choose one attribute of the data that effectively splits its set of samples into subsets enriched in one class or the other class. Its criterion is the normalized information gain (difference in entropy) that results in choosing an attribute for splitting the data. The attribute that has the highest normalized information gain is chosen to make the decision.

Pseudo Code:

1. Check for base cases.
2. For each attribute a calculate:

i. Normalized information gain from splitting on attribute

3. Select the best a, attribute that would have the highest information gain.
4. Create a decision node that splits on best of a, as rot node.
5. Recurs on the sub lists obtained by splitting on best of a and add those nodes as children node.

3. Experimental Result

To evaluate the effectiveness of our method, experiments on Wisconsin Breast Cancer Database WBCD is conducted. This database was obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg. This is a publicly available dataset on the Internet. Table.1 shows the descriptions of database.

Table.1. Descriptions of Database

No	Attributes	No of Attributes
1.	Number of instances	699
2	Number of attributes	10
3	Attributes 2 through 10	Instances
4	Classes	1. benign 2. malignant
5	Class distribution	1.Benign:458(65.5%) 2. Malignant: 241 (34.5%)

Attribute Information of WBCD Dataset are briefly summarized in Table.2.

Table.2. Attribute Information

No	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 -10
3.	Uniformity of Cell Size	1 -10
4.	Uniformity of Cell Shape	1 -10
5.	Marginal Adhesion	1 -10
6.	Single Epithelial Cell Size	1 -10
7.	Bare Nuclei	1 -10
8.	Bland Chromatin	1-10
9.	Normal Nucleoli	1-10
10.	Mitoses	1-10
11.	Class	(2 for benign, 4 for malignant)

4. Performance Evaluation

A. Accuracy

The accuracy result of Naive Bayes and C4.5 classification algorithm is shown in Figure.3.

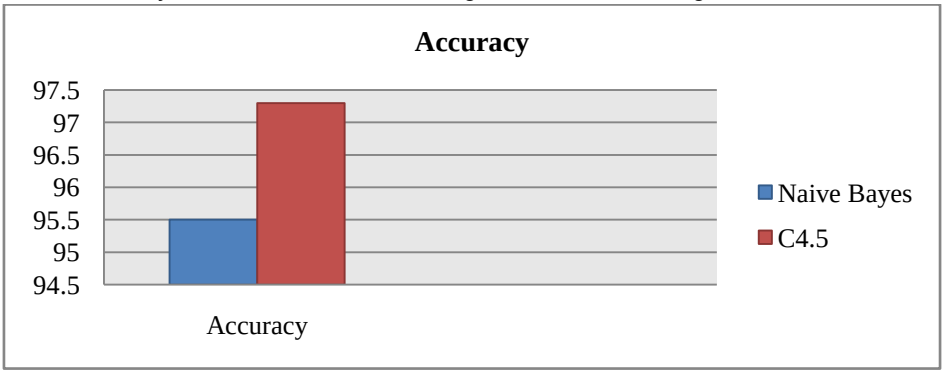

Fig.3 Accuracy of Naive Bayes and C4.5 Algorithm

B. Sensitivity and Specificity

The Sensitivity and Specificity of Naive Bayes and C4.5 classification algorithm are shown in Figure.4.

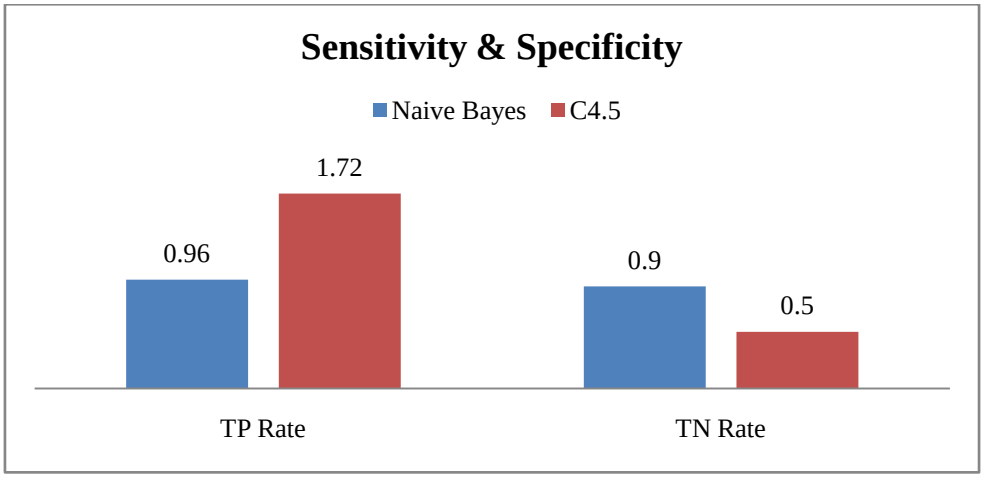

Fig5.4 Sensitivity and Specificity rate of Naive Bayes And C4.5 Algorithm

C. PPV, NPV and ROC

a.

The performance value of Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Receiver Operator Characteristic test (ROC) for Naive Bayes and C4.5 classifier are shown in Table 5.3.

Table5.3 performance value of PPV, NPV and ROC

	PPV	NPV	ROC
Naive Bayes Classifier	0.491	0.008	0.952
C4.5 Classifier	0.583	0.012	1.114

5. Performance Metrics

In this study, the accuracy of two data mining techniques is compared. Although such metrics are used more often in the field of information retrieval, it's considered as they are related to other existing metrics such as specificity and sensitivity. These metrics could be derived by using the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

A. Confusion Matrix

In the soft computing field, the confusion matrix is specific table layout illustrating a classification algorithm's performance. It is a tabular representation that permits more thorough analysis than accuracy. Each attribute of the matrix denotes the patterns in the anticipated data class whereas every tuple designates the patterns in the definite class.

1. True-Positive (TP) indicates the amount of 'positive 'patterns categorized as 'positive.'
2. False-Positive (FP) means the amount of 'negative 'patterns categorized as 'positive.'
3. False-Negative (FN) denotes the amount of 'positive 'patterns categorized as 'negative.'
4. True-Negative (TN) implies the amount of 'negative 'patterns categorized as 'negative.'

Table.4 below displays a confusion matrix layout using a two-class classifier is having the following cells as:

Table.4. A confusion matrix layout for a two-class classifier

Actual Class	**Predicted Class**	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

B. Accuracy Measures

A two-class confusion matrix defines several standard terms. The accuracy (i.e. classification accuracy) is the sum of the correctly classified examples divided by the total number of examples present. The following equation calculates this as:

$$\textbf{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

C. Sensitivity

A sensitivity analysis is a technique that is used to determine how a different value of an independent variable impacts a particular dependent variable under a given set of assumptions. Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are exactly identified as having the condition).

$$\textbf{Sensitivity}=\frac{TP}{TP+FN}$$

D. Specificity

Specificity (also called the true negative rate) is used to measure the proportion of negatives that are exactly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

$$\textbf{Specificity}=\frac{TN}{TN+FP}$$

E. Positive Predictive Value

The Positive Predictive Value (Precision) is the proportion of positive cases that are exactly identified. The following equation calculates this as:

$$\textbf{PPV}=\frac{TP}{TP+FP}$$

F. Negative Predictive Value

The Negative Predictive Value is the proportion of negative cases that are exactly identified. The following equation calculates this as:

$$\mathbf{NPV} = \frac{\mathbf{TN}}{\mathbf{TN+FN}}$$

G. Receiver Operator Characteristic test (ROC)

ROC is a plotting of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test.

$$\mathbf{ROC} = \frac{\mathbf{Sensitivity+Specificity}}{\mathbf{2}}$$

6. Conclusion

In this paper, the accuracy is evaluated based on the selected classifier algorithm like Naïve Bayes and C4.5 classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient ensemble classifiers for Medical applications. The performance of C4.5 classifier shows the high level performance compared with Naive Bayes classifiers. The values to measure the performance of the methods (i.e. accuracy, sensitivity, specificity) are derived from the confusion matrix and showed in Figure.3 and Figure.4. The values of Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Receiver Operator Characteristic (ROC) are showed in Table.3. It was found that C4.5 classifier model produced highest accuracy i.e. 97.3% which is so far highest. Other classifier like Naive Bayes were far less accurate compared to C4.5 classifier.

References

[1] V.N. Vapnik and A. Chervonenkis, “A note on one class of perceptrons”, Automation and Remote Control, 25, 1964.

[2] J.R.Quinlan, ”Induction of decision tree”. Journal of Machine Learning 1, 1986, Pg.no:81-106.

[3] V. N. Vapnik, “The Nature of Statistical Learning Theory”, Springer, New York, NY, USA, 1995.

[4] Mark A. Hall, Lloyd A. Smith, Feature Subset Selection: A Correlation Based Filter Approach, In 1997 International Conference on Neural Information Processing and Intelligent Information Systems (1997), pp. 855-858.

[5] Han and Kamber, - “Data Mining; Concepts and Techniques”, Morgan Kaufmann Publishers, 2000.

[6] H. Blockeel and J. Struyf, “Efficient algorithms for decision tree cross-validation”, Proceedings of the Eighteenth International Conference on Machine Learning (C. Brodley and A. Danyluk, eds.), Morgan Kaufmann, 2001, pp. 11-18.

[7] Witten H.I., Frank E., —Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.

[8] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Günes, A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS), In Proceedings of ICNC (2)'2005. pp.830~838.

[9] J. Han and M. Kamber, —Data Mining—Concepts and Techniquell (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[10] A.Bellachia and E.Guvan,“Predicting breast cancer survivability using data mining techniques”, Scientific Data Mining Workshop, inconjunction with the 2006 SIAM Conference on Data Mining, 2006.

[11] Asuncion and Newman,"UCI machine learning repository", 2007.

[12] Deisy.C, Subbulakshmi.B, Baskar S, Ramaraj.N, Efficient Dimensionality Reduction Approaches for Feature Selection, Conference on Computational Intelligence and Multimedia Applications, 2007.

[13] Chi C.L., Street W.H. and Wolberg W.H., “Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets”, Annual Symposium Proceedings / AMIA Symposium, 2007.

[14] Daniele Soria, Jonathan M. Garibaldi, Elia Biganzoli and Ian O. Ellis, “A Comparison of Three Different Methods for Classification of Breast Cancer Data”, “:https://www.researchgate.net/publication/221226489, Jan 2008”.G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

