ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Classification of Incomplete Multivariate Datasets using Memory Based Classifiers – A Proficiency Evaluation

C. Lakshmi Devasena[1]
[1]Associate Professor, Department of Computer Science and Engineering, Sphoorthy Engineering College, Hyderabad, India, devaradhe2007@gmail.com

Abstract : Classification is a steady practice for allocating a given piece of input into any of the known class. Classification is an important Machine Learning technique. Many classification problems exist in different application areas and need to be solved. This work evaluates the proficiency of different memory based classifiers for classification of Multivariate data set with Missing values. For the proficiency evaluation the data sets with missing values have been taken from UCI machine learning repository and evaluated using the open source machine learning tool. Different memory based classifiers has been compared and a practical guideline for selecting the renowned and more suited algorithm for a classification is presented. Apart from that some realistic criteria for relating and evaluating the best classifiers are discussed.

Key words : IB1 Classifier, IBk Classifier, K Star Classifier, LWL Classifier, Missing Values, Multivariate Dataset.

INTRODUCTION

In machine learning, classification refers to an algorithmic process for designating a given input data into one among the different classes given. An example would be the given software can be designated into classes namely "paid" or "open access". Algorithm which employs classification in its procedure is known as a classifier. The input data can be referred as an instance and the categories are known as classes. The distinctiveness of the instance can be described by a vector of features. These features can be ordinal, nominal, real-valued or integer-valued. Most of the data mining algorithms work only in terms of nominal data and require that real or integer-valued data be converted into groups. Classification is actually a supervised procedure that learns to sort out new instances based on the knowledge learnt from a previously classified training set of instances. The corresponding unsupervised practice is known as clustering. It necessitates grouping data into classes based on inherent similarity measure. In machine learning, classification systems induced from empirical data (examples) are first of all rated by their prognostic accuracy. In actuality, however, the interpretability or transparency of a classifier is often significant as well. This work evaluates the proficiency of memory-based classifiers to classify the Multivariate Datasets with missing values. The mutual dependence of attributes or variables causes distortion of the space. Many algorithms used for classification may not be applied on multivariate data with missing values. Motivated by the need of such requirement, in this work, the memory based classification techniques for their suitability to efficiently evaluate the multivariate datasets with missing values are compared for further utilization.

LITERATURE REVIEW

Many researchers have made the performance analysis of various classification techniques at different view-points. The effectiveness evaluation of Memory Based Classifiers for the classification of Multivariate Datasets without Missing Values is investigated in [1]. The effectiveness prediction of rule based classifiers for the classification of Iris Data set is described in [2] & [3]. Performance evaluation of different classifiers (Fuzzy C Means Classifier, Back Propagation Network, Adaptive Resonance Theory and Support Vector Machine) for the classification of Multivariate Coronary Artery Disease dataset is discussed in [4]. Proficiency analysis of different statistical classifiers like K-Nearest Neighbour Classifier, Probabilistic Neural Network and Naïve Bayesian Classifier for the sign language classification system is depicted in [5]. A detailed review of all different classifiers is explained in [6]. An algorithm named Progressive Temporal Class Rule Miner (PTCR-Miner) is proposed in [7] to achieve the classification of multivariate temporal data. Performance comparison of Parametric and Non-parametric Classifiers (Naïve Bayes, Multi layer perceptron, Logistic Regression and Bayesian Net) for the classification of Breast Feed Dataset is elaborated in [8].

DATASETS USED

Multivariate datasets with missing values selected for Proficiency evaluation of Memory-Based Classifiers are Labor Dataset, Breast Cancer Dataset and 1984 United States Congressional Voting Records Dataset from UCI Machine Learning Repository [11]. Labor dataset has sixteen attributes (Duration, Wage increase First Year, Wage increase second Year, Wage increase Third Year, cost of Living Adjustment, Working Hours, Pension, Stand by Pay, Shift differential, Education allowance, Statutory Holidays, Vacation, Long term disability assistance, Contribution to Dental Plan, Bereavement assistant and Contribution to Health Plan) and consists of 57 instances of two different classes (Bad and Good) and having 2% to 84% of missing values in all the attributes. Breast Cancer Dataset has nine attributes (age, menopause, tumor-size, inv-nodes,

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

node-caps, deg-malig, breast, breast-quad and irradiat) and contains missing values in two attributes namely, node-caps and breast-quad. This dataset consists of 286 instances of two classes namely No-recurrence Events and Recurrence Events. Congressional Voting Records Dataset has 16 Boolean attributes (Handicapped infants, Adoption of the budget resolution, Water Project cost sharing, Physician fee freeze, religious groups in schools, el-salvador-aid, anti-satellite-test-ban, mx-missile, aid to Nicaraguan contras, immigration, education spending, synfuels corporation-cutback, superfund right-to-sue, duty-free-exports, crime and export administration act-south-Africa) and consists of 435 instances of two classes namely Democrat and Republican. This congressional vote dataset has missing values in all attributes from 2% to 24%.

PROCEDURE FOR PAPER SUBMISSION

Different memory based Classifiers are evaluated to find the effectiveness of those classifiers in the classification of Multivariate Data sets with Missing Values.

IB1 Classifier

IB1 is Instance based nearest neighbour classifier. It uses normalized Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as this training instance. If several instances have the smallest distance to the test instance, the first one obtained is used. Nearest neighbour method is one of the effortless and uncomplicated learning/classification algorithms, and has been effectively applied to a broad range of problem areas [5].

To classify an unclassified vector X, this algorithm ranks the neighbours of X amongst a given set of N data (Xi, ci), i = 1, 2... N, and employs the class labels cj (j = 1, 2... K) of the K most similar neighbours to predict the class of the new vector X. In specific, the classes of the K neighbours are weighted using the similarity between X and its each of the neighbours, where the Euclidean distance metric is used to measure the similarity. Then, X is assigned the class label with the greatest number of votes among the K nearest class labels. The nearest neighbour classifier works based on the intuition that the classification of an instance is likely to be most similar to the classification of other instances that are nearby to it within the vector space. Compared to other classification methods such as Naive Bayes', nearest neighbour classifier does not rely on prior probabilities, and it is computationally efficient if the data set concerned is not very large.

IBk Classifier

IBK is an implementation of the k-nearest-neighbours classifier. Each case is considered as a point in multi-dimensional space and classification is done based on the nearest neighbours. The value of 'k' for nearest neighbours can vary. This determines how many cases are to be considered as neighbours to decide how to classify an unknown instance.

For example, for the 'Breast Cancer' dataset, IBk would consider the nine dimensional spaces for the nine input variables. An unclassified instance would be classified as belonging to the class of its closest neighbour using Euclidean distance metric. If 6 is used as the value of 'k', then 6 closest neighbors are considered. The class of the new instance is considered to be the class of the majority of the instances. If 6 is used as the value of k and 5 of the closest neighbors are of type 'Recurrence Events', then the class of the test instance would be assigned as 'Recurrence Events'. The time taken to classify a test instance with nearest-neighbour classifier increases linearly with the number of training instances kept in the classifier. It has a large storage requirement. Its performance degrades quickly with increasing noise levels. It also performs badly when different attributes affect the outcome to different extents. One parameter that can affect the performance of the IBK algorithm is the number of nearest neighbours to be used. By default it uses just one nearest neighbour.

K Star Classifier

A K Star is a memory-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. The use of entropy as a distance measure has several benefits. Amongst other things it provides a consistent approach to handling of symbolic attributes, real valued attributes and missing values. K* is an instance-based learner which uses measures like entropy [9].

Specification of K*

Let I be a (possibly infinite) set of instances and T a finite set of transformations on I. Each $t \in T$ maps instances to instances: $t: I \rightarrow I$. T contains a distinguished member σ (the stop symbol) which for completeness maps instances to themselves ($\sigma(a) = a$). Let P be the set of all prefix codes from T* which are terminated by σ. Members of T* (and so of P) uniquely define a transformation on I: $t(a) = tn(tn\text{-}1(\ldots t1(a)\ldots))$ where $t = t1\ldots tn$

A probability function p is defined on T*. It satisfies the following properties:

$$0 \leq \frac{p(\bar{t}u)}{p(\bar{t})} \leq 1$$
$$\sum_{u} p(\bar{t}u) = p(\bar{t})$$
$$p(\Lambda) = 1 \qquad (1)$$

As a consequence it satisfies the following:

$$\sum_{\bar{t} \in P} p(\bar{t}) = 1 \qquad (2)$$

The probability function P* is defined as the probability of all paths from instance 'a' to instance 'b':

$$P^*(b|a) = \sum_{\bar{t} \in P : t(a) = b} p(\bar{t}) \qquad (3)$$

It is easily proven that P* satisfies the following properties:

$$\sum_{b} P^*(b|a) = 1$$
$$0 \leq P^*(b|a) \leq 1 \qquad (4)$$

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

The K* function is then defined as:

$$K^*(b|a) = -\log_2 P^*(b|a) \quad (5)$$

K* is not strictly a distance function. For example, K*(a|a) is in general non-zero and the function (as emphasized by the | notation) is not symmetric. Although possibly counter-intuitive the lack of these properties does not interfere with the development of the K* algorithm below. The following properties are provable:

$$K^*(b|a) \geq 0$$
$$K^*(c|b) + K^*(b|a) \geq K^*(c|a) \quad (6).$$

LWL Classifier

LWL is a learning model that belongs to the category of memory based classifiers. Machine Learning Tools work by default with LWL model and Decision Stump in combination as classifier. Decision Stump usually is used in conjunction with a boosting algorithm.

Boosting is one of the most important recent developments in classification methodology. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data, and then taking a weighted majority vote of the sequence of classifiers thus produced. This simple strategy results in dramatic improvements in performance for simple classification problem. This inexplicable phenomenon can be understood in terms of well known statistical principles, namely additive modeling and maximum likelihood. Boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion, for two-class problem. The best estimate for the outputs is found using a local model that is a hiper-plane. Distance weighting the data training points corresponds to requiring the local model to fit nearby points well, with less concern for distant points:

$$C = \sum_i (x_i^T \beta - y_i)^2 \quad (7)$$

This process has a physical interpretation. The strength of the springs are equal in the unweighted case, and the position of the hiper-plane minimizes the sum of the stored energy in the springs (Equation 8). A factor of 1/2 in all our energy calculations is ignored to simplify notation. The stored energy in the springs in this case is C of Equation 7, which is minimized by the physical process.

$$w = \int dFx = \int Kdx.x = K \int xdx = \frac{Kx^2}{2} \quad (8)$$

The linear model in the parameters can be expressed as:

$$x_i^T \beta = y_i \quad (9)$$

In what follows we will assume that the constant 1 has been appended to all the input vectors xi to include a constant term in the regression. The data training points can be collected in a matrix equation: $X \beta = y$ (10)

where X is a matrix whose i^{th} row is x_i^T and y is a vector whose i^{th} element is y_i. Thus, the dimensionality of X is 'n x d' where n is the number of data training points and d is the dimensionality of x. Criterion given in equation 11 [10] is minimized by estimating the parameters using an unweighted regression. By solving the normal equations

$$(X^T X)\, \beta = X^T y \quad (11)$$

For β: $\beta = (X^T X) - X^T y$ (12)

From the point of view of efficiency or accuracy, inverting the matrix $X^T X$ is not the numerically best way to solve the normal equations, and usually other matrix techniques are used to solve Equation 11.

CRITERIA USED FOR CLASSIFICATION EVALUATION

The comparison between classifiers is made on the basis of the following measures.

Accuracy Classification

All classification result could have error rate and it may fail to classify correctly some instances. Accuracy can be calculated using the equation below.

Accuracy = (Instances Correctly Classified / Total Number of Instances Used)*100 % (13)

Mean Absolute Error

MAE is the average of difference between predicted and actual value in all test cases. The formula for calculating MAE is shown below in equation.

MAE = (|a1 – c1| + |a2 – c2| + … +|an – cn|) / n (14)

Here 'a' is the actual output and 'c' is the expected output.

Root Mean Square Error

RMSE is used to measure differences between values predicted by a model and the actually observed values. Taking the square root of the mean square error as shown in equation given below gives the RMSE.

$$\sqrt{\frac{(a_1-c_1)^2+(a_2-c_2)^2+\ldots+(a_n-c_n)^2}{n}} \quad (15)$$

Here is the actual output arrived is denoted by 'a' and the expected output is denoted by 'e'. The RMSE is the usual measure for numeric prediction.

Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system.

The classification accuracy, mean absolute error, time taken to test the model, root mean squared error and confusion matrices are calculated for each machine learning algorithm using the machine learning tool.

RESULTS AND DISCUSSION

This work has been performed using Open Access Machine learning tool to evaluate the effectiveness of all the memory- based classifiers for different multivariate datasets with Missing Values.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Data Set 1: Labor Data set

The performance of the memory based algorithms for Labor Data set containing missing values is shown in Table 1 in terms of Classification Accuracy, Time taken to test the Model, RMSE and MAE values. Comparison among different memory based classifiers based on the correctly classified instances is shown in Fig 1. Comparison among the different memory based classifiers based on MAE and RMSE values and the comparison graph is shown in Fig 2. Table 2 to Table 5 shows the confusion matrix arrived for these classifiers. The ranking has been done based on the classification accuracy, Time taken to test the Model, MAE and RMSE values. Based on the results arrived, IB1Classifier which has 100% accuracy and zero MAE and RMSE is best among all the classifiers followed by IBk Classifier, K Star and LWL Classifier.

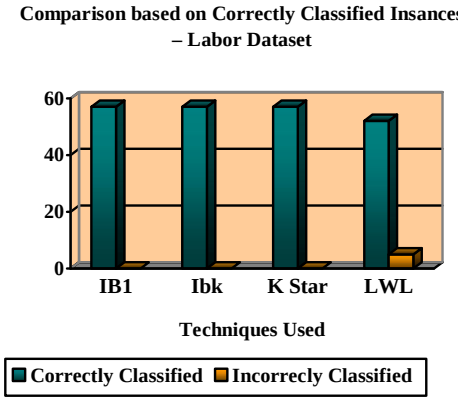

Fig 1: Comparison based on Correctly Classified Instances – Labor Dataset.

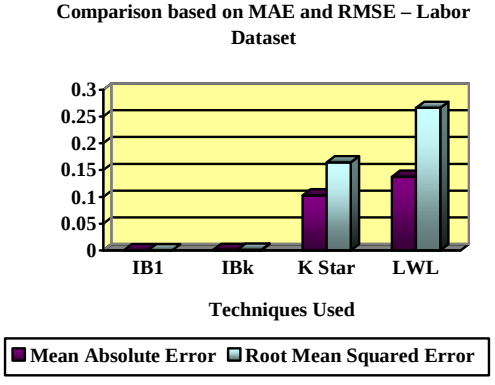

Fig 2: Comparison based on MAE and RMSE – Labor Dataset

Table 1: Overall performance of Memory Based Classifiers – Labor dataset

Classifier Used	Correctly Classified instances (0ut of 57)	Classification Accuracy	Time taken to test Model (in sec)	MAE	RMSE
IB1	57	100	0.06	0	0
IBk	57	100	0.06	0.169	0.169
K Star	57	100	0.11	0	0
LWL	52	91.23	0.18	0.168	0.2629

Table 2: Confusion Matrix of IB1Classifiers – Labor dataset

	Bad	Good
Bad	20	0
Good	0	37

Table 3: Confusion Matrix of IB2Classifiers – Labor dataset

	Bad	Good
Bad	20	0
Good	0	37

Table 4: Confusion Matrix of K Star Classifiers – Labor dataset

	Bad	Good
Bad	20	0
Good	0	37

Table 5: Confusion Matrix of LWL Classifiers – Labor dataset

	Bad	Good
Bad	15	5
Good	0	37

Data Set 2: Congressional Vote Data set

The performance of the memory based algorithms for Congressional Vote Data set containing missing values is shown in Table 6 in terms of Classification Accuracy, Time taken to test the Model, RMSE and MAE values. Comparison among different memory based classifiers based on the correctly classified instances is shown in Fig 3. Comparison among the different memory based classifiers based on MAE and RMSE values is shown in Fig 4. Table 7 to Table 10 shows the confusion matrix arrived for these classifiers. The entire ranking is done based on the classification accuracy, Time taken to test the Model, MAE and RMSE values. Based on the results arrived, IB1Classifier is best among all the classifiers followed by IBk Classifier, then by K Star Classifier and then LWL Classifier.

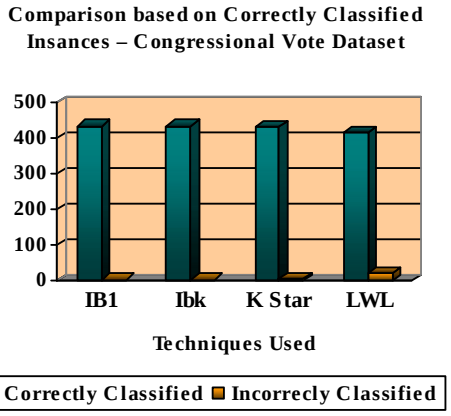

Fig 3: Comparison based on Correctly Classified Instances – Congressional Vote Dataset.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

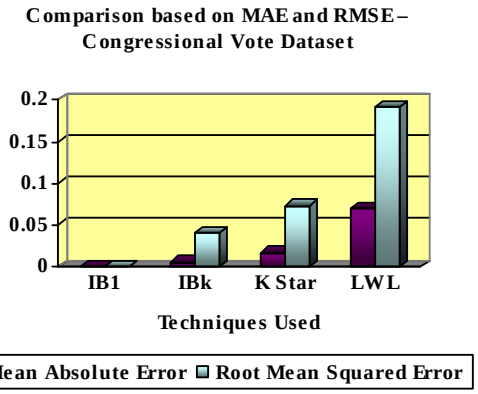

Fig 4: Comparison based on MAE and RMSE – Congressional vote Dataset

Table 6: Overall performance of Memory Based Classifiers – Labor dataset

Classifier Used	Correctly Classified instances (0ut of 57)	Classification Accuracy	Time taken to test Model (in sec)	MAE	RMSE
IB1	434	99.77	0.28	0	0
IBk	434	99.77	0.28	0.0049	0.0404
K Star	431	99.08	0.97	0.0167	0.0728
LWL	416	95.63	0.80	0.0691	0.1917

Table 7: Confusion Matrix of IB1Classifiers – Congressional Vote dataset

	Democrat	Republican
Democrat	267	0
Republican	1	167

Table 8: Confusion Matrix of IBk Classifiers – Congressional Vote dataset

	Democrat	Republican
Democrat	267	0
Republican	1	167

Table 9: Confusion Matrix of K Star Classifiers – Congressional Vote dataset

	Democrat	Republican
Democrat	264	3
Republican	1	167

Table 10: Confusion Matrix of LWL Classifiers – Congressional Vote dataset

	Democrat	Republican
Democrat	253	14
Republican	5	163

Data Set 3: Breast Cancer Data set

The performance of the memory based algorithms for Breast Cancer Data set containing missing values is shown in Table 11 in terms of Classification Accuracy, Time taken to test the Model, RMSE and MAE values. Comparison among the memory based classifiers based on the correctly classified instances is depicted in Fig 5. Comparison among different memory based classifiers based on MAE and RMSE values is given in Fig 6. The confusion matrix arrived for these classifiers are given in tables from Table 12 to Table 15. The ranking of classifiers is done based on the classification accuracy, Time taken to test the Model, MAE and RMSE values. Based on the results arrived, IB1Classifier got first position in ranking followed by IBk Classifier, K Star Classifier and LWL as shown in Table 11.

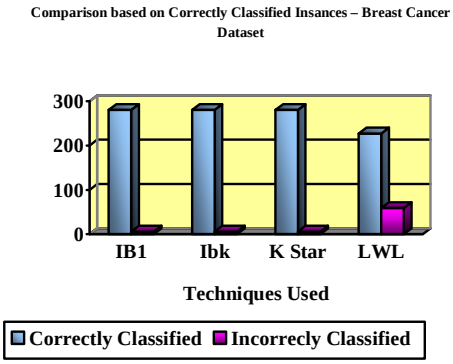

Fig 5: Comparison based on Correctly Classified Instances – Breast Cancer Dataset.

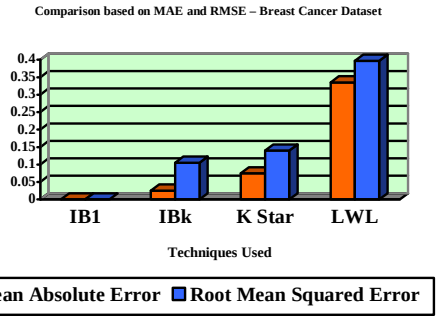

Fig 6: Comparison based on Correctly Classified Instances – Breast Cancer Dataset

Table 11: Overall performance of Memory Based Classifiers – Breast Cancer dataset

Classifier Used	Correctly Classified instances (0ut of 286)	Classification Accuracy	Time taken to test Model (in sec)	MAE	RMSE
IB1	280	97.90	0.10	0	0
IBk	280	97.90	0.10	0.0253	0.1053
K Star	280	97.90	0.35	0.0747	0.1399
LWL	227	79.37	0.60	0.3353	0.3971

Table 12: Confusion Matrix of IB1Classifiers – Breast Cancer dataset

	No-recurrence Events	Recurrence Events
No-Recurrence Events	200	1
Recurrence Events	5	80

Table 13: Confusion Matrix of IBk Classifiers – Breast Cancer dataset

	No-recurrence Events	Recurrence Events
No-Recurrence Events	200	1
Recurrence Events	5	80

Table 14: Confusion Matrix of K Star Classifiers – Breast Cancer dataset

	No-recurrence Events	Recurrence Events
No-Recurrence Events	200	1
Recurrence Events	5	80

Table 15: Confusion Matrix of IB1Classifiers – Breast Cancer dataset

	No-recurrence Events	Recurrence Events
No-Recurrence Events	200	1
Recurrence Events	5	80

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 148-153 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

CONCLUSION

This work investigated the proficiency of different memory based classifiers for classification of Multivariate data set with Missing values. For the proficiency evaluation, the data sets with missing values have been taken from UCI machine learning repository and experimented using the open source machine learning tool. Comparison of different memory based classifiers has been made and a renowned metrics are used to select the more suited algorithm for a classification of multivariate datasets. Apart from that some realistic criteria for relating and evaluating the best classifiers are discussed. After experimented different Memory Based Classifiers (IB1, IBk, K Star and LWL Classifiers) for the classification of different Multivariate Datasets with Missing values (Labor Dataset, Breast Cancer Dataset and Congressional Vote Dataset), it is concluded that IB1 Classifier performs best followed by IBk Classifier, K star Classifier and then by LWL Classifier.

ACKNOWLEDGEMENT

The author thanks the Management of Sphoorthy Engineering College and Faculties of CSE Department for the cooperation extended.

REFERENCES

[1] C. Lakshmi Devasena, "Effectiveness Prediction of Memory Based Classifiers for the Classification of Multivariate Data Set," *CS & IT-CSCP 2012*, Volume 2, pp. 413–424. DOI: 10.5121/csit.2012.2440.

[2] C. Lakshmi Devasena, Sumathi.T, Gomathi.V.V and Hemalatha.M, "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set," *Bonfring Int. J. Man Machine Interface*, 1(1): 5 – 9, 2011

[3] C. Lakshmi Devasena, T. Sumathi, V.V. Gomathi, R. Malarkodi and M. Hemalatha, 2011. "Predicting Effectiveness of Rule based Classifiers for a Classification Problem," *Proc. of Int. Conf. on Networks, Intelligence and Computing Technologies*, 1(2): 559 – 563. (ISBN: 978-81-8424-742-8).

[4] G. Nalinipriya, A. Kannan and P. Anandhakumar, "Performance Analysis of Classifiers for Multivariate Coronary Artery Disease Dataset using Renowned Metrics," *European Journal of Scientific Research*, Vol. 86, No 4, September, 2012, pp.565 – 572.

[5] M. Krishnaveni and V.Radha, "Performance evaluation of Statistical classifiers using Indian Sign language datasets," *International Journal of Computer Science, Engineering and Applications*, Vol.1, No.5, October 2011, pp. 167 – 175.

[6] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica 31*, pp, 249-268, 2007.

[7] Chao-Hui Lee and Vincent S. Tseng, "PTCR-MINER: An Effective Rule-Based Classifier on Multivariate Temporal Data Classification," *International Journal of Innovative Computing, Information and Control*, Vol. 7, No. 10, October 2011, pp. 5925-5938

[8] Yugal kumar and G. Sahoo, "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA," *Int. Jour. Information Technology and Computer Science*, 2012, 7, pp. 43-49.

[9] John G. Cleary, K*: An Instance-based Learner Using an Entropic Distance Measure.

[10] Christopher G. Atkeson, Andrew W. Moore and Stefan Schaal, Locally Weighted Learning, October 1996.

[11] UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.