Iraqi Journal of Statistical Sciences (22)2012
PP[26-40]

Support Vector Machine Procedure as a Data Mining Multi-class Classifier

Zakariya Y. Algamal[•]

Abstract

Support vector machine initially developed to perform binary classification. This paper presents a multi-class support vector machine classifier and ordinal regression to classify the type of bone mineral density. This paper compares the performance of four multi-class approaches, one-against-all, one-against-one, Weston and Watkins, and Crammer and Singer. Results from our real life data conclude that Crammer and Singer may be better approach depending on training error and the percentage of correctly classified test data. Also, we find that the training error becomes more less when the regulization parameter **C** and kernel parameter σ become large.

Keywords: Data mining support vector machine, quadratic programming, multi-class classification, ordinal regression.

أسلوب آلة المتجه الداعم بوصفه مصنفا متعدد الحالات في تنقيب البيانات

المستخلص

يستخدم اسلوب الة المتجه الداعم للتصنيف الثنائي عندما يكون متغير الاستجابة ذا صفتين. يهدف هذا البحث الى تقديم مفهوم اسلوب الة المتجه الداعم للتصنيف المتعدد الحالات وكذلك الانحدار الترتيبي لتصنيف نوع كثافة العظم. يقوم هذا البحث بمقارنة اربعة اساليب من اساليب المتجه الداعم للتصنيف المتعدد الحالات OAA, OAO, WW, و CS . وقد أظهرت النتائج

[•] Lecturer/ Dept. of Statistics and Informatics / College of Computer Science and Mathematics / University of Mosul.

Received Date 2/4/2012 Accept Date 21/5/2012

التي حصلنا عليها بان اسلوب CS يكون افضل اسلوب تصنيفي متعدد الحالات بالاعتماد على خطأ التدريب والنسبة المئوية للتصنيف الصحيح لمجموعة الاختبار ، كذلك وجدنا ان خطأ التدريب يصبح اقل كلما ازدادت قيمة المعلمة **C** و المعلمة σ.

1- Introduction

The amount of data in the world and in our lives seems ever increasing and there is no end in sight. Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Classification, one of the most common data mining tasks, is built to predict a categorical variable. Support vector machine (SVM) is a powerful data mining technique for classifying data. The SVM is a training algorithm for learning classification and regression rules from data (Seeja and Shweta, 2011). SVM is a modern learning system designed by Vapnik and his colleagues (Vapnik, 2010). Based on statistical learning theory, which explains the learning process from a statistical point of view, the SVM classifier crates a hyperplane that separates the data into two categories with the maximum margin. Originally, the SVM was a linear classifier based on the optimal hyperplane algorithm (Hu and Pan, 2007). This paper is composed of seven sections, the second section addresses the theoretical aspects of SVM. Section 3 contains the theoretical information about multi-class SVM. Details on Ordinal regression was given in section 4. Section 5 contains the description and results of the our Real data. Finally, section 6 addresses the conclusions.

2- Support Vector Machine Procedure

SVM is a group of supervised learning methods that can be applied to classification and regression (Ivanciuc, 2007). SVM was originally designed for binary classification (i.e. we have two categories of the response variable). Support vector learning is based on simple ideas, the simplicity comes from the fact that SVM applies a simple linear method to the data but in a high-dimensional feature space in non-linearly related to the input space (Karatzoglou and Meyer, 2006). The commonly seen binary classification problem can be divided into two cases, linearly separable and linearly inseparable. The solution to the former is easy to obtain, but kernel functions have to be introduced to solve the problems in the latter case (Liang et al, 2011). Suppose there is a data set of two classes of samples, in which each sample is denoted by $\mathbf{x}_i$ with the corresponding class label $\mathbf{y}_i$, that is,

$$\mathbf{x}_i \in \mathbf{R}^n \quad , \ \mathbf{y}_i \in \{-1,1\}, \quad \mathbf{i} = 1,2,\ldots\ldots,\mathbf{n} \qquad \ldots\ldots.(1)$$

Here, $\mathbf{x}_i$ is an n-dimensional vector with corresponding $\mathbf{y}_i$ equal to 1 if it belongs to a positive class or -1 if negative. In the linearly separable case, any hyperplane $\mathbf{f(x)}$ (Decision function) should meet the condition:

$$\begin{aligned} \mathbf{f(x_i) = w'x_i + b \geq 1 \ \ if \ \ y_i = 1} \\ \mathbf{f(x_i) = w'x_i + b \leq -1 \ \ if \ \ y_i = -1} \end{aligned} \qquad \ldots\ldots..(2)$$

Where $\mathbf{w}$ is the normalized weight vector with the same dimension as $\mathbf{x}_i$ and $\mathbf{b}$ is the normalized bias of the hyperplane. It should be noticed that $\mathbf{w}$ and $\mathbf{b}$ make $\mathbf{f(x)}$ equal to 1 or -1 if $\mathbf{x}$ is on the boundary, then the margin, the distance from the separating hyperplane to its nearest sample, between the two paralleled hyperplanes can be written:

$$\mathbf{margin} = \mathbf{2}\,\frac{|\mathbf{w'x+b}|}{\|\mathbf{w}\|} = \frac{\mathbf{2}}{\|\mathbf{w}\|} \qquad \ldots\ldots.(3)$$

Figure (1) shows the SVM structure. The object of support vector classification machine is to locate the optimal separatable hyperplane (that can maximize the margin subject to the equation (2). Therefore, the optimal separatable hyperplane can be converted to the following optimizing problem:

$$\begin{aligned}&\mathbf{max\,imize:}\ \frac{\mathbf{2}}{\|\mathbf{w}\|}\\&\mathbf{subject\ to:(w'x_i+b)y_i\geq 1}\end{aligned} \qquad \text{……..(4)}$$

With the help of the Lagrange multiplier method, we rewrite equation (4) in minimized the objective function:

$$\mathbf{L(w,b,\alpha)=\frac{1}{2}w'w-\sum_{i=1}^{n}\alpha_i\left[y_i(w'x_i+b)-1\right]} \qquad \text{……(5)}$$

where $\alpha_i\,(\alpha_i \geq \mathbf{0})$ is called the Lagrange multiplier. Driving it against $\mathbf{w}$ and $\mathbf{b}$, we can obtain the following two equations:

$$\mathbf{\frac{\partial L(w,b,\alpha)}{\partial w}=w-\sum_{i=1}^{n}y_i\alpha_i x_i=0} \qquad \text{…..(6)}$$

$$\mathbf{\frac{\partial L(w,b,\alpha)}{\partial b}=\sum_{i=1}^{n}y_i\alpha_i=0} \qquad \text{……(7)}$$

Putting the solution of equations (6) and (7) into equation (5), one can get the dual form of equation (5):

$$\mathbf{L(w,b,\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}y_iy_j\alpha_i\alpha_jx_i'x_j} \qquad \text{……..(8)}$$

Minimizing equation (8) is a convex quadratic programming problem under constraints

$$\mathbf{\sum_{i=1}^{n}y_i\alpha_i=0\ \ and\ \ \alpha_i\geq 0} \qquad \text{…….(9)}$$

All values of α_i that corresponding to the few samples on the boundary are positive and the others are equal to zero. The sample with $\alpha_i > \mathbf{0}$ is called the support vector (Liang et al, 2011).

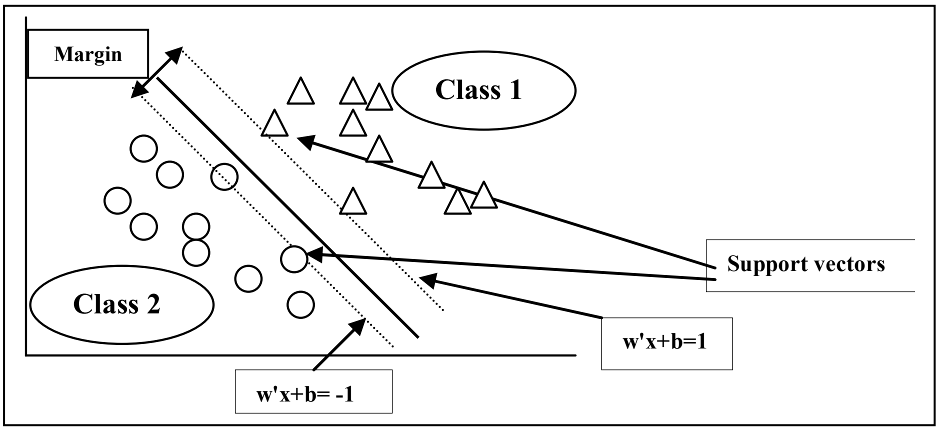

Figure (1): The Structure of SVM.

In many applications, a linearly inseparable case provides better accuracy. It uses feature functions $\boldsymbol{\phi(x)}$. The SVM extension to nonlinear data sets is based on mapping the input variable into a feature space of a higher dimension and then performing a linear classification in that higher dimensional space (Abe, 2010). By using the nonlinear vector function $\boldsymbol{\phi(x) = (\phi_1(x),..,\phi_\lambda(x))'}$ that maps the n-dimensional input vector $\mathbf{x}$ into the λ-dimensional feature space, the linear hyperplane $\mathbf{f(x)}$ in the feature space be

$$\mathbf{f(x) = w'\boldsymbol{\phi}(x) + b} \qquad ......(10)$$

Where $\mathbf{w}$ is a λ-dimensional vector (Abe, 2010). We use $\mathbf{k(x_i, x_j)}$ in training and classification instead of $\boldsymbol{\phi(x)}$ where it is called kernel function.

$$\mathbf{k(x_i, x_j) = \boldsymbol{\phi}(x_i)'\boldsymbol{\phi}(x_j) \ , \quad i, j = 1,2,......,n} \qquad ......(11)$$

Using equation (11), the minimization of the dual form in equation (8) in feature space becomes (Liang et al, 2011):

$$\mathbf{L(w,b,\boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j \, k(x_i, x_j)} \qquad ......(12)$$

The hyperplane can be calculated as:

$$f(x) = \text{sign}[\sum_{i}^{n} y_i \alpha_i k(x_i, x_j) + b] \qquad \ldots\ldots(13)$$

There are several popular kernel functions such as linear, polynomial, and radial basis.

3- Multi-class Support Vector Machine

Originally, SVM was developed to perform binary classification (i.e. two categories). However, applications of binary classification may be limited in many applications (Sangeetha and Kalpana, 2011). The SVM classifier has to be modified to work with multi-class classification problems (i.e. when there are three or more categories) (Monfrini and Guermeur, 2011). Many multi-class SVM (MSVM) classification approaches decompose the training data into several binary classes. Some like one-against-one, one-against-all. Others are depending on one single objective function like Weston and Watkins and Crammer and Singer. Figure (2) shows the basic structure of multi-class approaches.

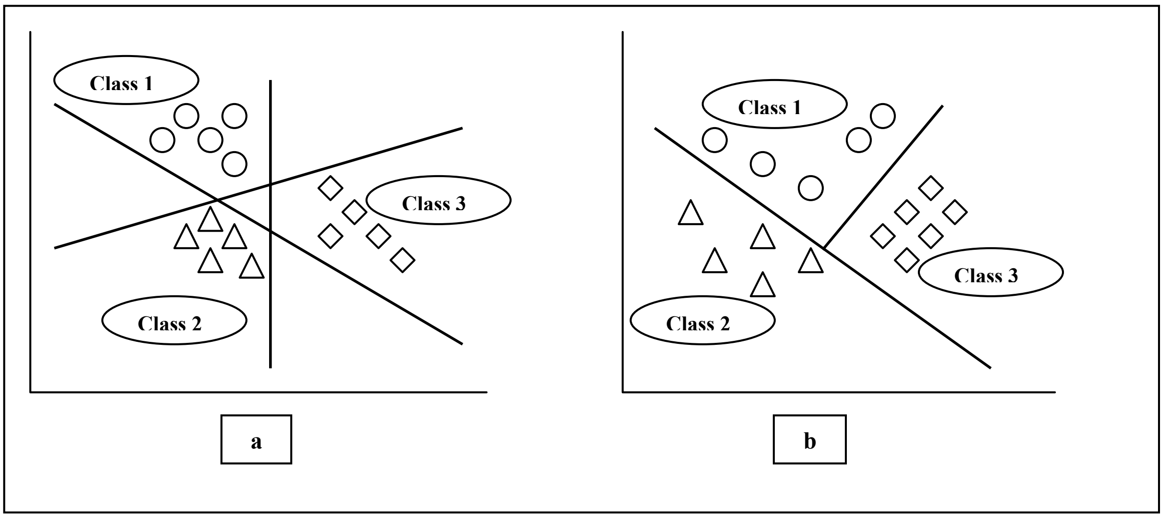

Figure (2):a- OAA and OAO Methods, b- WW and CS Methods

3-1 One-Against- All Approach

The one- against- all (OAA) support vector machine approach is the simplest of MSVM approaches. Assume that there are $\mathbf{k}$ classes (categories) that we want to classify. In this approach we construct $\mathbf{k}$ binary SVM classifiers to separate each class from the rest, like class 1 (positive) against all other classes (negative). Then the new objects are assigned to the class that has a positive vote and the largest distance to its hyperplane (Hsu and Lin, 2002), (Abe, 2010), and (Statnikov et al, 2011). Consider we have $\mathbf{k}$ - class, for a OAA approach we determine $\mathbf{k}$ direct decision functions that separate one class from the remaining class. So, there are $\mathbf{k}$ decision functions:

$$\begin{array}{l} (\mathbf{w}^{(1)})'\phi(\mathbf{x})+\mathbf{b}^{(1)} \\ (\mathbf{w}^{(2)})'\phi(\mathbf{x})+\mathbf{b}^{(2)} \\ \mathrm{M} \\ (\mathbf{w}^{(k)})'\phi(\mathbf{x}_1)+\mathbf{b}^{(k)} \end{array} \qquad ....(14)$$

and say $\mathbf{x}$ is in the class which has the largest value of the decision function,

class of $\mathbf{x} \equiv \arg\max_{i=1,2,..,k} [(\mathbf{w}^{(i)})'\phi(\mathbf{x})+\mathbf{b}^{(i)}]$ .....(15)

3-2 One-Against-One Approach

Another major method is called the one-against-one (OAO) approach. Here we construct binary SVM classifiers to separate each pair of classes: class 1 against class 2, class 1 against class 3, …., class $\mathbf{k-1}$ against $\mathbf{k}$. So, this approach constructs $\mathbf{k(k-1)/2}$ binary SVM classifiers, and then new objects are assigned to the class that has the majority of votes (Statnikov et al, 2011). Let the decision function for class $\mathbf{i}$ against class $\mathbf{j}$, with the maximum margin, be (Abe, 2010):

$$\mathbf{D_{ij}(x) = w'_{ij}\phi(x) + b_{ij}} \qquad \text{.........(16)}$$

where $\mathbf{w_{ij}}$ is the λ-dimensional vector and $\mathbf{D_{ij}(x) = -D_{ji}(x)}$. The regions

$$\mathbf{R_i = \{x | D_{ij}(x) > 0,\ j = 1,2,....,k,\ j \neq i\}\ for\ i = 1,2,...k} \qquad \text{.........(17)}$$

And if $\mathbf{x}$ is in $\mathbf{R_i}$, then $\mathbf{x}$ is considered to belong to class $\mathbf{i}$. The problem is that $\mathbf{x}$ may not be in any of $\mathbf{R_i}$. We classify $\mathbf{x}$ by voting. For the input vector $\mathbf{x}$ we calculate

$$\mathbf{D_i(x) = \sum_{i \neq j, j=1}^{k} sign(D_{ij}(x))} \qquad \text{..........(18)}$$

where

$$\mathbf{sign(x) = \begin{Bmatrix} 1 & for\ \ x \geq 0 \\ -1 & for\ \ x < 0 \end{Bmatrix}} \qquad \text{.........(19)}$$

And we classify $\mathbf{x}$ into the class

class of $\mathbf{x \equiv arg\,max_{i=1,2,..,k}\, D_i(x)}$ ....(20)

3-3 Weston and Watkins, and Crammer and Singer Approaches

The former approaches utilized binary SVM classifiers to make a multi-class classification. Other approaches make multi-class classification by considering all classes at once like the approach by Weston and Watkins (WW) (Weston and Watkins, 1999) and the one by Crammer and Singer (CS) (Crammer and Singer, 2000).These approaches use single optimization problem of size $\mathbf{(k-1)n}$ to obtain all weight vectors. They construct $\mathbf{k}$ two-class rules where the $\mathbf{m^{th}}$ function $\mathbf{w'_m\phi(x) + b}$ separates training vectors of the class $\mathbf{m}$ from the other vectors. Here there are $\mathbf{k}$ decision functions but all are obtained by solving one problem. Then the decision function is:

class of $\mathbf{x \equiv arg\,max_{m=1,2,..,k}[w_m \phi(x) + b_m]}$ ....(21)

4- Ordinal Regression Model

In many fields ordinal regression (OR) becomes the standard model for analyzing the effects of explanatory variables on multi-class response variable. In this, the response variables has ordered muli-class qualitative data (Normal, Osteopenia, and Osteporosis for our real data analysis). OR is a supervised learning of predicting categorical of ordinal scale, it lies somewhere between classification and regression. Many approaches have been developed to deal with ordinal regression, the one that we are interested in is to convert OR to a set of binary classification problems

(Xia et al, 2007). In binary classification usually we use the term "success" and its estimated probabilities for the response of interest, p(success), and the term "not success" with p(not success). In OR the term "success" can be conceived of many different ways. For our real data the number of class of the response variable be 3 $\mathbf{(k = 3)}$, so the binary classification is: (1) Normal against the rest (Osteopenia and Osteporosis together), (2) Normal and Osteopenia combined against Osteporosis. Here our success terms are Normal and Normal with Osteopenia together. More generally, if an ordinal response variable $\mathbf{y}$ has $\mathbf{k}$ classes $\mathbf{(k = 1,2,...,k)}$ then there are $\mathbf{(k-1)}$ ways to make binary classification. The appropriate model to solve the OR is the cumulative logit model (proportional odds model) (Kleinbaum and Klein, 2010). Because there are $\mathbf{k}$ classes, the model actually makes $\mathbf{(k-1)}$ classifications, each corresponding to the accumulation of probability across successive classes. For $\mathbf{k}$ classes of the response variable with probabilities $\pi_1, \pi_2, ..., \pi_k$ the cumulative logit classifier is defined as (Kleinbaum and Klein, 2010):

$$\mathbf{\log it = [p(y \le k)] = \ln[\frac{p(y \le k)}{1 - p(y \le k)}] \quad , k = 1,2,.....,k-1} \qquad ....(22)$$

In regression term equation (22) be

$$\mathbf{\log it\, p(y_i \le k) = \alpha_k + \sum_{i=1}^{p} \beta_i x_i \quad ,k = 1,2,....,k-1} \qquad ....(23)$$

The classification of the new data will be assigned to class $\mathbf{k}$ depending on the maximum probability of this class or category.

5- Application Case

5-1 Data Description

Osteoporosis is a major public health problem projected to generate an increasingly heavier social and economic toll in view of the ageing population worldwide. The data was taking from the study that conducted to find the affected variable on the Osteoporosis in women in Mosul city. The study consisted of 17 explanatory variables and their effect on the type of bone mineral density as a response variable on 344 women (Al-Jumaily, 2010).

5-2 Results

The response variable has three classes $\mathbf{(k = 3)}$ with Normal class (29.1%), Osteopenia (38.1%), and Osteoporosis (32.8%). The data were split into training data to build the MSVM (88.37%) and a sample was taken with (11.63%) from data to represent the test data to validate the MSVM. The performance of the four MSVM approaches discussed in the previous section (OAA, OAO, WW, and CS) has been studied. Also we used OR model as non MSVM. For all MSVM approaches, the radial basis function (RBF), $\mathbf{k(x_i, x_j) = exp(\frac{\|x_i - x_j\|^2}{-2\sigma^2})}$, is employed. The RBF dependent on Euclidean distance of $\mathbf{x_j}$ from $\mathbf{x_i}$ (one of these will be the support vector and the other will be the testing data point). Each binary classifier requires the selection of two hyperplane parameters: a regularization parameter $\mathbf{C}$, which is set to be **1**, **10**, and **100**, and a kernel parameter σ, which is set to be **0.1**, **0.5**, and **1**. Table 1 shows the OR classification results. The results of the MSVM approaches are reported in tables (2-4). We use R2.14 and WEKA programs to get the results.

Table (1): Ordinal regression classification results.

	Training error	Percentage of correctly classified data
OR	0.2301	51.5%

From table (1) we notice that the value of the percentage of correctly classified data still the same as compared with all multi-class SVM.

Table (2): MSVM approaches with $\mathbf{C = 1}$

	MSVM approaches	Training error	No. of SV	Percentage of correctly classified
$\sigma = \mathbf{0.1}$	OAA	0.132	299	61%
	OAO	0.134	298	62.5%
	CS	**0.107**	**295**	**62.5%**
	WW	0.151	296	57.5%
$\sigma = \mathbf{0.5}$	OAA	0.052	268	53%
	OAO	0.055	269	52.5%
	CS	**0.032**	**254**	**60%**
	WW	0.1809	265	40%
$\sigma = \mathbf{1}$	OAA	0.02	296	43%
	OAO	0.023	299	42.5%

	CS	**0.0197**	**298**	**55%**
	WW	0.046	299	40%

In table (2) we set the value of the regularization parameter **C** to be **1** and we vary the value of kernel parameter σ to be **0.1**, **0.5**, and **1.** One could observe that the training error gradually decreases when the kernel parameter σ gradually increases.

Table (3): MSVM approaches with $\mathbf{C = 10}$

	MSVM approaches	Training error	No. of SV	Percentage of correctly classified data
$\sigma = \mathbf{0.1}$	OAA	0.0562	254	56%
	OAO	0.0559	259	55%
	CS	**0.0493**	**242**	**60%**
	WW	0.2006	249	52.5%
$\sigma = \mathbf{0.5}$	OAA	0.0199	295	50%
	OAO	0.023	293	47.5%
	CS	**0.0197**	**285**	**52.5%**
	WW	0.023	287	47.5%
$\sigma = \mathbf{1}$	OAA	0.0166	296	42%
	OAO	0.0164	298	42.5%
	CS	**0.0131**	**295**	**52.5%**
	WW	0.0197	296	35%

Table (4): MSVM approaches with $\mathbf{C = 100}$

	MSVM approaches	Training error	No. of SV	Percentage of correctly classified data
$\sigma = \mathbf{0.1}$	OAA	0.0312	258	59%
	OAO	0.036	257	59%
	CS	**0.029**	**229**	**60%**

	WW	0.0361	241	58%
σ = 0.5	OAA	0.0099	295	50%
	OAO	0.0098	294	50%
	CS	**0.0098**	**277**	**55%**
	WW	0.0131	281	47.5%
σ = 1	OAA	0.0099	298	47%
	OAO	0.0098	297	45%
	CS	**0.0098**	**294**	**52.5%**
	WW	0.0131	294	35%

In table (3) and table (4) we set the value of the regularization parameter **C** to be **10** and **100,** respectively and we vary the value of kernel parameter σ to be **0.1**, **0.5**, and **1.** One could observe that the training error gradually decreases when the regularization parameter **C** gradually increases.

6- Conclusions

In this paper, we apply a classical multi class support vector machine approaches and ordinal regression to a real life multi-class data classification. The results are:

1- The Crammer and Singer (CS) approach may perform better than other multi-class classification approaches including OAA, OAO, WW, and OR for our application case, where the training error has the smallest value among the other approaches.

2- We conclude that when a regularization parameter **C** and a kernel parameter σ be large, the training error becomes less, although the percentage of correctly classified test data differs.

3- Although, the number of support vectors varies among these approaches, it has not any effect on the percentage of correctly classified test data.

4- The use of ordinal regression makes $\mathbf{(k-1)}$ binary classification since the data have order form which makes it uses compared with the other multi-class support vector machine approaches.

7- References

1. Abe, Sh., 2010, “Support Vector Machines for Pattern Classification”, 2nd ed., Springer-Verlang London Limited, NY.
2. Al-Jumaily, H., H., 2010, “Assessment for Osteoporotic Women in Mosul City”, MSc thesis, College of Nursing, Mosul University.
3. Crammer, K. and Singer, Y., 2000, “On the Learnability and Design of Output Codes for Multi Class Problems”, Computational Learning Theory, pp.35-46.
4. Hsu, C., W. and Lin, C., J., 2002, “A comparison of Methods for Multi-class Support Vector Machines”, IEEE Transactions on Neural Networks, Vol. 13, pp.415-425.
5. Hu, X. and Pan, Y., 2007, “Knowledge Discovery in Bioinformatics, Techniques, Methods, and Applications”, John Wiley & Sons, INC., NJ.
6. Ivanciuc, O., 2007, “Application of Support Vector Machines in Chemistry”, Reviews Computational Chemistry, Vol.23, pp.291-400.
7. Karatzoglou, A. and Meyer, D., 2006, “Support Vector Machines in R”, Journal of Statistical Software, Vol.15, Iss. 9, pp.1-28.
8. Kleinbaum, D.,G. and Klein, M., 2010, “Logistic Regression A self Learning Text”, 3rd ed., Springer Science + Business Media LLC, NY.

9. Liang, Y., Xu, Q., Li, H., and Cao, D., 2011, “Support Vector Machines and Their Application in Chemistry and Biotechnology”, Taylor and Francis Group, LLC., NY.
10. Monfrini, E. and Guermeur, Y., 2011, “A Quadratic Loss Multi-Class Support Vector Machine for Which a Rdius-Margin Bound Applies”, Informatica, Vol.22, No.1, pp.73-96.
11. Sangeetha, R. and Kalpana, B, 2011, “Performance Evaluation of Kernels in Multiclass Support Vector Machines”, International Journal of Soft Computing and Engineering, Vol.1, Iss.5, pp.138-145.
12. Seeja, K.,R. and Shweta, L., 2011, “Microarray Data Classification Using Support Vector Machines”, International Journal of Biometrics and Bioinformatics, Vol.5, Iss.1, pp.10-15.
13. Statnikov, A., Aliferis, C., F., Hardin, D., P., and Guyon, I., 2011, “A Gentle Introduction to Support Vector Machines in Biomedicine”, World Scientific Publishing Co, Pte., Ltd., Singapore.
14. Vapnik, V., 2010, “The Nature of Statistical Learning Theory”, 2nd ed., Springer-Verlage New York, Inc., NY.
15. Westone, J. and Watkins, C., 1999, “Support Vector Machine for Multi-Class Pattern Recognition”, Proceeding of the 7th European Symposium on Artificial Neural Networks.
16. Xia, F., Zhou, L., Yang, Y., and Zhang, W., 2007, “Ordinal Regression as Multiclass Classification”, International Journal of Intelligent Control and Systems, Vol.12, No.3, pp.230-236.