Defence Science Journal, Vol. 60, No. 4, July 2010, pp. 399-404

Statistical Measures to Determine Optimal Structure of Decision Tree: One versus One Support Vector Machine

Manju Bala and R.K. Agrawal*

Jawaharlal Nehru University, New Delhi-110 067
E-mail: manjugpm@gmail.com, *rka@mail.jnu.ac.in*

ABSTRACT

In this paper, one versus one optimal decision tree support vector machine (OvO-ODT SVM) framework is proposed to solve multi-class problems where the optimal structure of decision tree is determined using statistical measures, i.e., information gain, gini index, and chi-square. The performance of proposed OvO-ODT SVM is evaluated in terms of classification accuracy and computation time. It is also shown that proposed OvO-ODT SVM using all the three measures is more efficient in terms of time complexity for both training and testing phases in comparison to conventional OvO and support vector machine binary decision tree (SVM-BDT). Experiments on University of California, Irvine (UCI) repository dataset illustrates that ten cross-validation accuracy of our proposed framework is comparable or better in comparison to conventional OvO and SVM-BDT for most of the datasets. However, the proposed framework outperforms the conventional OvO and SVM-BDT for all the datasets in terms of both training and testing time.

Keywords: Support vector machine, decision tree, information gain, gini index, chi-square, SVM-BDT, binary decision tree

1. INTRODUCTION

Support vector machine (SVM) is most commonly used classifier for its better performance in comparison to other classifiers C4.5, artificial neural networks, bayesian classification in data mining and pattern recognition community. SVM based on statistical learning theory developed by Vapnik[1] was originally designed for binary classification. Most of the real-world applications, i.e., text categorisation[2], optical character recognition[3], speech recognition[4] are essentially multi-class classification problems. How to effectively extend SVM for multi-class classification is still an ongoing research issue[5]. The most common way to build a multi-class SVM is by constructing and combining several binary classifiers[6]. To solve multi-class classification problems, the whole problem is divided into a number of binary classification problems. The two representative ensemble schemes are one against all (OAA) and one versus one[5] (OvO). In literature, it has been stated that conventional OvO SVM approach has the problem of unclassifiable region. To resolve unclassifiable region for OvO SVM, Platt[7], *et al*. proposed decision directed acyclic graph (DDAG) scheme. Madzarov[8], *et al*. proposed support vector machine binary decision tree (SVM-BDT) architecture that uses binary SVMs for making decisions at each node of decision tree which takes advantage of both the efficient computation of the tree architecture and high accuracy of SVMs. The hierarchy of binary decision subtasks using SVMs is designed with clustering algorithms which employs Euclidean distance as similarity measure for creating two disjoint groups of patterns. Euclidean similarity measure does not take into account within class variability of patterns. Hence, it may not be suitable for measuring class separability between two different classes of patterns. In literature, other than Euclidean distance, many similarity measures were used to determine the class separability, each of them associates with some advantages and disadvantages. Among them, few more realistic and effective statistical measures used in literature are information gain, gini index and chi-square for measuring class separability.

In this paper, construction of OvO-ODT SVM has been proposed, where class separability is determined using statistical measures. The performance of the proposed method is evaluated in terms of classification accuracy and computation time required for training and testing. It is also shown theoretically that OvO-ODT SVM is more efficient in terms of computation time for both training and testing phases in comparison to conventional OvO and SVM-BDT.

2. DECISION TREE SUPPORT VECTOR MACHINE

The SVM is based on the idea of structural risk minimisation which minimises the generalisation error[1]. It classifies data by determining a set of support vectors which are members of the set of training inputs that outline a hyperplane in feature space[9]. Consider n training samples $\{(\vec{x_i}, y_i) \mid 1 \le i \le n\}$, where $\vec{x_i}$ represents i^{th} input feature vector and y_i is the corresponding target class. Given a set of n training instances, each represented as $(\vec{x_i}, y_i)$,

Received 5 October 2009, Revised 15 March 2010

the basic problem for training an SVM can be reformulated as:

$$\text{Maximise } J = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i\, y_j\, \alpha_i\, \alpha_j\, (x_i^T\ x_j) \qquad (1)$$

subject to $\sum_{i=1}^{n} y_i\, \alpha_i = 0$ and $\alpha_i \geq 0$, for $1 \leq i \leq n$

The computation of dot products between vectors without explicitly mapping to another space is performed by a kernel function $K(x_i, x_j)$. Use of a kernel function[10] enables the curse of dimensionality to be addressed and the solution implicitly contains support vectors that provide a description of the significant data for classification. Substituting $K(x_i, x_j)$ for $(x_i^T\ x_j)$ in Eqn. (1) produces a new optimisation problem:

Maximise

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i\, y_j\, \alpha_i\, \alpha_j\, K(x_i, x_j) \qquad (2)$$

Subject to $0 \leq \alpha_i \leq C,\ i = 1,...,n$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$, where C is trade-off variable.

Solving it for lagrangian multipliers α_i gives a decision function of the form

$$f(X) = \text{sgn}\left(\sum_{i=1}^{n} y_i\, \alpha_i\, K\ (x_i, x_j)\ +\ b \right) \qquad (3)$$

Conventional OvO SVM has the problem of unclassifiable region. To resolve unclassifiable region for OvO, Platt[7], *et al*. proposed DDAG approach. They have shown the existence of unclassifiable regions for *k*-class problem (k >2) which can lead to degradation of generalisation ability of classifier. In DDAG[7] scheme Vapnik Chernoff (VC) dimension, leave one out (LOO) error estimator, and Joachim's îá LOO measures were used for estimating the generalisation ability of pairwise classifier at each level of decision tree. The measures used are computationally extensive and does not take into consideration any statistical information about the data. Madzarov[8], *et al*. proposed SVM-BDT architecture. The hierarchy of binary decision subtasks using SVMs was designed with clustering algorithms. In SVM-BDT scheme, the classes are divided in two disjoint groups g_1 and g_2 using Euclidian distance as similarity measure. The two disjoint groups so obtained were then used to train a SVM classifier at the root node of the decision tree. The classes from the first and second clustering group were being assigned to left and right subtree, respectively. This process was continued recursively until there was only one class left in a group which defines a leaf in the decision tree.

3. PROPOSED OvO-ODT SVM FRAMEWORK USING STATISTICAL MEASURES

Euclidean similarity measure used in SVM-BDT does not take into account within class variability of patterns. Hence, it may not be suitable for measuring class separability between two different classes of patterns. To understand better picture of the overlap of the subspaces occupied by individual classes, statistical measure is used in pattern recognition community which measures the overlap or probabilistic distance between the two class probability distribution functions (PDFs). Hence, it constitutes a natural concept of measuring class separability. Among statistical measures, information gain is one of the most commonly used measures to measure class separability of patterns in literature. Information Gain[11] (IG) is a measure based on entropy[12] which indicates degree of disorder of a system. It measures reduction in weighted average impurity of the partitions compared with the impurity of the complete set of samples when one knows the value of a specific attribute. Thus, the value of IG signifies how the whole system is related to an attribute. IG is calculated as

$$IG(C \mid E) = H(C) - H(C \mid E) \qquad (4)$$

where $IG(C \mid E)$ is the information gain of the label C for a given attribute E, $H(C)$ represents the system's entropy and $H(C \mid E)$ is the system's relative entropy when the value of the attribute E is known. The system's entropy indicates its degree of disorder and is given by the following formula:

$$H(C) = -\sum_{i=1}^{m} p(C_i) \log p(C_i) \qquad (5)$$

where $p(C_i)$ is the probability of class i. The relative entropy is calculated as follows:

$$H(C \mid E) = \sum_{i=1}^{|E|} p(e_j) \left(-\sum_{i=1}^{m} p(C_i \mid e_j) \log p(C_i \mid e_j) \right) \qquad (6)$$

where $p(e_j)$ is the probability of value j for attribute e, and $p(C_i \mid e_j)$ is the probability of class label C_i with regard to e_j.

In OvO-ODT SVM, $k(k-1)/2$ independent binary SVM's are constructed for k-class problem. The optimal SVM model is selected on the basis of maximum value of IG that signifies more separability between patterns belonging to two different classes. IG for a given independent binary OvO SVM for training data containing n_i elements of class i and n_j elements of class j can be calculated as:

$$IG(i,j) = H(i,j) - \left[prob(i)H(t_p, f_p) + prob(j)H(f_n, t_n) \right] \qquad (7)$$

$$\text{where } H(x,y) = -x \log\left(\frac{x}{x+y} \right) - y \log\left(\frac{y}{x+y} \right) \qquad (8)$$

$$prob(i) = \frac{n_i}{(n_i + n_j)} \quad \text{and} \quad prob(j) = \frac{n_j}{(n_i + n_j)} \qquad (9)$$

t_p, f_p, t_n and f_n denote true positive, false positive, true negative and false negative, respectively.

The Gini index is another popular measure for feature selection proposed by Breiman[13], *et al*. It measures the

impurity of given set of training data D and can be calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{2} \left(prob(i) \right)^2 \quad (10)$$

For a binary split, a weighted sum of the impurity of each resulting partition is computed. The reduction in impurity that would be incurred by a particular binary split is calculated as:

$$\Delta Gini(i, j) = Gini(D) - Gini_{i,j}(D) \quad (11)$$

where $Gini_{i,j}(D) = \left[prob(i)Gini(L) + prob(j)Gini(R) \right]$ (12)

$Gini(L)$ and $Gini(R)$ are the Gini index on the left and right side of the hyperplane. OvO SVM model (i_k, j_k) that maximises the reduction in impurity *i.e.* Gini index is selected as splitting node in decision tree SVM at a particular level.

Chi-square[14] similarity measure is another criterion used for binary split in data mining and machine learning. It is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. In this case, the null hypothesis is that a random rule would place t_p patterns from class i and f_p tuples from class j independently in the left branch of decision tree and the remainder in the right branch of decision tree. The candidate decision rule would differ significantly from the random rule if the proportions differed significantly from those given by the random rule. The chi-square statistic χ^2 will be given by

$$\begin{aligned} \chi^2 = {} & g\left(t_p, \left(t_p + f_p\right) P_{pos}\right) + g\left(f_n, \left(f_n + t_n\right) P_{pos}\right) \\ & + g\left(f_p, \left(t_p + f_p\right) P_{neg}\right) + g\left(t_n, \left(t_n + f_n\right) P_{neg}\right) \end{aligned} \quad (13)$$

where $g\left(count, expect\right) = \dfrac{\left(count - expect\right)^2}{expect}$, P_{pos} is the probability of positive samples in left subtree and P_{neg} is the probability of negative samples in right subtree.

The higher the value of χ^2, the less likely it is that the null hypothesis is true. Thus, for a sufficiently high χ^2, one can reject the null hypothesis and can consider candidate rule is informative. Hence OvO SVM model (i_k, j_k) that maximises χ^2 is selected as splitting node in decision tree SVM at a particular level.

The outline for decision tree SVM using IG class separability measure for k-class is given below:

Input: Training data $X = \bigcup_{i}^{n} x_i$. The following steps are involved

Step 1. Generate the initial list $\{C_1, \ldots, C_k\}$

Step 2. Calculate $H(i, j)$ using Eqn (8) for $i = 1, \ldots, k$ and $j > i$

Step 3. Calculate, $H(t_p, f_p)$, $H(f_n, t_n)$, $prob(i)$ and $prob(j)$ using Eqn (8) and Eqn (9), respectively.

Step 4. Compute $IG(i, j)$ using Eqn (7).

Step 5. Determine class pair (C_i, C_j) for which $IG(i, j)$ takes maximum value from the list. If x_k, k=1, .., n belongs to class C_i then delete C_j from the list else delete class C_i.

Step 6. If the remaining classes are >2, repeat *Steps* 2-5 otherwise terminate the algorithm.

Similar computational steps are followed for other two measures to determine the structure of OvO-ODT SVM. To compute the time complexity of training phase, assume without loss of any generality that the number of samples in each class is approximately the same, i.e., n/k. To solve k-class problem using conventional OvO, $k(k–1)/2$ binary SVM classifiers are developed. Assuming the time complexity of building a SVM with n samples and d features is $O(n^2d)$, it can be shown that training time of conventional OvO, $T^{train}_{con\text{-}OvO\text{-}SVM}$, is $O(n^2d)$. In worst case, the decision tree generated in SVM-BDT is skewed if classes in two groups at every level are divided into uneven size. Under the assumption that group g_1 contains only one class and group g_2 contains remaining classes, the decision tree so generated will be of depth $(k–1)$ for k-class problem. Hence, the training time of SVM-BDT will be given by

$$\begin{aligned} T^{train}_{SVM\text{-}BDT\text{-}worst} &= \left(n^2 d + \left(\frac{n(k-1)}{k} \right)^2 d + \ldots + \left(\frac{2n}{k} \right)^2 d \right) \\ &\cong \left(n^2 dk \right) \end{aligned} \quad (14)$$

In SVM-BDT approach under best case, the class in two groups at every level is divided into approximately the same size. The decision tree so generated will be almost height balanced of maximum depth [log(k)]. The number of nodes in decision tree at depth i is 2^{i-1}, each containing $n/2^{i-1}$ samples. Hence, the training time for SVM-BDT in best case is given by

$$\begin{aligned} T^{train}_{SVM\text{-}BDT\text{-}best} = \Bigg(& n^2 d + 2\left(\frac{n}{2} \right)^2 d + 4\left(\frac{n}{2} \right)^2 d + \ldots \\ & + 2^{\log(k)-1} \left(\frac{n}{2^{\log(k)-1}} \right)^2 d \Bigg) \cong \left(n^2 d \right) \end{aligned} \quad (15)$$

However in general, the structure of decision tree generated using statistical measures is almost height balanced of maximum depth [log(k)]. There are 2^{i-1}nodes at i^{th} level and each node uses $(2n/k)$ samples. Hence, the training time for OvO-ODT SVM using statistical measure is given by

$$T^{train}_{OvO\text{-}ODT\text{-}SVM} \cong \sum_{i=1}^{\log(k)} 2^{i-1} \left(\frac{2n}{k} \right)^2 d \cong n^2 \left(\frac{d}{k} \right) \quad (16)$$

During testing phase of the conventional OvO, $k(k–1)/2$ decision functions are to be evaluated. Also, the majority voting is computed with $k(k–1)/2$ operation. Hence, the testing time $T^{test}_{Con\text{-}OvO}$ for each sample is given by $k(k–1)/2$. In worst case, the depth of SVM-BDT is $(k–1)$. The testing time $T^{test}_{SVM\text{-}BDT\text{-}worst}$ for each sample in

worst case is k. However in best case the depth of SVM-BDT is [log(k)] which requires [log(k)] testing time $T^{test}_{SVM\text{-}BDT\text{-}best}$ for each sample. Since, the maximum depth of OvO-ODT SVM is [log(k)], the testing time requires [log(k)] operations. According to the above analysis, it is evident that the training and testing time for OvO-ODT SVM will always require less computation time in comparison to conventional OvO and SVM-BDT.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed OvO-ODT SVM framework using information gain, gini index and chi-square. They have performed experiments on publically available UCI[15] benchmark datasets. Table 1 describes the datasets used in the experiments. All the experiments were performed on a computer having Pentium 4 dual-core processor with 1GB RAM. The kernel functions used in experiments are given in Table 2.

The classification accuracy is determined using ten cross-validations. The authors have used $\gamma = [2^{-11}, 2^{-10}, 2^{-9} \ldots 2^{0}]$ and trade-off variable C=1000. For a given kernel function, the authors determined the value of γ for which the maximum classification accuracy was achieved. Table 3 shows the comparison of maximum classification accuracy between conventional OvO SVM and SVM-BDT with the

Table 1. Description of datasets

Problem	#Train_data	# Class	#Attributes
Wine	178	3	13
Vehicle	846	4	18
Glass	214	6	9
Segmentation	210	7	19
Ecoli	336	8	7

Table 2. Kernel functions

Kernel Function	$K(x, x_i)$ for $\gamma > 0$
Gaussian	$exp\left(-\gamma \mid x - x_i \mid^2\right)$
Laplace	$exp\left(-\gamma \mid x - x_i \mid\right)$
Cauchy	$\left(1 \Big/ \left(1 + \gamma \mid x - x_i \mid^2\right)\right)$
Hypersecant	$2 \Big/ \left(exp\left(\gamma \mid x - x_i \mid\right) + exp\left(-\gamma \mid x - x_i \mid\right)\right)$
Square sync	$sin^2\left(\gamma \mid x - x_i \mid\right) \Big/ \left(\gamma \mid x - x_i \mid^2\right)$

Table 3. Comparison of classification accuracy (per cent)

Dataset	Kernel Choice	OvO	SVM-BDT	OvO-ODT SVM		
				χ^2	Gini	IG
Wine	Gaussian	82.58	81.88	82.58	82.58	**83.15**
	Laplace	**82.58**	81.76	**82.58**	**82.58**	**82.58**
	Cauchy	**82.02**	81.76	**82.02**	**82.02**	**82.02**
	Hypersecant	**93.26**	92.63	**93.26**	**93.26**	92.13
	Square sync	75.28	75.46	**76.97**	**76.97**	75.28
Vehicle	Gaussian	**76.83**	75.24	76.60	76.60	76.60
	Laplace	**77.42**	76.24	**77.42**	**77.42**	**77.42**
	Cauchy	**76.48**	74.65	**76.48**	**76.48**	**76.48**
	Hypersecant	**83.33**	82.98	83.22	**83.33**	**83.33**
	Square sync	**71.51**	71.08	70.922	70.80	70.8
Glass	Gaussian	72.43	65.42	70.56	**72.90**	71.03
	Laplace	75.70	68.69	**76.17**	**76.17**	75.24
	Cauchy	72.90	69.42	71.03	**73.36**	71.03
	Hypersecant	**71.96**	69.47	71.09	**71.96**	70.09
	Square sync	**66.36**	58.41	62.62	58.13	62.62
Segmentation	Gaussian	84.76	82.89	**85.24**	**85.24**	84.29
	Laplace	87.14	87.05	**87.62**	**87.62**	87.14
	Cauchy	**86.19**	85.05	**86.19**	**86.19**	85.71
	Hypersecant	90	90	**90.95**	**90.95**	89.05
	Square sync	**81.9**	73.80	79.52	79.05	80.95
Ecoli	Gaussian	**85.42**	81.34	**85.42**	**85.42**	**85.42**
	Laplace	**87.20**	85.78	**87.20**	**87.20**	86.91
	Cauchy	**85.42**	79.90	**85.42**	**85.42**	84.82
	Hypersecant	**85.42**	83.89	83.33	**85.42**	84.82
	Square sync	85.12	82.78	**87.20**	82.85	85.12

Table 4. Comparison of training and testing time (Seconds)

Dataset	Kernel choice	Training time (s)					Testing time (s)				
		OvO	SVM-BDT	OvO-ODT SVM			OvO	SVM-BDT	OvO-ODT SVM		
				χ^2	Gini	IG			χ^2	Gini	IG
Wine	Gaussian	2.15	2.15	**2.05**	2.13	2.11	0.08	0.04	**0.03**	**0.03**	**0.03**
	Laplace	2.98	2.83	2.67	2.18	**1.48**	0.07	0.04	**0.03**	**0.03**	**0.03**
	Cauchy	2.89	1.98	1.69	**1.34**	1.54	0.08	0.08	0.06	0.07	**0.03**
	Hypersecant	3.55	3.22	2.15	**2.13**	2.29	0.06	0.05	0.03	0.06	**0.02**
	Square sync	3.13	2.93	2.27	**2.18**	2.60	0.07	0.03	0.03	**0.02**	0.03
Vehicle	Gaussian	189.96	185.45	**155.43**	170.92	155.79	1.85	0.56	**0.46**	0.53	**0.46**
	Laplace	188.68	187.55	**134.34**	134.54	134.54	1.45	0.57	0.56	**0.35**	0.57
	Cauchy	186.44	186.57	144.34	144.54	**144.23**	1.73	0.73	0.46	**0.45**	0.73
	Hypersecant	198.69	181.75	**145.45**	148.78	145.75	1.64	0.76	0.56	**0.46**	0.76
	Square sync	189.75	189.57	159.72	159.65	**149.77**	1.63	0.66	**0.53**	0.59	0.66
Glass	Gaussian	12.31	11.88	**8.99**	10.06	9.05	1.20	0.32	**0.22**	0.28	**0.22**
	Laplace	11.90	10.53	9.84	5.45	**4.80**	1.21	0.66	**0.14**	0.18	0.18
	Cauchy	13.62	11.78	**7.62**	9.56	10.71	1.43	0.89	**0.15**	0.17	0.17
	Hypersecant	8.36	7.35	6.36	7.46	**5.93**	1.21	0.79	**0.15**	0.18	0.19
	Square sync	8.93	8.04	5.93	**4.92**	5.42	1.25	0.99	**0.11**	0.14	0.19
Segmentation	Gaussian	3.00	2.98	**2.39**	2.45	2.49	1.40	0.55	**0.42**	0.54	0.44
	Laplace	3.15	2.90	**2.34**	2.52	2.42	1.35	0.89	0.53	0.43	**0.33**
	Cauchy	3.10	2.89	2.13	2.12	**2.10**	1.78	0.76	0.34	**0.32**	0.33
	Hypersecant	4.91	4.89	4.57	4.56	**4.21**	1.24	0.54	0.36	**0.31**	**0.31**
	Square sync	3.99	3.81	3.61	3.21	**3.03**	1.54	0.64	0.38	**0.32**	0.35
Ecoli	Gaussian	26.48	25.99	**17.55**	21.97	20.07	1.55	0.88	**0.54**	0.60	0.96
	Laplace	24.38	23.45	17.34	**16.87**	17.34	1.82	1.78	1.05	1.12	**1.02**
	Cauchy	27.85	25.83	21.85	**19.85**	24.85	2.31	1.20	0.99	0.98	**0.92**
	Hypersecant	20.98	20.23	19.78	**16.78**	19.78	1.82	1.15	**1.03**	1.10	1.04
	Square sync	23.65	21.94	20.85	**20.35**	20.85	1.88	1.33	1.04	**1.09**	1.03

proposed OvO-ODT SVM using all the three statistical measures. The best classification accuracy for each dataset is shown in bold. From Table 3 it can be observed that the classification accuracy of the proposed OvO-ODT SVM method is better or equivalent to conventional OvO SVM and SVM-BDT for different choices of kernel functions for most of datasets using all three statistical measures. Among the three measures used for determining the structure of decision tree, neither of them is clear winner over the others.

Table 4 shows the computation time for training and testing phase for all kernels with $\gamma = 2^{-11}$ and trade-off variable C=1000. The minimum training and testing time for each dataset is shown in bold. It can be observed from Table 4 that time required for training OvO-ODT SVM using all the three statistical measures is less in comparison to conventional OvO and SVM-BDT. Among the three measures used for building OvO-ODT SVM, neither of them is a clear winner over the other measures. It can also be observed that the proposed OvO-ODT SVM is also computationally better than conventional OvO and SVM-BDT in terms of testing time for all the three measures.

5. CONCLUSIONS

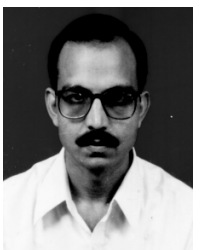
In this paper, OvO-ODT SVM framework is proposed to solve multi-class problems where the optimal structure of decision tree is determined using statistical measures, i.e., information gain, gini index, and chi-square. It has also been shown theoretically that the computation time of training and testing of OvO-ODT SVM using statistical measures is better in comparison to conventional OvO SVM and SVM-BDT. The performance of the proposed framework is evaluated in terms of classification accuracy and computation time of training and testing phases. Experimental results on UCI repository dataset demonstrate better or equivalent performance of the proposed OvO-ODT SVM scheme in comparison to conventional OvO SVM and SVM-BDT in terms of classification accuracy for most of the datasets using all the three measures employed for construction of decision tree. The experimental results obtained also demonstrate that the training and testing time of proposed OvO-ODT SVM takes less computation time in comparison to conventional OvO SVM and SVM-BDT.

REFERENCES

1. Vapnik, V.N. Statistical Learning Theory. New York, John Wiley & Sons, 1998.
2. Joachim's, T.; Cristianini, N.; & Taylor, J. S. Composite Kernels for hypertext categorisation. *In* The International Conference on Machine Learning, 2001.
3. Mori, S.; Suen, C.Y. & Yamamot, K. Historical review of OCR research and development. *Proceedings IEEE*, 1992, **80**, 1029-058.
4. Schmidt, M. Identifying speaker with support vector networks. *In* Interface '96 Proceedings*,* 1996, Sydney.
5. Rifkin, R. & Klautau, A. In defence of one-vs-all classification. *J. Machine Learning ,* 2004, **5**, 101-41.
6. Hsu, C.W. & Lin, C.J. A comparison of methods for multi-class support vector machine. *IEEE Trans. Neural Networks*, 2002, **13**(2), 415-25.
7. Platt, Cristianini N. & Shawe-Taylor, J. Large margin DAGSVM's for multi-class classification. *Adv. Neural Info. Proc. Syst.,* 2000, **12**, 547-53.
8. Gjorgji, M.; Dejan, G. & Ivan, C. A multi-class SVM classifier utilising binary decision tree. *Informatica,* 2009, **33**, 233-41.
9. Kittler, J. & Hojjatoleslami, A. A weighted combination of classifiers employing shared and distinct representations. *IEEE Comp. Vision Pattern Recog.*, 1998, 924-29.
10. Scholkopf, B. & Smola, A. Learning with kernels. Cambridge, MA, MIT Press, 2002.
11. Bala, M. & Agrawal, R. K. Evaluation of decision tree SVM framework using different statistical measures. *ARTcom,* 2009, 341-45.
12. Quinlan, J.R. Induction of decision trees. *Machine Learning ,* 1986, **1**, 81-106.
13. Breiman, L.; Friedman, J.; Ohlsen, R. & Stone, C. Classification and regression trees. Belmont, CA, Wadsworth, 1984.
14. Duda, R. O.; Hart, P. E. & Stork, D. E. Pattern classification, Ed 2. John Wiley & Sons, 2000.
15. www.ics.uci.edu/~mlearn.MLRepository.html.

Contributors

Ms Manju Bala doing her PhD at the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. She received BE(Comp. Sci. and Engg.) from Maharishi Dayanand University and MTech (Comp. Sci.) from IASE Deemed University, Rajasthan. Presently, she is working as an Assistant Professor in Computer Science department, I.P. College for Women, University of Delhi, New Delhi. Her current area of research is pattern recognition.



Mr Ramesh Kumar Agrawal obtained MTech (Computer Application) from Indian Institute of Technology Delhi, New Delhi. He has done his PhD (Computational Physics) from Delhi University. Presently, he is working as an Associate Professor at the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current areas of research are: Classification, feature extraction and selection for pattern recognition problems in domains of image processing, security, and bioinformatics.