AN EMPIRICAL COMPARISON OF SUPERVISED LEARNING ALGORITHMS IN DISEASE DETECTION

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] , sasirekaraj@yahoo.co.in[2]

[3]Department of Mathematics, Dr M.G.R University, Chennai-95, India

lellapalliarunakishore@gmail.com

ABSTRACT

In this paper empirical comparison is carried out with various supervised algorithms. We studied the performance criterion of the machine learning tools such as Naïve Bayes, Support vector machines, Radial basis neural networks, Decision trees J48 and simple CART in detecting diseases. We used both binary and multi class data sets namely WBC, WDBC, Pima Indians Diabetes database 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 disease detection.

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]. Machine learning refers to a system that has the capability to automatically learn knowledge from experience and other ways [2]. 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 [3]

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 WBC, WDBC, Pima Indians diabetes database and 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. Related work is given in Section 2, Section 3 gives a brief description about the data mining algorithms and section 4 gives the description about the datasets used for this experiment. Section 5 gives the results obtained and the concluding remarks are given in Section 6 to address further research issues.

2. RELATED WORK

Large number of data mining algorithms has been developed in recent years for extraction of knowledge in databases. Of these many are supervised learning algorithms. These algorithms are mostly used for classification tasks. In a comparison of 10 learning algorithms over 11 datasets after calibration with Platt's method or isotonic regression SVM perform comparably to neural nets and nearly as well as boosted trees [4]. Gorman et al [5] reported that back

DOI : 10.5121/ijitcs.2011.1408

propagation outperformed nearest neighbour for classifying sonar targets. Shadmehr et al [6] showed that the performance of Bayes algorithm is better. Kirkwood et al [7] developed a symbolic algorithm ID3 which performed better than discriminant analysis for classifying the gait cycle of artificial limbs. Spikvoska et al [8] found that a HONN (higher order neural network) performed better than ID3. Atlas et al [9] showed that back propagation performed better than Cart. Mitchell et al [10] compared many algorithms on the MONK's problem. Ripley [11] compared neural networks and decision trees on the Tsetse fly data. King et al[12] Statlog is the first comprehensive study that analyzed different data mining algorithms on large real world data sets. LeCun et al 13] compared several learning algorithms on a handwriting recognition problem. Cooper et al 14] evaluated supervised learning methods on real medical data set using accuracy. Bauer et al [15] did empirical analysis about different statistical methods such as bagging and boosting. Lim et al [16] compared decision trees and other methods using accuracy as the main criterion. Perlich et al [17] conducted comparison between decision trees and logistic regression. Provost et al [18] examined the issue of predicting probabilities of decision trees including smooth and bagged trees. Witten et al [19] presented the comparison of different tools and techniques of data mining. Present research work is dedicated to analyze five supervised learning methods over four disease datasets with accuracy, precision, recall and Matthews correlation coefficient as performance criterion.

3. DATA MINING ALGORITHMS

3.1. 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 [20], [21]. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

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

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'.

3.2. 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 [3] 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 [22].

3.2.1 CART

The classification and regression trees (CART) methodology proposed by [23] 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 [23] [24] [25]. 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.

3.2.2 J48

Decision tree J48 [26] implements Quinlan's C4.5 algorithm [27] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. 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

3.3 Radial Basis Neural Networks

Radial Basis Function (RBF networks) is the artificial neural network type for application of supervised learning problem [28]. 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 [29]. 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 [28].

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 [30]. A mathematical justification of this can be found in the paper by Cover [31].

3.4 Support Vector Machines

Support vector machines (SVM) are a class of learning algorithms which are based on the principle of structural risk minimization (SRM) [32] [33]. 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 [34],[35],[36]. 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.[37]

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 [38]. 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 [38]. For our experiments we used Gaussian RBF kernel. A Gaussian RBF kernel is formulated as

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

4. DATASETS DESCRIPTION

4.1 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

4.2 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)

4.3 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

4.4 Pima Indians Diabetes Database

The Pima Indian diabetes database, donated by Vincent Sigillito is available in UCI machine learning depository. A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases is a collection of medical diagnostic reports of 768 examples.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

Number of Instances: 768
Table 4 shows the class Distribution: (class value 1 is interpreted as "tested positive for diabetes" and 0 is interpreted as “tested negative for diabetes”)

Table. 4 Class distribution of Pima Indians Diabetes Database

CLASS	NUMBER OF INSTANCES
0	500
1	268

Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

The paper [39] dealing with this data base uses an adaptive learning routine that generates and executes digital analogs of perceptron-like devices, called ADAP. They used 576 training instances and obtained a classification of 76% on the remaining 192 instances.

5 RESULTS

Experiments were conducted in WEKA with 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier[40]. From the confusion matrix to analyze the performance criterion for the classifiers in disease detection accuracy, precision, recall and Mathews correlation coefficient (MCC) have been computed for all datasets. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. Recall is the percentage of positive labelled instances that were predicted as positive. MCC measures the correlation of the actual and predicted class. In general, MCC gives a more balanced measure for the performance than the typically used values sensitivity and specificity [41],[42].MCC is a special case of the linear correlation coefficient, and therefore also scales between +1 (perfect correlation) and -1 (anti correlation), with 0 indicating randomness.

Accuracy, precision, recall and MCC 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.

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

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

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \qquad (7)$$

Table 4, 5, 6 and 7 shows the accuracy(ACC), MCC percentage for WBC, Pima, WDBC datasets and Breast tissue 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 5. ACC, MCC for WBC dataset

Algorithm	**ACC (%)**	**MCC**
Naïve Bayes	96.50	0.92
RBF networks	96.66	0.92
Trees-J48	94.59	0.88
Trees-CART	94.27	0.87
SVM-RBF kernel	96.84	0.94

Table 6. ACC, MCC for Pima dataset

Algorithm	**ACC (%)**	**MCC**
Naïve Bayes	76.30	0.46
RBF networks	75.39	0.43
Trees-J48	73.82	0.41
Trees-CART	75.13	0.43
SVM-RBF kernel	96.74	0.92

Table 7. ACC, MCC for WDBC dataset

Algorithm	**ACC (%)**	**MCC**
Naïve Bayes	92.61	0.84
RBF networks	93.67	0.88
Trees-J48	92.97	0.85
Trees-CART	92.97	0.84
SVM-RBF kernel	98.06	0.95

Table 7.Precision, Recall for WBC dataset

Algorithm	Precision (%)	Recall (%)
Naïve Bayes	98.7	95.7
RBF networks	98.7	95.9
Trees-J48	95.7	95.7
Trees-CART	96.4	94.4
SVM-RBF kernel	98.7	97.2

Table 8.Precision, Recall for Pima dataset

Algorithm	Precision (%)	Recall (%)
Naïve Bayes	80.2	84.4
RBF networks	77.6	86.8
Trees-J48	79.0	81.4
Trees-CART	77.6	86.8
SVM-RBF kernel	96.5	98.6

Table 9 Precision, Recall for WDBC dataset

Algorithm	Precision (%)	Recall (%)
Naïve Bayes	0.90	0.89
RBF networks	0.93	0.90
Trees-J48	0.89	0.91
Trees-CART	0.91	0.89
SVM-RBF kernel	0.99	0.95

Table 10 ACC for Breast tissue dataset

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

Table 11 and 12 shows the percentage of Precision and Recall for Breast tissue dataset.

Table 11 Precision (%) for Breast tissue

Algorithm	*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

Table 12. Recall (%) for Breast tissue

Algorithm	**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 13 MCC for Breast tissue

Algorithm	***Car***	***Fad***	***Mas***	***Gla***	***Con***	***Adi***
Naïve Bayes	0.97	0.96	0.89	0.96	0.88	0.91
RBF Networks	0.89	0.88	0.83	0.92	0.91	0.94
Trees-J48	0.97	0.92	0.83	0.96	0.91	0.91
Trees-CART	1.0	0.96	0.93	0.92	0.91	0.97
SVM-RBF kernel	0.97	0.96	1.0	1.0	1.0	1.0

From the results we can see that the percentage of accuracy, precision, recall, MCC 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 five supervised learning classifiers such as Naïve Bayes, SVM RBF kernel, RBF neural networks, Decision trees J48 and Simple CART on four real world datasets. As the real world datasets may have irrelevant noisy features they require lot of pre-processing to achieve satisfactory classification accuracy. The aim of this study is find out the classifier which can perform well on the real world data sets. In this study all the classifiers are used to classify the datasets namely WBC, WDBC, Pima diabetes and Breast tissue obtained from UCI machine learning depository without any pre-processing techniques. The experiments were conducted in WEKA with 10 fold cross validation. The results are compared and found that SVM RBF Kernel is excellent in performance than other classifiers with respect to accuracy, sensitivity, specificity and precision for both binary and multiclass datasets. Although other classifiers perform well in classification the behaviour varies differently for each dataset. SVM RBF Kernel always outperforms other classifiers for all datasets. In this comparative study more concentration is given towards the accuracy of the classifiers as this is concerned with disease detection. In future work accuracy as well as complexity of the algorithms will be calculated. Also 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] 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.

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

[4] R Caruana, and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms" *Proceedings of the 23rd international conference on Machine learning*, 2006.

[5] R P. Gorman and T.J. Sejnowski, “Analysis of hidden units in a layered network trained to classify sonar targets”, *Neural networks,* 1 (Part 1) 75-89, 1988.

[6] R. Shadmehr and Z. D’Argenio, “A comparison of a neural network based estimator and two statistical estimators in a sparse and noisy environment”, In IJCNN-90 *Proceedings of the international joint conference on neural networks,* 289-292,Ann Arbor, Mi, IEEE Neural Networks Council, 1990.

[7] C. Kirkwood, B. Andrews and P. Mowforth, “Automatic detection of gait events: a case study using inductive learning techniques”, *Journal of biomedical engineering,* 11(23): 511-516, 1989.

[8] L. Spikovska and M.B. Reid, “An empirical comparison of id3 and honns for distortion invariant object recognition” In Tai-90:tools for artificial intelligence: *Proceedings of the 2nd international IEEE conference,* Los Alamitos, CA, IEEE Computer Society Press, 1990.

[9] L. Atlas, J. Connor and D. Park, “An performance comparison of trained multilayer perceptrons and trained classification trees”, In Systems, man and cybernetics: *Proceedings of the 1989 IEEE conference,* 915-920, Cambridge, Ma. Hyatt Regency, 1991.

[10] T. Mitchell, B. Buchanan, G. Dejon, T. Diettrich, P. Rosenbloom and A. Waibel, “Machine Learning*” Annual Review of Computer Science,* vol 4, 417-433, 1990.

[11] B. Ripley,“Statistical aspects of neural networks”, *Chaos and Networks- Statistical and Probabilistic Aspects,* Chapman and Hall, 1993.

[12] R. King, C. Feng and A. Sutherland, “Statlog: Comparison of classification algorithms on large real world problems”, *Applied Artificial Intelligence,* 9, 1995.

[13] Y. LeCun, L.D. Jackel, L. Bottou, A. Brunot, C. Cortes, J.S. Denker, H. Drucker, L. Guyon, U.A. Mutter, E. Sackinger, P. Simard and V. Vapnik, “Comparison of learning algorithms for handwritten digit recognition”, *International Conference on Artificial Neural Networks,* 53-60, Paris, 1995.

[14] G.F Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B.G Buchanan, R. Cruana, M.J. Fine, C. Glymour, G. Gordon, B.H. Hanusa, J.E. Janosky, C. Meek, T. Mitchell, T. Richardson and P. Spirtes, “An evaluation of machine learning methods for predicting pneumonia mortality”, *Artificial intelligence in Medicine,* 9, 1997.

[15] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting and variants”, *Machine Learning,* 36, 1999.

[16] T.S Lim, W.Y Loh and Y.S Shih, “A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms”, *Machine Learning,* 40, 203-228, 2000.

[17] C. Pertich, F. Provosi and J.S Simono, “Tree induction vs Logistic regression: a learning curve analysis”, *J. Mach. Learn. Res.,* 4, 211-255, 2003

[18] F. Provost, D. Jensen and T. Oates, “Effiecient progressive sampling”, Fifth ACM SIGKDD, *International conference on knowledge Discovery and Data Mining,* San Diego, USA, 1999.

[19] I.H. Witten and E. Frank, “*Data Mining: Practical Machine learning tools and techniques with java implementations”,* Morgan Kaufmann, 2000.

[20] 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*.

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

[22] 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.

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

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

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

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

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

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

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

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

[31] 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.

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

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

[34] 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.

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

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

[37] 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.

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

[39] Smith, J.,W., Everhart, J.,E., Dickson, W.,C., Knowler, W.,C. and Johannes, R.,S., Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in *Proceedings of the Symposium on Computer Applications and Medical Care*, IEEE Computer Society Press, 261-265, 1988.

[41] P. Baldi, S. Brunak, Y. Chauvin, et al. Assessing the accuracy of prediction algorithms for classification: and overview. *Bioinformatics*, 5(5):412–424, 2000.

[42] . W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

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**.**