KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST CANCER

S. Aruna[1], Dr S.P. Rajagopalan[2] and L.V. Nandakishore[3]

[1,2]Department of Computer Applications, Dr M.G.R University, Chennai-95, India
arunalellapalli@yahoo.com[1] , sasirekharaj@yahoo.co.in[2]
[3]Department of Mathematics, Dr M.G.R University, Chennai-95, India
lellapalliarunakishore@gmail.com

ABSTRACT

In this paper, we study the performance criterion of machine learning tools in classifying breast cancer. We compare the data mining tools such as Naïve Bayes, Support vector machines, Radial basis neural networks, Decision trees J48 and simple CART. We used both binary and multi class data sets namely WBC, WDBC and Breast tissue from UCI machine learning depositary. The experiments are conducted in WEKA. The aim of this research is to find out the best classifier with respect to accuracy, precision, sensitivity and specificity in detecting breast cancer.

KEYWORDS

J48, Naïve Bayes, RBF neural networks, Simple Cart, Support vector machines.

1. INTRODUCTION

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases [1]. The term Data Mining or Knowledge Discovery in databases has been adopted for a field of research dealing with the automatic discovery of implicit information or knowledge within databases [2]. Machine learning refers to a system that has the capability to automatically learn knowledge from experience and other ways [3]. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends [4]

In this paper we analyze the performance of supervised learning algorithms such as Naïve Bayes, SVM Gaussian RBF kernel, RBF neural networks, Decision trees.J48and simple CART. These algorithms are used for classifying the breast cancer datasets WBC, WDBC, Breast tissue from UCI Machine learning depository (http://archive.ics.uci.edu/ml).We conducted our experiments using WEKA tool. These algorithms have been used by many researchers and found efficient in some aspects. The goal of this research is to find the best classifier which outperforms other classifiers in all the aspects.

This paper is organized as follows. Section 2 gives a brief description about the data mining algorithms and section 3 gives the description about the datasets used for this experiment. Section 4 gives the results obtained and the concluding remarks are given in section 5 to address further research issues.

D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 37–45, 2011.

DOI: 10.5121/csit.2011.1205

2. DATA MINING ALGORITHMS

2.1. Naive Bayes

Naive Bayes classifier is a probabilistic classifier based on the Bayes theorem, considering a strong (Naive) independence assumption. Thus, a Naive Bayes classifier considers that all attributes (features) independently contribute to the probability of a certain decision. Taking into account the nature of the underlying probability model, the Naive Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect [5], [6]. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}. \qquad (1)$$

where P is the probability, C is the class variable and F_1.....F_n are Feature variables F_1 through F_n The denominator is independent of C.

2.2. Decision trees CART and J48

Decision trees are supervised algorithms which recursively partition the data based on its attributes, until some stopping condition is reached [4] Decision Tree Classifier (DTC) is one of the possible approaches to multistage decision-making. The most important feature of DTCs is their capability to break down a complex decision making process into a collection of simpler decisions, thus providing a solution, which is often easier to interpret [7].

The classification and regression trees (CART) methodology proposed by [8] is perhaps best known and most widely used. CART uses cross-validation or a large independent test sample of data to select the best tree from the sequence of trees considered in the pruning process. The basic CART building algorithm is a greedy algorithm in that it chooses the locally best discriminatory feature at each stage in the process. This is suboptimal but a full search for a fully optimized set of question would be computationally very expensive. The CART approach is an alternative to the traditional methods for prediction [8] [9] [10]. In the implementation of CART, the dataset is split into the two subgroups that are the most different with respect to the outcome. This procedure is continued on each subgroup until some minimum subgroup size is reached.

Decision tree J48 [11] implements Quinlan's C4.5 algorithm [12] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labelled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class.

J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation

2.3. Radial Basis Neural Networks

Radial Basis Function (RBF networks) is the artificial neural network type for application of supervised learning problem [13]. By using RBF networks, the training of networks is relatively fast due to the simple structure of RBF networks. Other than that, RBF networks are also capable of universal approximation with non-restrictive assumptions [14]. The RBF networks can be implemented in any types of model whether linear on non-linear and in any kind of network whether single or multilayer [13].

The design of a RBFN in its most basic form consists of three separate layers. The input layer is the set of source nodes (sensory units). The second layer is a hidden layer of high dimension. The output layer gives the response of the network to the activation patterns applied to the input layer. The transformation from the input space to the hidden-unit space is nonlinear. On the other hand, the transformation from the hidden space to the output space is linear [15]. A mathematical justification of this can be found in the paper by Cover [16].

2.4. Support Vector Machines

Support vector machines (SVM) are a class of learning algorithms which are based on the principle of structural risk minimization (SRM) [17] [18]. SVMs have been successfully applied to a number of real world problems, such as handwritten character and digit recognition, face recognition, text categorization and object detection in machine vision [19],[20],[21]. SVMs find applications in data mining, bioinformatics, computer vision, and pattern recognition. SVM has a number of advanced properties, including the ability to handle large feature space, effective avoidance of over fitting, and information condensing for the given data set.etc.[22]

Each kind of classifier needs a metric to measure the similarity or distance between patterns. SVM classifier uses inner product as metric. If there are dependent relationships among pattern's attributes, such information will be accommodated through additional dimensions, and this can be realized by a mapping [23]. In SVM literature, the above course is realized through kernel function

$$k(x, y) = \langle \phi(x), \phi(y) \rangle \qquad (2)$$

Kernels can be regarded as generalized dot products [23]. For our experiments we used Gaussian RBF kernel. A Gaussian RBF kernel is formulated as

$$k(x, y) = \exp\left[\frac{- \| x - y \|^2}{2\sigma^2} \right] \qquad (3)$$

3. DATASETS DESCRIPTION

3.1 Wisconsin Diagnostic Breast Cancer Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Number of instances: 569, Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)

Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" -1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius. All feature values are recoded with four significant digits. Class distribution: 357 benign, 212 malignant

3.2 Wisconsin Breast Cancer Dataset

This has 699 instances (Benign: 458 Malignant: 241) of which 16 instances has missing attribute values removing that we have 683 instances of which 444 benign and 239 are malignant. Features are computed from a digitized image of a Fine Needle Aspiration (FNA) of a breast mass. Table 1 presents the description about the attributes of the WBC dataset

Table 1. Description about the attributes of the WBC dataset

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)

3.3 Breast Tissue Dataset

This is a dataset with electrical impedance measurements in samples of freshly excised tissue from the Breast. It consists of 106 instances. 10 attributes: 9 features+1class attribute. Six classes of freshly excised tissue were studied using electrical impedance measurements. Table 2 presents the details about the 6 classes and number of cases that belong to those classes.

Table 2. Description about the 6 classes of breast tissue dataset

Class	# of cases
Car Carcinoma	21
Fad Fibro-adenoma	15
Mas Mastopathy	18
Gla Glandular	16
Con Connective	14
Adi Adipose	22

Impedance measurements were made at the frequencies: 15.625, 31.25, 62.5, 125, 250, 500, 1000 KHz. These measurements plotted in the (real, -imaginary) plane constitute the impedance spectrum from where the features are computed. Table 3 presents the description about the attributes of the breast tissue dataset

Table 3. Description about the attributes of the breast tissue dataset

Id	Attribute	Description
1	I0	Impedivity (ohm) at zero frequency
2	PA500	phase angle at 500 KHz
3	HFS	high-frequency slop e of phase angle
4	DA	impedance distance between spectral ends
5	AREA	area under spectrum
6	A/DA	area normalized by DA
7	MAX IP	maximum of the spectrum
8	DR	distance between I0 and real part of the maximum frequency point
9	P	length of the spectral curve

4 RESULTS

From the confusion matrix to analyze the performance criterion for the classifiers in detecting breast cancer, accuracy, precision (for multiclass dataset), sensitivity and specificity have been computed to give a deeper insight of the automatic diagnosis [24]. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. The sensitivity is the measure of the ability of a prediction model to select instances of a certain class from a data set. The specificity corresponds to the true negative rate which is commonly used in two class problems. Accuracy, precision, sensitivity and specificity are calculated using the equations 4, 5, 6 and 7 respectively, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

Table 4, 5, 6 shows the accuracy percentage for WBC, Breast tissue and WDBC datasets respectively. From the results we can see that all the classifiers except SVM-RBF kernel have varying accuracies but SVM-RBF kernel always has higher accuracy than the other classifiers for both binary and multiclass datasets.

Table 4. Accuracy percentage for WBC dataset

Algorithm	**Accuracy (%)**
Naïve Bayes	96.50
RBF networks	96.66
Trees-J48	94.59
Trees-CART	94.27
SVM-RBF kernel	96.84

Table 5. Accuracy percentage for Breast tissue dataset

Algorithm	**Accuracy (%)**
Naïve Bayes	94.33
RBF networks	92.45
Trees-J48	95.28
Trees-CART	96.22
SVM-RBF kernel	99.00

Table 6. Accuracy percentage for WDBC dataset

Algorithm	**Accuracy (%)**
Naïve Bayes	92.61
RBF networks	93.67
Trees-J48	92.97
Trees-CART	92.97
SVM-RBF kernel	98.06

WBC and WDBC datasets are binary class datasets whereas breast tissue is a 6 class dataset. Hence for binary class datasets for measuring the performance criterion sensitivity and specificity are calculated, for multiclass dataset sensitivity and precision are calculated. Table 7 and 8 shows the performance criterion such as for WBC and WDBC datasets respectively. Table 9 and 10 shows the percentage of sensitivity and precision for Breast tissue dataset.

Table 7. Performance Criteria for WBC

Algorithm	**Sensitivity(%)**	**Specificity(%)**
Naïve Bayes	95.7	97.8
RBF Networks	95.9	97.8
Trees-J48	94.4	92.6
Trees-CART	94.4	93.9
SVM-RBF kernel	97.2	97.8

Table 8. Performance Criteria for WDBC

Algorithm	**Sensitivity(%)**	**Specificity(%)**
Naïve Bayes	89.6	94.3
RBF Networks	90.0	95.7
Trees-J48	91.5	93.8
Trees-CART	89.1	95.2
SVM-RBF kernel	95.7	99.4

Table 9. Performance Criteria (sensitivity) for Breast tissue

Algorithm	**Sensitivity(%)**					
	Car	**Fad**	**Mas**	**Gla**	**Con**	**Adi**
Naïve Bayes	100	100	88.8	93.7	92.8	90.9
RBF Networks	100	80.0	88.8	93.7	92.8	95.4
Trees-J48	95.2	93.3	94.4	100	92.8	95.4
Trees-CART	100	93.3	94.4	93.7	92.8	100
SVM-RBF kernel	100	93.3	100	100	100	100

Table 10 Performance Criteria (precision) for Breast tissue

Algorithm	**Precision(%)**					
	Car	***Fad***	***Mas***	***Gla***	***Con***	***Adi***
Naïve Bayes	95.4	93.7	94.1	100	86.6	95.2
RBF Networks	91.3	100	84.2	93.7	92.8	95.4
Trees-J48	100	93.3	94.4	94.1	92.8	95.4
Trees-CART	100	100	94.4	93.7	92.8	95.6
SVM-RBF kernel	95.4	100	100	100	100	100

From the results we can see that the percentage of sensitivity, specificity and precision of SVM-RBF kernel is higher than that of other classifiers. SVM-RBF kernel always outperforms than the other classifiers in performance for both binary and multiclass datasets.

5 CONCLUSION

In this paper we compared the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM RBF kernel, RBF neural networks, Decision trees J48 and Simple CART. The aim of this research is to find a best classifier. All the classifiers are used to classify the breast cancer datasets namely WBC, WDBC and Breast tissue. The experiments were conducted in WEKA. The results are compared and found that SVM RBF Kernel outperforms other classifiers with respect to accuracy, sensitivity, specificity and precision for both binary and multiclass datasets. In future work we propose to analyze the linear and non linear SVM with and without dimensionality reduction techniques.

REFERENCES

[1] Julie M. David and Kannan Balakrishnan, (2010) "Significance of Classification Techniques In Prediction Of Learning Disabilities", *International Journal of Artificial Intelligence & Applications (IJAIA),* Vol.1, No.4.

[2] S.J. Cunningham, G Holmes, (1999) "Developing innovative applications in agricultural using data mining". *In: The Proceedings of the Southeast Asia Regional Computer Confederation Conference.*

[3] D.K. Roy, L.K. Sharma, (2010) "Genetic k-Means clustering algorithm for mixed numeric and categorical data sets", *International Journal of Artificial Intelligence & Applications*, Vol 1, No. 2, pp 23-28.

[4] H. Jiawei and K. Micheline, (2008) *Data Mining-Concepts and Techniques*, Second Edition, Morgan Kaufmann - Elsevier Publishers, ISBN: 978-1-55860-901-3.

[5] S. Belciug, (2008) "Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection", *Proc. 8th International conference on Artificial Intelligence and Digital Communications*.

[6] F. Gorunescu, (2006) *Data Mining: Concepts, models and techniques*, Blue Publishing House, Cluj Napoca.

[7] T. Pang-Ning, S. Michael, K. Vipin, (2008), *Introduction to Data Mining*, Low Price edn. Pearson, Education, Inc., London, ISBN 978-81-317-1472-0.

[8] L. Breiman, J. Friedman., R. Olshen, C. Stone, (1984), *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[9] D.Steinberg., and P.L. Colla, (1995) "CART: Tree-Structured Nonparametric Data Analysis", *Salford Systems:* SanDiego, CA.

[10] D.Steinberg., and P.L. Colla, (1997) "CART-Classification and Regression Trees", *Salford Systems:* San Diego, CA.

[11] Ross J. Quinlan, (1992) " Learning with Continuous Classes". *5th Australian Joint Conference on Artificial Intelligence*, Singapore, 343-348.

[12] Ross Quinlan, (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

[13] M. J. L. Orr, (1996) *Radial Basis Function Networks*. Edinburgh, Scotland.

[14] J. Park, and I.W. Sandberg, (1993) "Approximation and Radial-Basis-Function Networks", *Neural Computation.* 5, 305-316.

[15] S. Haykin, (1994) *Neural Networks a Comprehensive Foundation*, New Jersey, PrenticeHall.

[16] T. M. Cover, (1965) "Geometrical and Statistical Properties of Systems of Linear with Applications in Pattern Recognition," *IEEE Transactions on Electronic Computers EC-14*, pp. 326-334.

[17] Vladimir N. Vapnik. (1998) *Statistical Learning Theory*. New York: Wiley.

[18] Vladimir N. Vapnik. (1995) *The Nature of Statistical Learning Theory*, New York: Springer-Verlag,

[19] C. Campbell, N. Cristianini, and J.Shawe-Taylor, (1999) "Dynamically Adapting Kernels in Support Vector Machines", *Advances in Neural Information Processing Systems*,**Vol.** 11. MIT Press, 204-210.

[20] Corinna Cortes,Valdimir Vapnik, (1995) "Support-Vector Networks" *Machine Learning*, 20, 273-297.

[21] M,Pontil ,A.Verri, (1998) "Support Vector Machines for 3 D object recognition**."** *IEEE T Pattern Anal* , 20(6):637-646.

[22] You et al, (2010) "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network**"** *BMC Bioinformatics,* 11:343.

[23] Bernhard Schölkopf and Alex Smola, (2002) *Learning with kernels*. MIT Press, Cambridge, MA

[24] D.G. Altman, J.M. Bland, (1994) "Diagnostic tests Sensitivity and specificity". *BMJ* 308 (6943):1552.

Authors

S. Aruna is a graduate in Siddha Medicine from Dr M.G.R Medical University, Guindy, Chennai. Inspired by her software knowledge Prof Dr L.V.K.V Sarma (Retd HOD, Maths, I.I.T. Chennai) insisted her to do higher studies in computer science. The author completed her PGDCA as University first and gold medallist and MCA with first class, distinction. The author is presently a research scholar in Dr M.G.R University, Maduravoil, Chennai-95 in Dept of Computer Applications under Dr S.P. Rajagopalan PhD. She is also School first and Science first in class X. Her research interests include Semi supervised learning, SVM, Bayesian classifiers and Feature selection techniques**.**

Dr S.P.Rajagopalan PhD is Professor Emeritus in Dr M.G.R University, Maduravoil, Chennai-95, India. He was former Dean, College Development council, Madras UniversiChennai, India. Fifteen scholars have obtained PhD degrees under his supervision. One hundred and sixty papers have been published in National and International journals. Currently 20 scholars are pursuing PhD under his supervision and guidance.

Mr L.V. NandaKishore MSc, MPhil (Mathematics), Assistant Professor, Dept of Mathematics, Dr M.G.R University, Maduravoil, Chennai-95, currently doing his PhD in Madras University. He has many publications on Bayesian estimation, asset pricing and cliquet options. His research interest includes stochastic processes, asset pricing, fluid dynamics, Bayesian estimation and statistical models**.**