Optimal Hierarchical Structure Design of Decision Tree SVM Using Distance Based Approach

Manju Bala
IP College for Women,
University of Delhi
Manjugpm@gmail.com

***Abstract* -** In literature multi-class SVM is constructed using One against All, One against One and Decision tree based SVM using Euclidean and Mahalanobis distance. To maintain high generalization ability, the most separable classes should be separated at the upper nodes of decision tree. In this paper, A deterministic quantitative model based on distance based approach (DBA) method has been developed and applied for evaluation, optimal selection SVM model for the first time. DBA recognizes the need for relative importance of criteria for a given application, without which inter-criterion comparison could not be accomplished. It requires a set of model selection criteria like information gain, gini index, chi-squared, chernoff-bound, kullbak divergence and scatter-matrix-based class separability in kernel-induced space, along with a set of SVM Models and their level of criteria for optimal selection, and successfully presents the results in terms of a merit value which is used to rank the SVM models. One real dataset from distinct published papers have been used for demonstration of DBA method. The result of this study will be a selection of SVM Model at the root node of decision tree One Versus One (OvO) SVM based on the Euclidean composite distance of each alternative to the designated optimal SVM Model. It is shown that the Optimal Decision Tree (ODT) SVM requires less computation time in comparison to conventional One against All SVM. Experimental results on UCI repository dataset demonstrates better or equivalent performance of our proposed decision tree scheme in comparison to conventional One versus One (OvO) SVM in terms of classification accuracy for most of the datasets. The proposed scheme outperforms conventional One versus One SVM in terms of computation time for both training and testing phase using DBA approach employed for determining the structure of decision tree.

Index Terms*—*Distance based approach, Model selection, Model selection criteria, SVM, Kernel function.

ACRONYM

SVM	Support Vector Machine
OvO	One vesus One
ODT	Optimal Decision Tree
IG	Information Gain
Gini Index	GI
CB	Chernoff Bound
KD	Kullback Divergence
SC	Chi-squared Ratio of interclass and intra class scatters in kernel-induced

NOTATION

δ	distance metric
χ^2	Chi-squared
r_{ij}	indicator value for alternative SVM i for attribute j
$\overline{r}_j$	average of attribute j
S_j	standard deviation of attribute j
CD	Euclidean composite distance

*****

I. Introduction

In the past two decades valuable work has been carried out in the area of text categorization [1], optical character recognition [2], speech recognition [3], handwritten digit recognition [4] etc. All such real-world applications are essentially multi-class classification problems. Various classification techniques have been suggested in data mining and machine learning i.e. C4.5, Artificial neural networks, Bayesian classification, Support Vector Machines etc.

Support Vector Machines (SVM) is based on statistical learning theory developed by Vapnik [5], [6]. It is originally formulated for binary-class problems. Larger and more complex classification problems have subsequently been solved with SVM. How to effectively extend it for multi-class classification is still an ongoing research issue [7]. The most common way to build a multi-class SVM is by constructing and combining several binary classifiers [8]. To solve multi-class classification problems, we divide the whole pattern into a number of binary classification problems. The two representative ensemble schemes are One against One (OvO) and One against All (OAA) [7]. It has been reported in literature that both conventional OvO and OAA SVMs suffer from the problem of unclassifiable region [9], [10]. To resolve unclassifiable region Platt, Cristianini and Shawe-Taylor [10] proposed decision tree OvO SVM formulation to overcome unclassifiable region. Takahashi and Abe [9] proposed decision tree OAA SVM based on class separability measures i.e. Euclidean distance between class centers and Mahalanobis distance. In literature, other than Euclidean distance and Mahalanobis distance a large number of distance metrics were used to determine the class separability, each having its own advantages and disadvantages. Few more realistic and effective statistical measures used in literature are gini index, scatter-matrix-based class separability in kernel-induced space, chi-squared, kullbak divergence, chernoff-bound and information gain for measuring class separability.

In this paper, we propose construction of OvO ODT-SVM where class separability is determined using CD obtained by mingling various class separability criteria through DBA approach. The remainder of the paper is organized as follows: Section II reveals the existing literature for different types of decision tree based approaches for OvO SVM. In section III, various class separability criteria are identified. The distance based approach (DBA) method is explained in Section IV and Section V describes the algorithm used for DBA based OvO ODT-SVM model. The demonstration with the help of illustrated examples to develop a procedure mingling various class separability criteria for comprehensive selection of the alternative SVM models at each nonleaf node of decision tree is described in Section VI. Finally, the conclusions are given in Section VII.

II. Literature Review

The most common way to build a multi-class SVM is by constructing and combining several binary classifiers [8]. To solve multiclass classification problems, we divide the whole classification problem into a number of binary classification problems. The two representative ensemble schemes are OvO and OAA [7].

Convetional OvO SVM has the problem of unclassifiable region.To resolve unclassifiable region for OvO SVM (Decision directed Acyclic graph (DDAG) SVM) Platt, Cristianini and Shawe-Taylor [10] proposed decision tree OvO SVM formulation. They have shown with an example three-class problem the existence of unclassifiable regions which can lead to degradation of generalization ability of classifier. In general, the unclassifiable region is visible and generalization ability of classifier is not good for k-class problem where k >2.

In DDAG OvO scheme [10], VC dimension, LOO error estimator and Joachim's ξα LOO measures were used for estimating the generalization ability of pairwise classifier at each level of decision tree. During training at the top node, a pair (C_i, C_j) that has the highest generalization ability is selected from an initial list of classes (C_1,…,C_k). Then it generates the two lists deleting C_i or C_j from the initial list. If the separated classes include the plural classes, at the node connected to the top node, the same procedure is repeated for the two lists till one class remains in the separated region. This means that after only k-1 steps just one class remains, which therefore becomes the prediction for the current test sample.

Madzarov, Gjorgjevikj and Chorbev [11] proposed binary tree architecture (SVM-BDT) that uses SVMs for making binary decisions in the nodes 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. In proposed scheme SVM-BDT, the classes are divided in two disjoint groups g_1 and g_2 using Euclidian distance as distance measure. The two disjoint groups so obtained are then used to train a SVM classifier in the root node of the decision tree. The classes from first and second clustering group are being assigned to left and right subtree respectively. This process continues recursively until there is only one class is left in a group which defines a leaf in the decision tree. gini index, scatter-matrix-based class separability in kernel-induced space, chi-squared, kullbak divergence, chernoff-bound and information gain used separately in the construction of decision tree (i.e. SVM-BDT and Takahashi and Abe [3] OAA SVM formulation) 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. So there was a need to find a robust class separability measure called Euclidean composite distance measure obtained by DBA approach to determine the hierarichal structure of OvO ODT-SVM.

III. Class Separability Criteria for Support vector Machine Models

A model can be judged according to its ability to reproduce higher generalization ability. For optimally selection of a SVM model as the root node from a set of available models, there are a set of comparison criteria is available to compare models quantitatively. The comparison criteria we used are described as follows:

1) Gini Index (GI)

The Gini index is another popular measure for feature selection in the field of data mining proposed by [14]. It measures the impurity of given set of training data D and can be calculated as

$$GI(D) = 1 - \sum_{i=1}^{2}(prob(i))^2 \quad (1)$$

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 in OvO ODT-SVM between two classes of dataset classes i and j is calculated as below

$$\Delta GI(i,j) = GI(D) - GI_{i,j}(D) \quad (2)$$

where $GI_{i,j}(D) = [prob(i)GI(L) + prob(j)GI(R)]$

GI(L) is the gini index on the left side of the hyper plane and GI(R) is that on the right.

2) Ratio of interclass and intra class scatters in kernel-induced (SC)

To measure class variability of patterns, the ratio of interclass and intra class scatters in kernel-induced feature space can also be used which better depicts the physical relationship of data in input space and thereby providing high generalization ability of classifier based on decision tree. The scatter-matrix-based measure (S) of training set D in original space is defined as

$$S = \frac{tr\left(S_b\right)}{tr\left(S_w\right)} \tag{3}$$

where S_w is the within class scatter matrix and S_b is the between class scatter matrix, defined as

$$S_b = \left(m_i - m_j\right)\left(m_i - m_j\right)' \text{ where}$$

$$m_i = \frac{1}{n_i}\sum_{x \in C_i} x \text{ and } m_j = \frac{1}{n - n_i}\sum_{x \notin C_i} x \tag{4}$$

$$S_w = Q_i + Q_j \tag{5}$$

where Q_i and Q_j given as

$$Q_i = \frac{1}{n_i}\sum_{x \in C_i}\left(x - m_i\right)\left(x - m_i\right)'$$

$$Q_j = \frac{1}{n - n_i}\sum_{x \notin C_i}\left(x - m_j\right)\left(x - m_j\right)'$$

Using kernel trick, samples from class i and class j are implicitly mapped from R^{d} to a feature space, F. Let $\emptyset(\cdot) : \mathrm{R}^{\mathrm{d}} \rightarrow \mathrm{F}$ denote the mapping and $\mathrm{k}_\theta\left(\mathrm{x}_\mathrm{i}, \mathrm{x}_\mathrm{j}\right) = \langle \emptyset(\mathrm{x}_\mathrm{i}), \emptyset\left(\mathrm{x}_\mathrm{j}\right) \rangle$ denote the kernel function, where θ is the set of kernel parameters and $\langle \cdot, \cdot \rangle$ is the inner product. K denotes the kernel matrix and $\{\mathrm{K}\}_{\mathrm{i,j}}$ is defined as $\mathrm{k}_\theta\left(\mathrm{x}_\mathrm{i}, \mathrm{x}_\mathrm{j}\right)$. Let $\mathrm{K}_{\mathrm{A,\,B}}$ be kernel matrix computed with the samples from A and B denote two subsets of training sample set D. Let $\mathrm{S}_\mathrm{b}^\emptyset$ and $\mathrm{S}_\mathrm{w}^\emptyset$ denotes the between class scatter matrix and within class scatter matrix in F, respectively and defined as follows

$$\mathrm{S}_\mathrm{b}^\emptyset = \sum_{\mathrm{i}=1}^{2} \mathrm{n}_\mathrm{i}\left(\mathrm{m}_\mathrm{i}^\emptyset - \mathrm{m}^\emptyset\right)\left(\mathrm{m}_\mathrm{i}^\emptyset - \mathrm{m}^\emptyset\right)^\mathrm{T} \tag{6}$$

$$\mathrm{S}_\mathrm{w}^\emptyset = \sum_{\mathrm{i}=1}^{2} \sum_{\mathrm{x} \in \mathrm{D}_\mathrm{i}}\left(\emptyset(\mathrm{x}) - \mathrm{m}_\mathrm{i}^\emptyset\right)\left(\emptyset(\mathrm{x}) - \mathrm{m}_\mathrm{i}^\emptyset\right)^\mathrm{T} \tag{7}$$

Where $\mathrm{m}_\mathrm{i}^\emptyset$ denotes the mean of training samples from class i and $\mathrm{m}^\emptyset$ is the mean of all the training samples in F. $m_i^\emptyset$ and $m_j^\emptyset$ denotes the mean vectors of training samples from the classes i and j in F. let H is vector whose elements are all “1”. Its size will be decided by the context.

$$m_i^{\emptyset\,T} m_i^\emptyset = n_i^{\,-2} \cdot H^T K_{D_i,D_i} H \tag{8}$$

$$m_j^{\emptyset\,T} m_j^\emptyset = n_j^{\,-2} \cdot H^T K_{D_j,D_j} H \tag{9}$$

$$m_i^{\emptyset\,T} m_j^\emptyset = \left(n_i n_j\right)^{-2} \cdot H^T K_{D_i,D_j} H \tag{10}$$

$$tr\left(S_b^\emptyset\right) = tr\left[\sum_{i=1}^{2} n_i\left(m_i^\emptyset - m^\emptyset\right)\left(m_i^\emptyset - m^\emptyset\right)^T\right]$$

$$= \frac{H^T K_{D_i,D_i} H}{n_i} + \frac{H^T K_{D_j,D_j} H}{n_j} - \frac{H^T K_{D,D} H}{n} \tag{11}$$

$$tr\left(S_w^\emptyset\right) = tr\left[\sum_{i=1}^{2}\sum_{x \in D_i}\left(\emptyset(x) - m_i^\emptyset\right)\left(\emptyset(x) - m_i^\emptyset\right)^T\right]$$

$$= tr\left(K_{D,D}\right) - \frac{H^T K_{D_i,D_i} H}{n_i} - \frac{H^T K_{D_j,D_j} H}{n_j} \tag{12}$$

Now the class separability in a feature space F is obtained as

$$SC = \frac{tr\left(S_b^\emptyset\right)}{tr\left(S_w^\emptyset\right)} \tag{13}$$

3) Chi-squared (χ^2)

Chi-squared [15] is another criterion used for binary split in data mining and machine learning, is statistical test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. We are interested in determining whether a particular decision rule is useful or informative. In this case, the null hypothesis is that a random rule would place $\mathrm{t_p}$ patterns from class i and $\mathrm{f_p}$ 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-squared statistic χ^2 will be given by

$$\chi^2 = \mathrm{g}\left(\mathrm{t_p}, (\mathrm{t_p} + \mathrm{f_p})\mathrm{P_{pos}}\right) + \mathrm{g}\left(\mathrm{f_n}, (\mathrm{f_n} + \mathrm{t_n})\mathrm{P_{pos}}\right) + \mathrm{g}\left(\mathrm{f_p}, (\mathrm{t_p} + \mathrm{f_p})\mathrm{P_{neg}}\right) + \mathrm{g}\left(\mathrm{t_n}, (\mathrm{f_n} + \mathrm{t_n})\mathrm{P_{neg}}\right) \tag{14}$$

$$\text{Where } \mathrm{g}(\mathrm{count}, \mathrm{expect}) = \frac{(\mathrm{count} - \mathrm{expect})^2}{\mathrm{expect}} \tag{15}$$

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, the difference between the expected and observed distributions is statistically significant; one can reject the null hypothesis and can consider candidate rule is informative.

4) Kullbak Divergence (KD)

In order to obtain a quantitative measure of how separable are two classes, a distance measure can be easily extracted from some parameters of the data. A very important aspect of probabilistic distance measures is that a number of these criteria can be analytically simplified in the case when the class conditional p.d.f.s $p(\mathbf{X}_k \mid C_i)$ follows multivariate normal distribution. The class conditional p.d.f.s $p(\mathbf{X}_k \mid C_i)$ of k-dimensional sample $\mathbf{X_k}$ = [x_1, x_2, …, x_k] for a given class C_i , i=1, 2 is given by

$$p(\mathbf{X}_k \mid C_i) = \frac{1}{(2\pi)^{d/2} \mid \boldsymbol{\Sigma}_k^i \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}_k - \boldsymbol{\mu}_k^i)^t (\boldsymbol{\Sigma}_k^i)^{-1} (\mathbf{X}_k - \boldsymbol{\mu}_k^i) \right] \tag{16}$$

where $\boldsymbol{\mu}_k^i$ is a mean vector and $\boldsymbol{\Sigma}_k^i$ is a covariance matrix for class C_i .

In literature, for multivariate normal distribution for two classes with k attributes, KD measure is given as follows [16]:

$$J_k^D = \frac{1}{2}(\boldsymbol{\mu}_k^2 - \boldsymbol{\mu}_k^1)^t \left(\left(\Sigma_k^1\right)^{-1} + \left(\Sigma_k^2\right)^{-1}\right)(\boldsymbol{\mu}_k^2 - \boldsymbol{\mu}_k^1)$$
$$+\frac{1}{2} tr\left(\left(\Sigma_k^1\right)^{-1} \Sigma_k^2 + \left(\Sigma_k^2\right)^{-1} \Sigma_k^1 - 2I_k\right) \tag{17}$$

5) Chernoff Bound (CB)

In literature, for multivariate normal distribution for two classes with k attributes, CB measure is given as follows [16]:

$$J_k^c = \frac{1}{2}\beta(1-\beta)(\boldsymbol{\mu}_k^2 - \boldsymbol{\mu}_k^1)^t \left[(1-\beta)\Sigma_k^1 + \beta\Sigma_k^2\right]^{-1} (\boldsymbol{\mu}_k^2 - \boldsymbol{\mu}_k^1)$$
$$+\frac{1}{2}\log \frac{\left|(1-\beta)\Sigma_k^1 + \beta\Sigma_k^2\right|}{\left|\Sigma_k^1\right|^{1-\beta} \left|\Sigma_k^2\right|^{\beta}} \tag{18}$$

6) Information Gain (IG)

Among staistical measures information gain (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 we know the value of a specific attribute. Thus, the value of IG signifies how the whole system is related to an attribute. IG is calculated using: It is popularized in machine learning by Quinlan [13].

$$\mathrm{IG(C|E) = H(C) - H(C|E)} \tag{19}$$

where IG(C|E) is the information gain of the label for a given attribute E, H(C) the system's entropy and is H(C|E) the system's relative entropy when the value of the label E is known.

The system's entropy indicates its degree of disorder and is given by the following formula:

$$\mathrm{H(C)} = \sum_{i=1}^{m} p(c_i) \log p(c_i) \tag{20}$$

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

$$\mathrm{H(C|E)} = \sum_{i=1}^{|E|} p(e_j)\left(-\sum_{i=1}^{m} p(c_i|e_j)\right) \log p\left(c_i|e_j\right)\Big) \tag{21}$$

where $p(e_j)$ is the probability of value j for attribute e, and $p(c_i|e_j)$ is the probability of c_i with regard to e_j

IV. Distance Based Approach (DBA) Method

The development of the distance based approach (DBA) method begins with defining the optimal state of the overall objective, and specifies the ideally good values of attributes involved in the process. The optimal state of the objective is represented by the optimum model, the OPTIMAL. The vector OP $(r_1, r_2, \ldots, r_n)$ is the set of "optimum" simultaneous attributes values. In an n-dimensional space, the vector OP is called the optimal point. For practical purposes, the optimal good value for attributes is defined as the best values which exist within the range of values of attributes. The OPTIMAL, then, is simply the SVM that has all the best values of attributes.

It may happen that a certain SVM has the best values for all attributes, this is very unlikely. Instead, a variety of alternatives may be used to simulate the optimal state. For this reason, the OPTIMAL has not to be considered as feasible alternatives, but it is used only as reference to which other alternatives are quantitatively compared. The numerical difference resulting from comparison represents the effectiveness of alternatives to achieve the optimal state of objective. Hence, here, the decision problem is to find a feasible solution which is as close as possible to the optimal point. The objective function for finding such a solution can be formulated as:

$$Minimize \ \ \delta\left\{Alt(x), OPTIMAL\right\} \tag{22}$$

$$Subject\ to\ \mathrm{x}\ \varsigma\ \mathrm{X}$$

where { $Alt(x)$ } and δ represent a SVM alternative in the n-dimensional space, and the distance from the optimal point, respectively. Thus the problem and its solutions depend on the choice of optimal point, OPTIMAL, and the distance metric, δ , used in the model. In two dimensional spaces, this solution function can be illustrated as in Fig. 2, where H is the feasible region and the OP is the optimal point.

The DBA method determines the point in H region which is "the closest" to the optimal point and is graphically explained in Fig. 3 for two dimensional cases. Note that the lines $(Alt-OP)_{X1}$, and $(Alt-OP)_{X2}$ are parallel to X1 and X2 axis respectively. Consequently, $(Alt-OP)_{X1} =\mid OP_{X1} - Alt_{X1} \mid$ and $(Alt-OP)_{X2} =\mid OP_{X2} - Alt_{X2} \mid$. Based on Pythagoras theorem, in two dimensional space, δ is:

$$\delta = \left[(OP_{X1} - Alt_{X1})^2 + (OP_{X2} - Alt_{X2})^2 \right]^{1/2} \qquad (23)$$

In general terms, the “distance δ ” can be formulated as:

$$\delta = \left[\Sigma (OP_{ij} - Alt_{ij})^2 \right]^{1/2} \qquad (24)$$

where i=1, 2, 3, 4... n = alternative SVMs
j=1, 2, 3... m = selection attributes

To implement the above approach, let us assume that we have a complete set of SVMs consisting of 1, 2, 3,...n SVMs, and 1,2,3...m selection attributes corresponding to each alternative SVM, $Alt_1(r_{11}, r_{12}, ..., r_{1m})$, $Alt_2(r_{21}, r_{22}, ..., r_{2m})$, $Alt_n(r_{n1}, r_{n2}, ..., r_{nm})$, and the OPTIMAL $(r_{b1}, r_{b2}, ..., r_{bm})$ where r_{bm} = the best value of attribute ‘m’. The whole set of alternatives can be represented by the following matrix,

$$[r] = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \\ r_{b1} & r_{b2} & \cdots & r_{bm} \end{bmatrix} \qquad (25)$$

Thus, in this matrix; a vector in an m-dimensional space represents every SVM alternative. In order to ease the process, and in the same time to eliminate the influence of different units of measurement, the matrix is standardized using Z formula as:

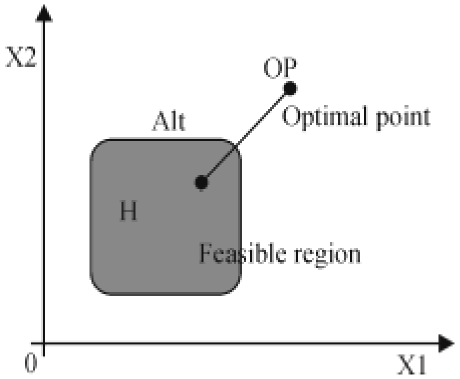

Fig. 2. Distance Based Approach

$$Z_{ij} = \frac{r_{ij} - \overline{r}_j}{S_j} \qquad (26)$$

here, $\overline{r}_{ij} = \frac{1}{n} \sum_{i=1}^{n} r_{ij}$; and (27)

$$S_j = \left[\frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)^2 \right]^{1/2} \qquad (28)$$

where i = 1, 2, 3, ... , n and j = 1, 2, 3, … , m.

$\overline{r}_j$ and S_j represent the average value and standard deviation of each attribute for all alternative SVMs whereas m and n represent number of different SVM attributes and number of alternate SVMs, respectively.

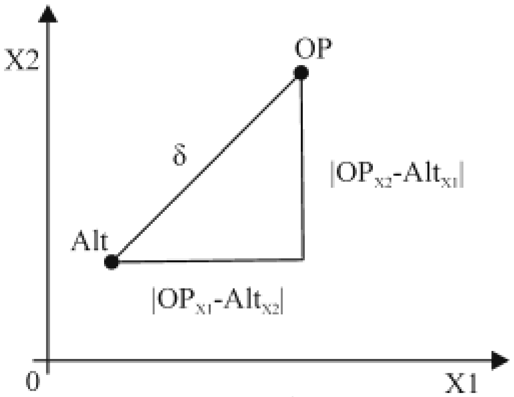
X2
OP
δ
|OP_{X2}-Alt_{X1}|
Alt
|OP_{X1}-Alt_{X2}|
0
X1

Fig. 3. Distances of real vector

$$[Z_{std}] = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1m} \\ Z_{21} & Z_{22} & \cdots & Z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nm} \\ Z_{OP1} & Z_{OP2} & \cdots & Z_{OPm} \end{bmatrix} \qquad (29)$$

where $Z_{11} = \frac{r_{11} - \overline{r}_1}{S_1}$, $Z_{12} = \frac{r_{12} - \overline{r}_2}{S_2}$, $Z_{1m} = \frac{r_{1m} - \overline{r}_m}{S_m}$,

The next step is to obtain the difference from each alternative to the reference point, the OPTIMAL, by subtracting each element of optimal by correspondence element in the alternative set. This results in another interim matrix:

$$[Z_{dis}] = \begin{bmatrix} Z_{OP1} - Z_{11} & Z_{OP2} - Z_{12} & \cdots & Z_{OPm} - Z_{1m} \\ Z_{OP1} - Z_{21} & Z_{OP2} - Z_{22} & \cdots & Z_{OPm} - Z_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ Z_{OP1} - Z_{n1} & Z_{OP2} - Z_{n2} & \cdots & Z_{OPm} - Z_{nm} \end{bmatrix} \qquad (30)$$

Finally the Euclidean composite distance, CD, between each alternative SVM to the optimal state, OPTIMAL, is derived from the following formula:

$$CD_{OP-Alt} = \left[\sum_{j=1}^{m} (Z_{OPj} - Z_{ij})^2 \right]^{1/2} \qquad (31)$$

Within any given set of SVM’s alternatives, this distance of each alternative to every other is obviously a composite distance. In other words, it can be called as the mathematical expression of several distances on each of several dimensions in which SVMs can be compared.

V. Distance Based Approach (DBA) OvO ODT-SVM

The outline for OvO ODT-SVM using DBA method for k-class is given below:

1. Generate the initial list {C_1, …, C_k)
2. Calculate Euclidean composite distance, CD using eqn (31) for i = 1, …, k and j > i
3. Determine class pair (C_i, C_j) for which CD takes maximum value from the list. If X belongs to class C_i then delete C_j from the list else delete class C_i.
4. If the remaining classes exist, repeat Steps 2-3 otherwise terminate the algorithm.

VI. Experimental Results

The objective of this demonstration is to test the suitability of the developed Distance Based Approach method so that a comprehensive ranking of the alternative SVMs could be made combining various attributes relevant to SVMs for a data set. Table I describes the datasets used in our experiments. All the experimental tests were performed on a computer having Pentium 4 dual core processor with 1GB RAM. The kernel functions used in experiments are given in Table II. To see the effectiveness of our proposed DBA based OvO ODT-SVM, we compared our methods with conventional OvO SVM. We have used five kernel functions with value of C = 1000 and $\gamma = [2^{-11}, 2^{-10}, 2^{-9} \ldots 2^{0}]$. The classification accuracy is determined using ten cross-validations. For a given kernel function and C, we determine the value of γ for which the maximum classification accuracy is achieved.

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	$\exp(-\gamma \mid x - x_i \mid^2)$
Laplace	$\exp(-\gamma \mid x - x_i \mid)$
Cauchy	$(1 \big/ (1 + \gamma \mid x - x_i \mid^2))$
Hypersecant	$2 \big/ (\exp(\gamma \mid x - x_i \mid) + \exp(-\gamma \mid x - x_i \mid))$
Square sync	$\sin^2 (\gamma \mid x - x_i \mid) \big/ (\gamma \mid x - x_i \mid)^2$

Example – 1: A dataset (Glass) having 06 classes has been taken to build the OvO ODT-SVM from the open literature for evaluation, optimal SVM Model selection at every root node of decision tree to get its hierarchical structure based on six criteria namely IG, Gini_index, chi-squared, CB, KD and SC as described in section III.

As in OvO ODT-SVM, $\frac{k(k-1)}{2}$ independent binary SVM's are constructed for k-class problem. Therefore for the glass dataset at beginning there is set of 15 SVM models from which one has to choose one optimal SVM model. To illustrate the DBA OvO ODT-SVM approach , the estimated and optimal values of the six criteria of these 15 SVM models are given in Table 3 for gauss kernel with C=1000 and $\gamma = [2^{-11}]$.

TABLE 3

DATABASE FOR ESTIMATED AND OPTIMAL VALUES OF ATTRIBUTES FOR EACH ALTERNATE SVMS FOR GLASS

Model	Gini_index	SC	Chi-squared	KD	CB	IG
m12	0.1385	0.055	36.294	0.0294	2.3701	0.1459
m13	0.2531	0.9867	0.7400	0.2001	15.6782	0.9204
m14	0.4136	0.27545	84.2398	0.3023	10.5684	0.6041
m15	0.2000	1.1361	7.1000	0.0000	0.0000	0.3521
m16	0.3107	0.0488	0.0000	0.0070	2.3807	0.4896
m23	0.1662	0.5293	55.2873	0.2194	5.662	0.2454
m24	0.3588	1.086	84.2702	0.0096	5.8591	0.4892
m25	0.0978	0.5624	39.4628	0.0000	0.0000	0.1586
m26	0.0296	0.0706	14.1100	0.0462	4.0166	0.4725
m34	0.4178	0.5686	32.6122	1.8780	8.4256	0.6086
m35	0.4875	0.4643	19.000	0.0000	0.0000	0.6806
m36	0.4882	1.0409	26.0000	0.2695	17.5684	0.6813
m45	0.3599	0.2914	29.0598	0.0000	0.0000	0.5456
m46	0.4640	2.5888	36.9591	0.3025	13.4647	0.6567
m56	0.4537	1.1402	23.0000	0.0000	0.0000	0.6461

From the comparison of rankings of the fifteen SVMs based on the values of all these six criteria as given in Table 3, it is observed that the ranking of the SVMs varies with respect to the criterion for selection. In order to avoid this problem it is proposed to apply DBA to rank the SVMs based on all these six criteria taken collectively. In the present method each criterion is considered as an individual selection attribute for the evaluation and comparison of SVMs. The matrix $[r_a]$ can represent the adjusted matrix of the process with the attribute values given above. Note the best numerical value of some criteria is smaller than that of the worst level. To avoid confusion and difficulties in performing the analysis, those values have been adjusted using following case:

Case - I: When bigger value of the attribute represents fitting well to the actual data i.e. is the best value:

Attribute Adjusted Value = Attribute Value - Attribute Minimum value in the database ($r_a = r_i - r_{\min}$).

The adjusted matrix thus obtained is shown as Matrix ($[r_a]$). From Eq. (27), the average values of the attributes are 0.280, 0.67437, 0.426, 0.2185, 0.733 and 0.367. The standard deviation of each attribute obtained using Eq. (28) is 0.152, 0.6542, 6.300, 0.475, 6.090 and 0.214 respectively.

Finally the Euclidean composite distance, CD, between each alternative SVMs to the optimal state, OPTIMAL, is derived from Eq. 31. Table 4 shows the composite distance value and the ranking of the alternate SVMs based on the contributing criteria. The overall ranking is based on composite distance value of each of the alternate SVM that is determined considering all six contributing attributes together using DBA. The alternate SVM with highest composite distance value is given rank no. 01 that with second lowest composite distance value is given rank no. 02, and so on. The results, so obtained, depict that the M12 model is ranked at number one based on the analysis using six criteria. Hence SVM model M12 is chosen as the root node of the decision tree. Now the left subtree will contain a set of classes {1, 3, 4, 5, 6} and right subtree will contain a set of classes {2, 3, 4, 5, 6}.

The same step wise procedure has been followed for ranking of SVMs OvO models generated on both side i.e. left subtree and right subtree of the decision tree. This process is recursively repeated until one is left with a set of two classes a root node of subtrees of the decision tree.

TABLE 4
SVMS RANKING BASED ON DBA FOR GLASS DATASET

Model Name	Sum	Composite Distance (CD) Value	Rank
M12	58.0808	7.621079	1
M13	21.0907	4.592466	13
M14	27.2352	5.218733	11
M15	48.124	6.937144	5
M16	52.466	7.243345	4
M23	41.5658	6.447151	8
M24	29.2114	5.404754	10
M25	55.7064	7.46367	2
M26	55.2048	7.429991	3
M34	17.9834	4.24068	14
M35	41.8872	6.472029	7
M36	23.2047	4.817129	12
M45	44.447	6.666859	6
M46	16.2219	4.02764	15
M56	35.9553	5.996277	9

The inorder traversal of DBA OvO ODT-SVM with M12 as root node as follows:
{1, M13, 3, M14, 3, M34, 4, M16, 3, M36, 6, M34, 4, M46, 6, M15, 3, M36, 6, M34, M45, 3, M36, 6, M35, 5, M56, 6, M12, 2, M24, 4, M23, 3, M34, 4 M25, 3, M34, 4, M45, 3, M35, 5, 3, M36, 6, M34,4, M46,6, M45, 3, M36, M35, 5, M56, 6.}

In this way an optimal classifier is generated for a particular choice of kernel function with its corresponding given parameters C and γ. For every kernel function we have generated 12 OvO ODT-SVM model and the model with best classification accuracy is reported in Table 5. To analyze the proposed DBA OvO ODT-SVM approach it is compared with other single criterion based OvO ODT-SVMs methods proposed in literature[11].

VII. CONCLUSION

This paper addresses the issue of optimal selection of SVM models at the root node of OvO ODT-SVM based on a number of conflicting criteria taken all together. The decision has unrestricted choices in exploring the influences of various different set of model selection criteria to final decision. As soon as a complete set of criteria for SVMs selection, along with the set of alternative SVMs and their level of criteria are formulized, and efficient rationalization process around multi-attribute decision model DBA can be performed. It is well established that no SVM model is optimal for all contributing criteria. This model allows a decision maker to perform, not just a general analysis, but also other various focused analyses regarding his or her personal preferences. The distance based approach method uses a relatively simple mathematical formulation and straight forward matrix operation; it is capable of solving complex multi-attributes decision problems, incorporating both quantitative and qualitative factors.

REFERENCES

[1] Thorsten Joachim's, N. Cristianini and J. Shawe Taylor, "Composite Kernels for Hypertext categorization", Proceedings of the International Conference on Machine Learning, 2001.
[2] S. Mori, C. Y. Suen and K. Yamamota, "Historical review of OCR research and development", Proceedings of the IEEE, vol. 80, pp. 1029-1058, 1992.
[3] M. Schmidt, "Identifying speaker with support vector networks", In Interface '96 Proceedings, Sydney, 1996.
[4] J. Weston and C. Watkins, "Multiclass Support Vector Machines", presented at the Proc. ESANN99, M. Verleysen, Ed., Brussels, Belgium, 1999.
[5] C. Corts and V.N. Vapnik, "Support Vector Networks", Machine Learning, Vol. 20, pp. 273-297, 1995.
[6] V. N. Vapnik, Statistical Learning Theory. New York: John Wiley & Sons, 1998.
[7] R. Rifkin and A. Klautau, "In Defence of One-Vs.-All Classification", Journal of Machine Learning, vol. 5, pp. 101-141, 2004.
[8] C. W. Hsu and C. J. Lin, "A comparison of methods for Multiclass Support vector machine", IEEE Transactions on Neural Networks, Vol. 13 (2), pp. 415-425, 2002.
[9] F. Takahashi and S. Abe, "Decision-tree-based multiclass support vector machines", Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP '02), volume 3, pages 1418–22, Singapore, 2002.
[10] Platt, N. Cristianini, J. Shawe-Taylor. Large margin DAGSVM's for multiclass classification. *Advances* in Neural Information Processing System. Vol. 12, pp. 547–553, 2000.
[11] Gjorgji, M.; Dejan G. & Ivan C. A Multi-class SVM Classifier Utilizing Binary Decision Tree. Informatica, 2009, 33, 233-241
[12] Shannon C. E., "A Mathematical Theory of Communication", Bell System Tech. J., Volume 27, pp. 379-423, 623-659, 1948.
[13] Quinlan J. R., "Introduction of Decision Trees', Machine Learning, Volume 1, pp. 81-106.
[14] L. Breiman, J. Friedman, R. Ohlsen and C. Stone, "Classification and regression trees", Wadsworth, Belmont , CA, 1984.
[15] R. Duda and P. Hart, "Pattern classification and scene analysis", J. Wiley, New York, 1973.
[16] Devijver P A and Kittler J, "Pattern Recognition: A statistical Approach," Prentice Hal1, 1982.