International Journal of Computing and Business Research (IJCBR)

ISSN (Online) : 2229-6166

Volume 6 Issue 2 March 2015

KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS

Dr. S. Vijayarani[1], Mr.S.Dhayanand[2]

Assistant Professor[1], M.Phil Research Scholar[2]

Department of Computer Science, School of Computer Science and Engineering,

Bharathiar University, Coimbatore, Tamilnadu, India[1, 2]

Abstract The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective analysis, diagnosis and decision making. Discovery of hidden patterns and relationships often goes idle. Advanced data mining techniques can help and provides remedy to handle this situation. Data mining is the process of extraction of hidden information from the voluminous dataset. Various data mining techniques are clustering, classification, association analysis, regression, summarization, time series analysis and sequence analysis, etc. Data mining techniques plays a vital role in different domains such as text mining, graph mining, medical mining, multimedia mining and web mining. The objective of this research work is to predict kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN). The aim of this work is to compare the performance of these two algorithms on the basis of its accuracy and execution time. From the experimental results it is observed that the performance of the ANN is better than the other algorithm.

Keywords: Data Mining; Data mining techniques; Kidney disease; Support Vector Machine; Artificial Neural Network.

1. Introduction

Data mining is the process of extracting hidden information from the large dataset. Data mining is used in several domains such as image mining, text mining, sequential mining, web mining, graph mining, spatial mining and so on. Data mining techniques are used in various applications like fault diagnosis, anomaly detection, medical diagnosis, e-mail filtering, face recognition and oil spill detection [11]. Data mining techniques such as classification, clustering and association rule and etc. plays a great role in extracting unknown knowledge from the databases.

Classification is a data mining technique used to predict group membership for data instances. Classification is comparable to clustering in that it also segment information retrieval into distinct segment called classes. In order to predict the outcome [12], the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute [6].

This work predominantly focused on, prediction of four types of kidney diseases (Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure and Chronic Glomerulonephritis. Kidney diseases are predicted using data mining algorithms such as Support Vector Machine(SVM) and Artificial Neural Network (ANN).

The remaining portion of the paper is organized as follows. Related works are discussed in Section 2. The proposed methodology is given in Section 3. Section 4 analyzes the experimental results. Section 5 gives the conclusion.

2. Literature Review

Abhishek et.al [1] have used two neural network techniques, Back Propagation Algorithm (BPA), Radial Basis Function (RBF) and one non-linear classifier Support Vector Machine (SVM) and compared in accordance with their efficiency and accuracy. They used WEKA 3.6.5 tool for implementation to find the best technique among the above three algorithms for Kidney Stone Diagnosis. The main purpose of their thesis work was to propose the best tool for medical diagnosis, like kidney stone identification, to reduce the diagnosis time and improve the efficiency and accuracy. From the experimental results they concluded, the back propagation (BPA) significantly improved the conventional classification technique for use in medical field.

Andrew Kusiak et al [2] have used data preprocessing, data transformations, and data mining approach to elicit knowledge about the interaction between many of measured parameters and patient survival. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining were interpreted for their medical significance. They have introduced a new concept in their research work, it have been applied and tested using collected data at four dialysis sites. The approach presented in their paper reduced the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most significant parameters discovered.

Ashfaq Ahmed K et.al, [3] have presented a work using machine learning techniques, namely Support Vector Machine [SVM] and Random Forest [RF]. These were used to study, classify and compare cancer, liver and heart disease data sets with varying kernels and kernel parameters. Results of Random Forest and Support Vector Machines were compared for different data sets such as breast cancer disease dataset, liver disease dataset and heart disease dataset. The results with different kernels were tuned with proper parameter selection. Results were better analyzed to establish better learning techniques for predictions. It is concluded that varying results were observed with SVM classification technique with different kernel functions.

Sadik Kara et.al [13] had concentrated on the diagnosis of optic nerve disease through the analysis of pattern electroretinography (PERG) signals with the help of artificial neural network (ANN). Implemented Multilayer feed forward ANN trained with a Levenberg Marquart (LM) backpropagation algorithm. The end results were classified as healthy and diseased. The stated results shown that the proposed method PERG could make an effective interpretation.

Koushal Kumar et.al [10] used Learning vector quantization (LVQ), two layers feed forward perceptron trained with back propagation training algorithm and Radial basis function (RBF) networks for diagnosis of kidney stone disease. They have compared the performance of all three neural networks on the basis of its accuracy, time taken to build model, and training data set size. They used Waikato Environment for Knowledge Analysis (WEKA) tool for execution. Finally from the experimental results, authors concluded that multilayer perceptron trained with back propagation is best algorithm for kidney stone diagnosis.

3. Methodology

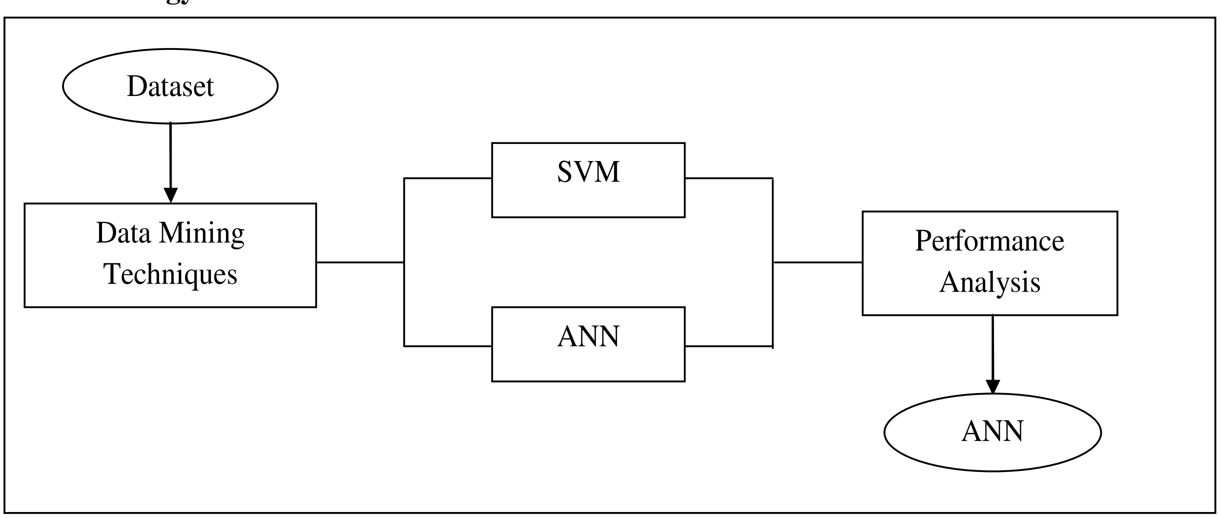

Figure 1. System Architecture

Dataset

The dataset is collected from several medical labs, centres and hospitals. From this the synthetic kidney function test (KFT) dataset have been created for analysis of kidney disease. This dataset contains five hundred and eighty four instances and six attributes are used in this comparative analysis. The attributes in this KFT dataset are Age, Gender, Urea, Creatinine and Glomerular Filtration Rate (GFR). This dataset consists of renal affected disease information.

Data mining techniques

There are many different methods used to perform data mining tasks. Data mining techniques not only require specific types of data structures, but also imply certain types of algorithmic approaches. Some of the common data mining techniques are classification, clustering, prediction, sequential pattern etc. There are two types of techniques are parametric and nonparametric techniques. Parametric model describes the relationship between input and output through the use of algebraic equation where some parameters are not specified. Nonparametric techniques are more appropriate for data mining applications. A nonparametric model is one that is data driven. No explicit equations are used to determine the model. Nonparametric techniques include neural networks, decision trees and genetic algorithms [5].

Support Vector Machine (SVM)

Support Vector Machines, a new method for the classification of both linear and nonlinear data. In a casing, a support vector machine (SVM) is an algorithm that works as follows. It uses a nonlinear mapping to renovate the unique training data into a higher dimension [15]. Surrounded by this new dimension, it examines for the linear optimal separating hyperplane i.e., a "decision boundary" sorting out the tuples of one class from another. With a suitable nonlinear mapping to a necessarily high dimension [14], data from two classes can always be separated by a hyperplane. The SVM finds the hyperplane using support vectors and margins [8]. Although the training time of even the fastest SVMs can be exceedingly slow, they are extremely accurate, outstanding to their ability to model complex nonlinear decision boundaries [16]. They are much less prone to over fitting than other methods. The Support vectors initiate also provide a compact description of the learned model. SVMs can be used for prediction along with classification. They have been applied to a several areas, including handwritten digit recognition, object recognition, and speaker identification and also benchmark time-series prediction tests.

The pseudo code for the perceptron learning, taken from [4], is given in Table 1

Table 1. Pseudo code for SVM

Require: A linear separable set S , learning rate $\eta \in \Re^+$
Step1 $w_0 = 0; b_0 = 0; k = 0;$
Step 2 $R = \max_{1 \le i \le l} \| x_i \|$
Step3 while at least one mistake is made in the for loop do
Step4 for $i = 1,...,l$ do
Step5 if $y_i(< w_k, x_i > +b_k) \le 0$ then
Step6 $w_{k+1} = w_k + \eta y_i x_i$
Step7 $b_k + \eta y_i \Re^2$ (updating bias[1])
Step8 $k = k + 1$
Step 9 end if
Step10 end for
Step11 end while
Step12 Return w_k, b_k, where k is the number of mistakes

Artificial Neural Network (ANN)

Backpropagation absorbs by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value. The target value may be the known class label of the training tuple (for classification problems) or a continuous value (for prediction). For each training tuple, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual target value. These modifications are made in the "backwards" direction [9], i.e., from the output layer, through each hidden layer down to the first hidden layer, hence the name backpropagation. Even though it is not guaranteed, in general the weights will finally converge, and the learning process stops. The steps involved are expressed in terms of inputs, outputs, and errors. Though, once become familiar with the process, each step is inherently simple. The algorithm is summarized in Table2.

Table 2: Pseudo code for Backpropagation

Input:

- *D*, a data set consisting of the training tuples and their associated target values;
- *l*, the learning rate;
- *network*, a multilayer feed-forward network

Output: A trained neural network.

Methods:

(1) Initialize all weights and biases in *network*;
(2) while terminating condition is not satisfied {
(3) for each training tuple ***X*** in *D* {
(4) // Propagate the inputs forward:
(5) for each input layer unit *j* {
(6) *Oj* = *Ij*; // output of an input unit is its actual input value
(7) for each hidden or output layer unit *j* {
(8) $I_j = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit *j* with respect to the previous layer, *i*
(9) $O_j = \frac{1}{1+e^{-1}j};\}$ // compute the output of each unit *j*
(10) // Backpropagate the errors:
(11) for each unit *j* in the output layer
(12) $Err_j = O_j(1-O_j)(T_j - O_j)$; // compute the error
(13) for each unit *j* in the hidden layers, from the last to the first hidden layer
(14) $Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, *k*
(15) for each weight *wi j* in *network* {
(16) $\Delta w_{ij} = (l) Err_J O_i$; // weight increment
(17) $w_{ij} = w_{ij} + \Delta w_{ij};\}$ // weight update
(18) for each bias θ_j in *network* {
(19) $\Delta\theta_j = (l) Err_j$; // bias increment
(20) $\theta_j = \theta_j + \Delta\theta_j;\}$ // bias update
(21) } }

IV Experimental Results

This work is performed in Matlab tool, developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, FORTRAN and Python. The experimental comparison of SVM and ANN are done based on the performance measures of classification accuracy and execution time.

Classification Accuracy

Table 3 represents the performance of classification accuracy measure of the datasets using classification algorithms such as SVM and ANN.

Table 3: Accuracy Measure

Algorithms	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	TP Rate
SVM	76.32	23.68	0.763
ANN	87.70	12.3	0.877

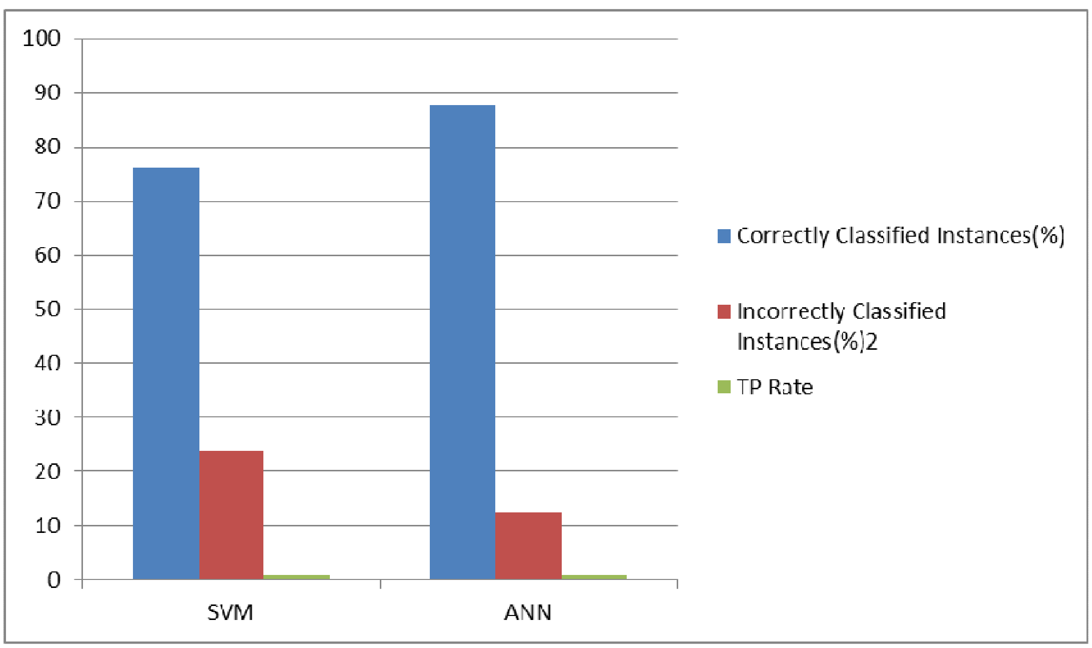

Figure 2: Accuracy Measure

Figure 2 represents the accuracy measure for the data mining algorithms namely SVM and ANN. From the experimental result, ANN performs best in classifying process than SVM algorithm.

Execution Time

Table 4 represents the execution time required for SVM and ANN algorithms

Table 4: Execution time Analysis

Algorithms	**Execution Time in Seconds**
SVM	3.22
ANN	7.26

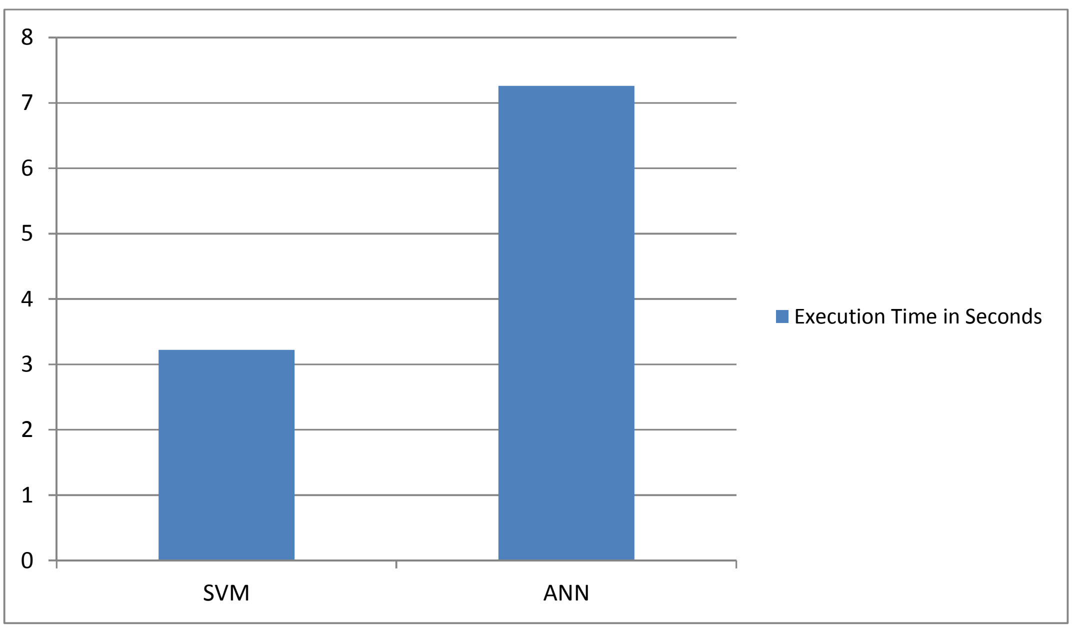

Figure 3: Execution Time of Classification Algorithms

Figure 3 represents the time taken for execution process. SVM performs with minimum period of execution time than the ANN algorithm.

Table 5 represents, the number of patients affected by the kidney diseases.

International Journal of Computing and Business Research (IJCBR)

ISSN (Online) : 2229-6166

Volume 6 Issue 2 March 2015

Table 5. Classification of Kidney Diseases

Data Mining Algorithms → / Kidney Disease	SVM	ANN
Normal	435	439
Acute Nephritic Syndrome	45	43
Chronic Kidney disease	42	42
Acute Renal Failure	19	16
Chronic Glomerulonephritis	42	43

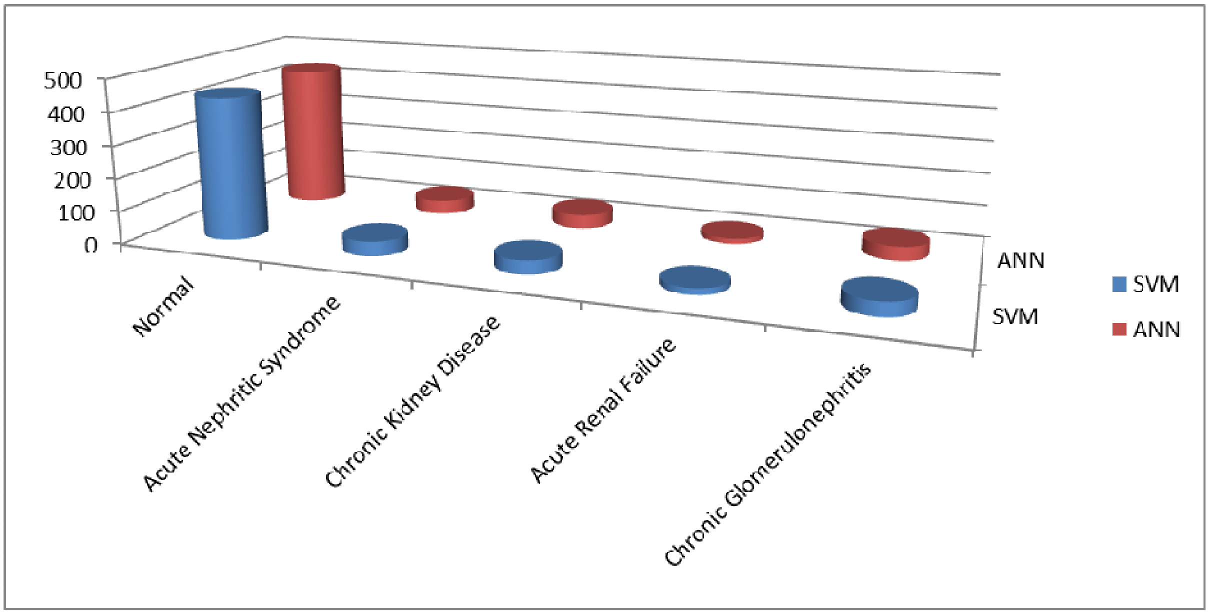

Figure 4: Classification of Kidney Diseases

Figure 4 represents the Kidney diseases classified by SVM and ANN algorithms. Based on chart analysis, ANN gives the overall best classification result than other algorithm.

V. Result and Discussion

The algorithm which has the higher accuracy with the minimum execution time has chosen as the best algorithm. In this classification, each classifier shows different accuracy rate. ANN has the maximum classification accuracy, hence, it is considered as the best classification algorithm. For execution time factor, the SVM algorithm needs minimum execution time.

VI Conclusion

In this research work classification process is used to classify four types of kidney diseases. Comparisons of Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms are done based on the performance factors classification accuracy and execution time. From the results, it can be concluded that the ANN achieves increased classification performance, yields results that are accurate, hence it is considered as best classifier when compared with SVM classifier algorithm. Feasibly, SVM classifier classifies the data with minimum execution time. In future, ANN algorithm is enhanced to minimize the execution time.

Reference

1. Abhishek, Gour Sundar Mitra Thakur, Dolly Gupta, (2012) "Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis", International Journal of Computer Science and Information Technologies, Vol. 3 (3) , page no 3900-3904
2. AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) "Predicting survival time for kidney dialysis patients: a data mining approach", Elsevier Publication, Computers in Biology and Medicine 35, page no 311–327
3. Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain,(2013) "*Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques"*, International Journal of Computer Applications Volume 69– No.11, May page no 12-16
4. Cristianini. N and Shawe-Taylor. J, (2000) "*An Introduction to support vector machines and other kernel based learning methods"*. Cambridge University Press
5. Dipak V. Patil, Bichkar. R.S, (2011) "An Optimistic Data Mining Approach for Handling Large Data Set using Data Partitioning Techniques", International Journal of Computer Applications (0975 – 8887), Volume 24– No.3, June, page no 29- 33
6. Ishani Mandli, Mahesh Panchal, (2014) "Selection of Most Relevant Features from High Dimensional Data using IG-GA Hybrid Approach", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, February, pg. 827-830
7. Jainesh Patel, Neha R Soni, (2014) "Survey of Supervised and Unsupervised Algorithms in Email Management", International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March, page no 1241-1243
8. Jiawei han and Micheline Kamber, (2006) "*Data mining Concepts and Techniques,* Second edition", Morgan Kaufmann publishers, by Elsevier Inc. All rights reserved.
9. Khyati Chaudhary, Bhawna Mallick, "Exploration of Data mining techniques in Fraud Detection: Credit Card", International Journal of Electronics and Computer Science Engineering, Volume1,Number 3, page no 1765-1771
10. Koushal Kumar, Abhishek, (2012) "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal of Information Technology and Computer Science, 7 July, page no 20-25

11. Mircea GH, Robert J. Howlett, Lakhmi G. Jain, (2004) "*Knowledge-Based Intelligent Information and Engineering Systems"*, 8th International Conference, KES 2004, Wellington, New Zealand, Sep 20-25. Proceedings, e-book Springer publication
12. Ramya. S, Senthil Kumar. R, (2014) "*Survey of Accident Severity Estimation Using Data Mining Techniques"*, International Journal of Computer Science & Engineering Technology, Vol. 5 No. 10 Oct page no 1041- 1044
13. Sadik Kara, Aysegul Guvenb, Ayse OztUrk Onerc, (2006) "*Utilization of artificial neural networks in the diagnosis of optic nerve diseases",* Elsevier Publication, Computers in Biology and Medicine 36 page no 428–437
14. Shaohua Teng, Hongle Du, Naiqi Wu, Wei Zhang, Jiangyi Su, (2010) A Cooperative Network Intrusion Detection Based on Fuzzy SVMs, JOURNAL OF NETWORKS, VOL. 5, NO. 4, APRIL page no 475-483
15. Sonali Sonavane, Alka Khade, Gaikwad. V.B, (2013) "Novel Approach for Localization of Indian Car Number Plate Recognition System using Support Vector Machine", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August, page no 179-183
16. Sunitha Devi1. P, Sowjanya. CH, Sunitha. K.V.N, (2014) "A Review of Supervised Learning Based Classification for Text to Speech System", International Journal of Application or Innovation in Engineering & Management, Volume 3, Issue 6, June page no 79-86