Identifying Efficient Kernel Function in Multiclass Support Vector Machines

R.Sangeetha
Ph.D Research Scholar
Department of Computer Science
Avinashilingam Deemed University for Women

Dr.B.Kalpana
Associate Professor
Department of Computer Science
Avinashilingam Deemed University for Women

ABSTRACT

Support vector machine (SVM) is a kernel based novel pattern classification method that is significant in many areas like data mining and machine learning. A unique strength is the use of kernel function to map the data into a higher dimensional feature space. In training SVM, kernels and its parameters have very vital role for classification accuracy. Therefore, a suitable kernel design and its parameters should be used for SVM training. In this paper, we present certain kernel functions for multiclass support vector machines and propose the appropriate and optimal kernel for one-versus-one (OAO) and one-versus-all (OAA) multiclass support vector machines. The performance of the one-versus-one and one-versus-all multiclass SVM are illustrated by empirical results and it is evaluated by the parameters like support vectors, support vector percentage, classification error, training error and CPU time. The experimental results demonstrate the ability to use more generalized kernel function and it goes to prove that the polynomial kernel's efficiency in terms of high classification accuracy for several datasets.

General Terms

Pattern Classification, Data Mining, Machine Learning

Keywords

Support Vector Machine, Multiclass Classification, Kernel function, One versus One, One versus All.

1. INTRODUCTION

Classification and Prediction [1, 2] are thriving research problems in machine learning and data mining .Support vector machine [5,6,7], a new computational learning method based on Vapnik–Chervonenkis theory [3, 4] to solve multidimensional function estimation. Basically, SVM classifier is developed for binary classification and later on it is extended to multiclass support vector machine which is the flattering topic in the field of research. In multiclass SVM, multiclass labels are decomposed into several two class labels and it trains a svm classifier to solve the problems and then reconstruct the solution of the multiclass problem from outputs of the classifiers [9], such as OAO-SVM and OAA-SVM.

Classification time and Computational complexity for the multiclass SVM classifier depend on the number of support vectors required for the multiclass SVM. Number of support vector increases, it leads to increase in computational requirements such as addition, multiplication and floating point. In SVM classification, the required memory to store the support vectors is directly proportional to the number of support vectors. Hence, support vectors must be reduced to speed up the classification and to minimize the computational and hardware resources required for classification.

The paper is organized as follows. Section 2 explains the multiclass support vector machines. Section 3 describes the kernels and its parameters. Section 4 comprises the experimental results. Lastly, Section 5 concludes with future work on achievable prospects in this area..

2. SUPPORT VECTOR MACHINES

SVM is based on the structural risk minimization principle (SRM), which was proposed by Vapnik [3] and its generalization is optimal. Initially, SVM is developed for binary classification and later it extended multiclass classification. Its core concept is to derive the hyperplane to separate the two classes. Consider a set of training examples (x_i, y_i), i = 1,…. l, $x_i \in R^n$, $y_i \in \{+1, -1\}$, and try to train a function $y_i = f(x_i)$ that predicts the classification y_i of unknown data x_i, with minimal error. Support vector machines use a function ψ to map data into a higher dimensional space and construct a separating hyperplanes in feature space. One of the hyperplanes that maximizes the margin is an optimal separating hyperplane. Binary classification and its kernel selection are explained in [12, 13].

2.1 Multiclass Support Vector Machines [15]

Many real world problems like circuit diagnosis, natural language processing come under the category of multiclass support vector machines. Multiclass SVM can be solved by combining the binary classification decision functions. Multiclass SVM is of two types namely, ***One versus One*** decomposition and ***One versus All*** decomposition.

The OAA decomposition [10] transforms the multiclass problem into a series of *c* binary subtasks that can be trained by the binary SVM. Let the training set $T_{XY}^{y} = \{(x_1, y_1^{'}),....,(x_l, y_{l\}}^{'})$ contain the modified hidden states defined as

$$y_i^{'} = \begin{cases} 1 & for \quad y = y_i \,, \\ 2 & for \quad y \neq y_i \end{cases} \qquad (1)$$

The discriminant functions

$$f_y(x) = \langle \alpha_y \bullet K_s(x) \rangle + b_y \,, \quad y \in Y \,, \qquad (2)$$

are trained by the binary SVM solver from the set T_{XY}^{y} , $y \in Y$

The OAO decomposition [10] transforms the multi-class problem into a series of $g = c(c - 1)/2$ binary subtasks that can be trained by the binary SVM. Let the training set $T_{XY}^{y} = \{(x_{1}^{'}, y_{1}^{'}),...., (x_{l_j}^{'}, y_{l_j}^{'})$ contain the training vectors $\boldsymbol{x}_i \in I^j=\{i: y_i = y^1 \vee y_i = y^2\}$ and the modified the hidden states defined as

$$y_i^{'} = \begin{cases} 1 \;\; for \;\; y_j^1 = y_i \,, \\ 2 \;\; for \;\; y_j^2 \neq y_i \,, \end{cases} \quad i \in I^j \qquad (3)$$

The training set T_{XY}^{j} , $j = 1,2,...g$ is constructed for all $g=c(c-1)/2$ combinations of classes $y_j^1 \in Y \;\&\; y_j^2 \in Y \setminus \{y_j^1\}$. The binary SVM rules q_j , $j = 1, . . . , g$ are trained on the data T_{XY}^{j} .

Table 1. Types of Kernels

Kernels	Function
Laplacian	$K(x,y) = \exp\left(-\frac{\lVert x-y \rVert}{\sigma}\right)$
Rational Quadratic	$K(x,y) = 1 - \frac{\lVert x-y \rVert^2}{\lVert x-y \rVert^2 + c}$
Multiquadratic	$k(x,y) = \sqrt{\lVert x-y \rVert^2 + c}$
Wave	$K(x,y_j) = \frac{\theta}{\lVert x-y \rVert} \sin \frac{\lVert x-y \rVert}{\theta}$
Power	$K(x,y) = -\lVert x-y \rVert^d$
Log	$K(x,y) = -\log \lVert x-y \rVert^d + 1)$
Bessel	$K(x,y) = \frac{J_{V+1}(\sigma \lVert x-y \rVert)}{\lVert x-y \rVert^{-n(v+1)}}$
Cauchy	$K(x,y) = \frac{1}{1 + \frac{\lVert x-y \rVert^d}{d}}$
Wavelet	$K(x,y) = \prod_{m=1}^{N} h\left(\frac{x_i - c}{a}\right) h\left(\frac{y_i - c}{a}\right)$

3. KERNELS IN MULTICLASS SVM

Support vector machine is the first kernel based learning algorithm. Kernel function determines the characteristics of OAO and OAA SVM model and level of non linearity. A necessary and sufficient condition for a simple inner product kernel to be valid is that it must satisfy Mercer's theorem [11]. In general, kernels are of two types namely *Local* and *Global* kernels. Data that are close to each other in local kernels influence on the kernel points and data that are far away from each other in global kernels influence on the kernel points. Commonly used kernels like Linear, Polynomial, RBF, Sigmoid are discussed in [12, 13] and used in this paper. Also, there are some more kernels which are represented in Table 1.Kernel functions and its transformation largely depends on the domain .So, selecting a suitable kernel function with its parameter is a major research area in multiclass support vector machine.

Table 2. Data Sets Used

Data Sets	Size	Features	Class Distribution
Pentagon	99	2	5
Iris	150	4	3
Wine	270	13	3

4. EXPERIMENTAL RESULTS

In this Section, we evaluate the performance of one-versus-one and one-versus-all multiclass SVM using different kernel functions on two benchmark datasets(Iris, Wine) taken from UCI machine learning repository and a pentagon dataset taken from [10].Brief outline of the datasets is given in table 2.Kernels are evaluated using the performance metrics like support vectors, support vector percentage, training error, classification error and CPU.Here,five fold cross validation is used to split the training dataset and test dataset. For each method in multiclass SVM, the optimal regularization parameter *C* and the kernel parameters are estimated by repeating classifications. The classification accuracy of SVM methods based on classification error and training error with these optimal parameters are compared. The tables 3.1, 3.2, 3.3 show the results for 3-class, 5-class OAA and OAO SVM. In that only few kernels give good classification performance with low classification error.

Linear kernel $K(x_i,x_j) = 1 + x_i^T x_j$ is a simple kernel function based on the penalty parameter ***C***, since parameter *C* controls the trade-off between frequency of error c and complexity of decision rule [7]. Also, it reduces the support vectors, training error and classification error by incrementing the parameter C.But it is not suitable for large datasets.

Polynomial kernel $K(x_i,x_j) = (1 + x_i^T x_j)^p$ also known as ***global kernel***, is non-stochastic kernel estimate with two parameters i.e. ***C*** and polynomial degree ***p***. Each data from the set $\mathbf{x_i}$ has an influence on the kernel point of the test value $\mathbf{x}_j$, irrespective of its the actual distance from $\mathbf{x}_j$ [14], It gives good classification accuracy with minimum number of support vectors and low classification error.

Radial basis function $K(x_i,x_j) = \exp(-\gamma \lVert x_i - x_j \rVert^2)$ also known as ***local kernel***, is equivalent to transforming the data into an infinite dimensional Hilbert space .Thus, it can easily solve the non-linear classification problem. It has an effect on the data points in the neighborhood of the test value [14]. RBF gives similar result as polynomial with minimum training error but for some cases the number of support vector and classification error increases.

Exponential radial basis function $K(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|}{2\sigma^2})$ gives piecewise linear solution. ***Gaussian radial basis function*** $K(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$ deals with data that has conditional probability distribution approaching gaussian function. ***RBF kernels*** perform better than the linear and polynomial kernel. However, it is difficult to find an optimum parameters $\boldsymbol{\sigma}$ and equivalent ***C*** that gives better result for a given problem. ***Sigmoid kernel*** $K(x_i,x_j)=\tanh(kx_i^T x_j-\delta)$ is not efficient as other kernel function, because it lacks the necessary condition of a valid kernel. Parameters κ and δ must be chosen properly to obtain high classification accuracy.

Metrics of the Kernels based on the parameter values in Table 3.1, 3.2 and 3c are graphically portrayed in the Annexure I to analyze the data. In kernel function, number of support vector increases then the classification accuracy diminishes. Figures [1-6] represent the support vector, support vector % for OAO and OAA SVMs. Figures [7-12] symbolize Train error and test error for OAO and OAA SVMs. After analyzing all the features of the kernel function using figures [1-12], appropriate and optimal kernels for our datasets are polynomial kernel, RBF kernel .They have minimum number of support vectors, minimum value as classification error and training error and good classification accuracy.

5. CONCLUSION AND FUTURE WORK

From the empirical results, we present the performance of multiclass SVM using different kernels on three different datasets and a comparison is made. Here, we are attempted to explore the best choice among SVM kernels namely linear, polynomial, radial basis function (RBF) and sigmoid kernels . Different degree of the polynomial kernels and different widths of the RBF kernel are evaluated. As a result, the efficient kernel for multiclass SVM classifier is polynomial kernel for these datasets.

In multiclass SVM, OAO / OAA model's quality is ascertained by its ability to learn from the data and to predict unknown data i.e. ***learning capacity*** and ***generalization ability***. These two important characteristics of SVM are determined by an optimal and efficient kernel function and its parameter selection. We have to select a kernel function that should satisfy both these properties. Generalization ability gets better for certain polynomial degrees and radial basis with suitable parameter selection gives good learning capacity. Single parameter value of kernel cannot decide these two properties. So, two kernel functions with these characteristics are selected and checked whether the merits of two kernels can be combined to form a hybrid kernel. High classification accuracy can be achieved by optimizing the kernel function and tuning its kernel parameters.

6. REFERENCES

[1] Han,J., Kamber,M.2006. Data Mining—Concepts and Technique, 2nd ed. San Mateo, CA: Morgan Kaufmann.

[2] Tan,P.N.,Steinbach,M. and Kumar,V.2005. Introduction to Data Mining. Reading, MA: Addison-Wesley.

[3] Vapnik, V. 1999. An overview of statistical learning theory. IEEE Trans. on Neural Networks.

[4] Cristianini, N. and Shawe-Taylor, J. 2000. Introduction to Support Vector Machines. Cambridge University Press.

[5] Schölkopf, B. and Smola, A.2001. Leaning with Kernels. MIT Press.

[6] Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 56–89.

[7] Corinna Cortes and Vapnik, V. 1995. Support-Vector Networks, Machine Learning.

[8] Manikandan , J. and Venkataramani, B. 2010.Study and evaluation of a multi-class SVM classifier using diminishing learning technique, Neurocomputing , doi:10.1016/j.neucom.2009.11.042

[9] Anna Wang, Wenjing Yuan, Junfang Liu, Zhiguo Yu, Hua Li, 2009. A novel pattern recognition algorithm: Combining ART network with SVM to reconstruct a multi-class classifier, Computers and Mathematics with Applications, 1908_1914.

[10] Vojtech Franc, Václav Hlavá. 2009. Statistical Pattern Recognition Toolbox for Matlab.

[11] Ralf Herbrich December 2001.Learning kernel classifiers: theory and algorithms, MIT Press, Cambridge, Mass, ISBN 026208306X.

[12] Sangeetha, R., Kalpana, B. 2010. A comparative study and choice of an appropriate kernel for support vector machines. In: Das, V.V., Vijaykumar, R. (eds.) ICT 2010. CCIS, vol. 101,pp. 549–553. Springer, Heidelberg (2010)

[13] Sangeetha, R., Kalpana, B.2010. Optimizing the Kernel Selection for Support Vector Machines using Performance Measures. In: A2CWiC 2010, ISBN: 978-1-4503-0194-7.

[14] Smits ,G.F. and Jordaan, E.M. 2002. Improved SVM Regression using Mixtures of Kernels, IJCNN '02. Proceedings of the International Joint Conference on Neural Networks.

[15] Weston, J. and Watkins, C. Multi class support vector machines, Technical Report.

Appendix

Table 3.1 Iris Dataset

Kernels	Parameter	One versus One					Parameter	One versus All				
		SV	SV%	TE	CE	CPU (S)		SV	SV%	TE	CE	CPU (s)
Linear	C=10	16	13.33	0.0167	0.5	0.03	C=10	70	58.33	0.0583	0.3333	0.14
	C=100	11	7.5	0.0	0.6333	0.01	C=10000	63	52.5	0.0167	0.3000	28
Polynomial	C=1, p = 1.5	23	19.1	0.0167	0.3333	0.04	C=10,p=2	15	12.5	0.08	0.1	0.34
	C=1,p=2.5	16	13.33	0.1416	0.4333	0.125	C=100,p=2	10	8.3	0.0	0.2	0.09
RBF	C=1, γ = 0.5	40	33.33	0.058	0.2667	0.03	C=10, γ =1.5	20	16.67	0.033	0.0667	0.04
	C=1, γ =1.5	31	25.8	0.025	0.5333	0.05	C=10, γ =1	23	19.1	0.025	0.1333	0.04
ERBF	C=1,σ=1.5	47	39	0.0167	0.0333	0.031	C=1, σ =0.5	31	25.8	0.008	0.0667	0.015
	C=10, σ =2.5	28	23.33	0.0167	0.2667	0.03	C=100, σ =2	17	14.16	0.1667	0.1	0.06
GRBF	C=10, σ =2	31	25.8	0.025	0.1333	0.03	C=10, σ =0.05	45	37.5	0.008	0.2333	0.12
	C=10, σ =1.5	26	21.6	0.0167	0.4	0.03	C=100, σ =0.05	44	36.67	0.008	0.2	0.28
Sigmoid	C=1, k=1,δ=2	46	38.3	0.0583	0.2667	0.063	C=1000,k=1,δ =3	12	10	0.0	0.2	0.218
	C=1000,k =5,δ=2	40	33.33	0.3083	0.0333	0.016	C=1000,k=2,δ =5	11	9.16	0.0	0.2667	0.313

Table 3.2 Pentagon Dataset

Kernels	Parameter	One versus One					Parameter	One versus All				
		SV	SV%	TE	CE	CPU (S)		SV	SV%	TE	CE	CPU (s)
Linear	C=10	25	31.65	0.0	0.25	0.015	C=10	50	63.3	0.013	0.05	0.109
	C=100	20	25.32	0.0	0.25	0.03	C=1000	40	50.63	0.0	0.05	0.078
Polynomial	C=1000,p=3	18	22.78	0.0	0.25	0.0	C=100,p=1.5	19	24.05	0.0	0.1	0.1
	C=1000,p=6	15	18.98	0.0	0.25	0.031	C=1000,p=1.5	17	21.51	0.0	0.1	0.171
RBF	C=10, γ = 0.005	15	18.98	0.0	0.4	0.0	C=100, γ =0 .5	43	54.43	0.367	0.3	0.156
	C=100, γ =0.5	20	25.32	0.0	0.3	0.12	C=100, γ =6	21	26.58	0.025	0.1	0.09
ERBF	C=100, σ =1.5	21	26.58	0.0	0.25	0.02	C=100, σ =0.5	30	37.97	0.0	0.05	0.046
	C=1000, σ =0.5	28	35.44	0.02	0.8	0.016	C=inf, σ =2	19	24.05	0.0	0.1	0.109
GRBF	C=100, σ =0.05	38	48.1	0.0	0.5	0.031	C=10, σ =0.5	32	40.5	0.101	0.45	0.031
	C=1000, σ =2	18	22.78	0.04	0.3	0.015	C=inf, σ =0.5	17	21.51	0.316	0.25	0.031
Sigmoid	C=10, k=1, δ =2	28	35.44	0.01	0.3	0.0	C=100,k=1,δ=1	20	25.32	0.0	0.1	0.046
	C=100,k=0.5,δ=1	20	25.32	0.03	0.25	0.031	C=inf,k=2,δ=0.5	19	24.05	0.0	0.1	0.187

Table 3.2 Wine Dataset

Kernels	Parameter	One versus One					Parameter	One versus All				
		SV	SV%	TE	CE	CPU (S)		SV	SV%	TE	CE	CPU (s)
Linear	C=1	33	22.91	0.0625	0.9412	85.12	C=10	35	24.3	0.528	0.8542	121
	C=100	31	21.52	0.0694	0.9118	121.1	C=100	39	27.08	0.253	0.8574	114
Polynomial	C=10,p=0.5	44	30.55	0.1458	0.1471	0.687	C=10,p=2	8	5.55	0.319	0.0882	0.156
	C=100,p=0.25	45	31.25	0.2656	0.1471	0.153	C=100,p=2	8	5.55	0.319	0.0882	0.171
RBF	C=100, γ =0.0005	68	47.22	0.0486	0.4118	0.703	C=100,γ =0.00005	65	45.13	0.09	0.6765	2.359
	C=100, γ =0.05	75	52.08	0.0069	0.0588	0.859	C=1000,γ =0.00005	55	38.19	0.09	0.6765	62
ERBF	C=100, σ =8	70	48.6	0.0277	0.5588	0.812	C=100, σ =6	78	54.16	0.006	0.705	9.875
	C=100, σ =2.5	85	59.02	0.0138	0.2647	0.328	C=100, σ =10	72	50	0.006	0.6765	11.26
GRBF	C=1000, σ =8	80	55.55	0.0277	0.1765	0.593	C=1000, σ =8	90	62.5	0.013	0.8529	2.234
	C=1000, σ =6	80	55.55	0.0208	0.0882	48	C=1000, σ =6	100	69.44	0.0	0.8876	1.187
Sigmoid	C=100, k=2, δ =4	112	77.77	0.9	0.5902	0.562	C=100, =2,δ=4	122	84.72	0.58	0.8532	1.234
	C=100, k=2,δ =2	110	76.38	0.91	0.5902	0.531	C=100 k=2,δ=2	122	84.72	0.59	0.8532	1.25

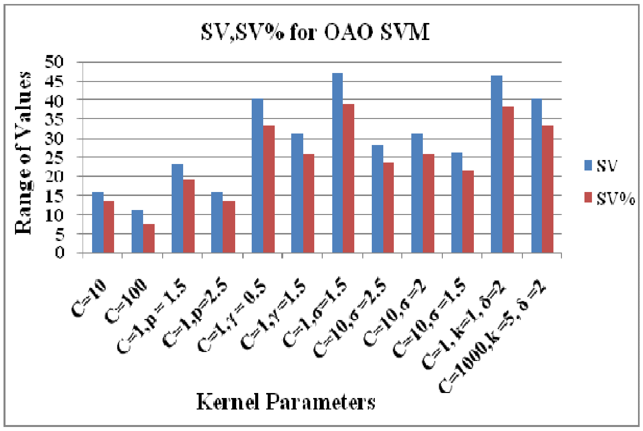

Fig.1: OAO -Support Vectors with its % for Iris Dataset

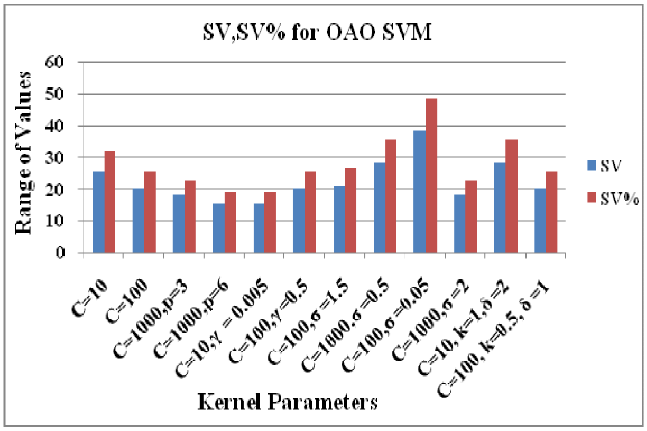

Fig.2: OAO - Support Vectors with its % for Pentagon Dataset

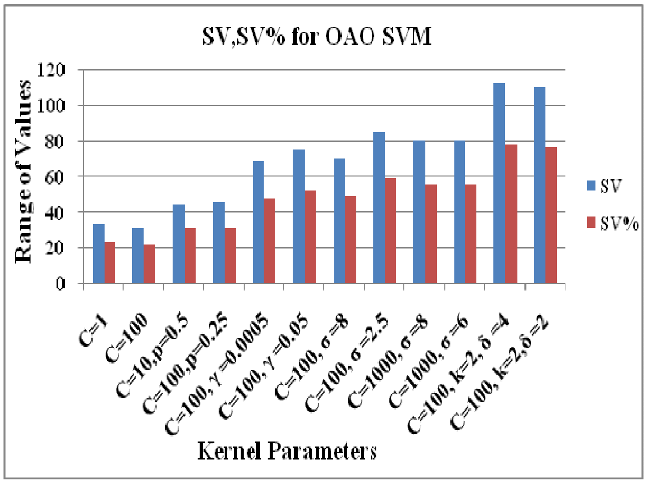

Fig.3: OAO - Support Vectors with its % for Wine Dataset

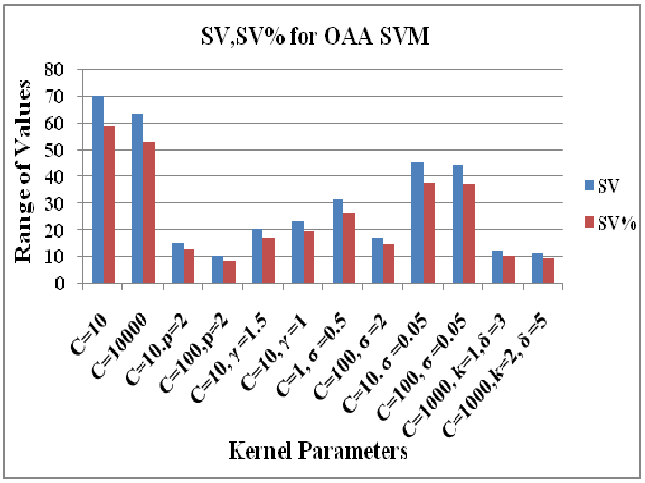

Fig.4: OAA - Support Vectors with its % for Iris Dataset

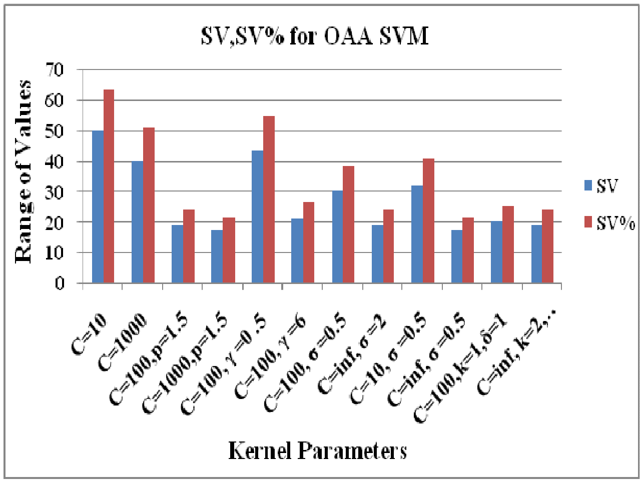

Fig.5: OAA - Support Vectors with its % for Pentagon Dataset

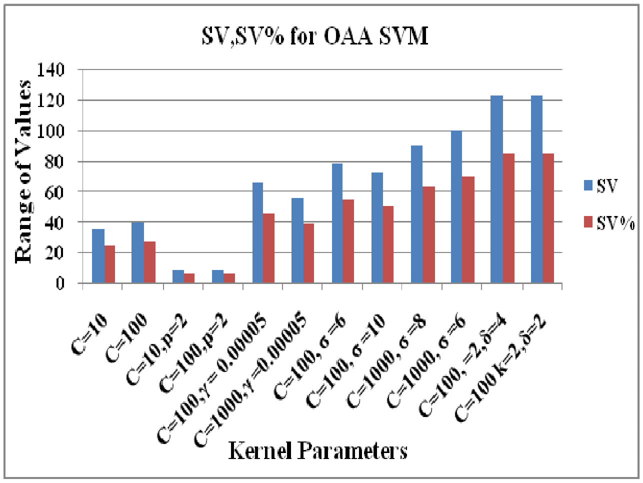

Fig.6: OAA - Support Vectors with its % for Wine Dataset

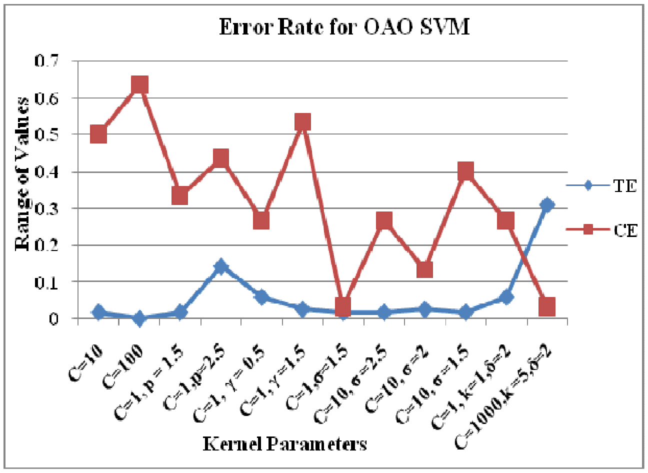

Fig.7: OAA – Error Rate for Iris Dataset

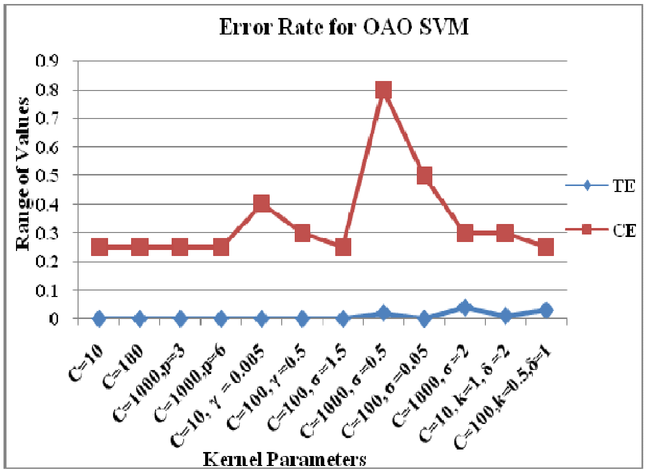

Fig.8: OAA - Error Rate for Pentagon Dataset

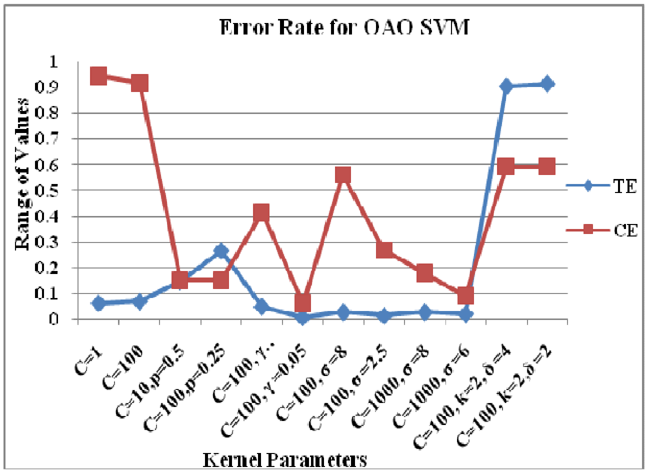

Fig.9: OAA - Error Rate for Wine Dataset

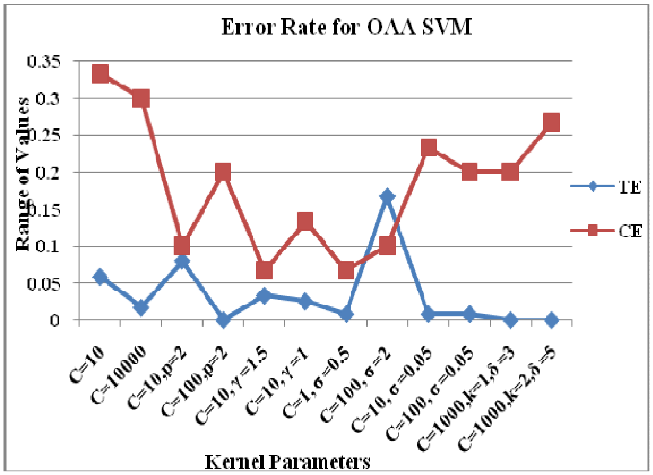

Fig.10: OAA - Error Rate for Iris Dataset

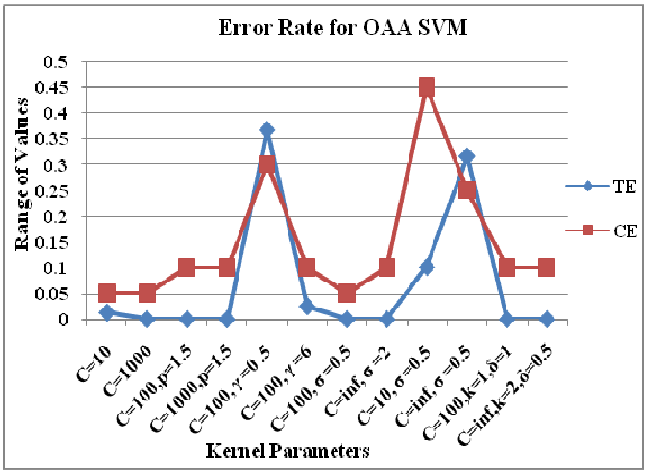

Fig.11: OAA - Error Rate for Pentagon Dataset

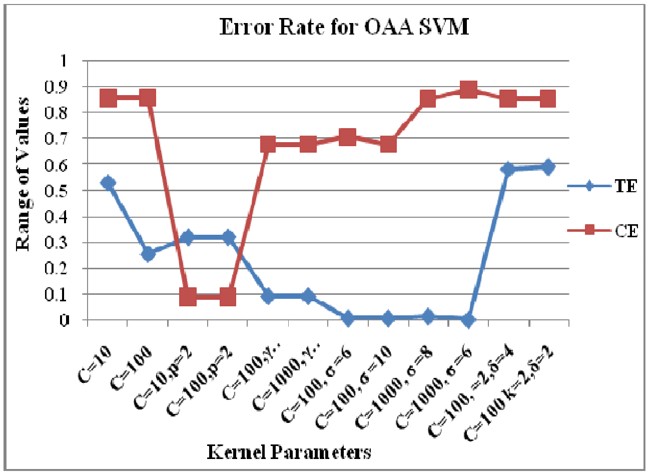

Fig.12: OAA - Error Rate for Wine Dataset