Minimum Enclosing Spheres Formulations for Support Vector Ordinal Regression

S K Shevade
Computer Science and Automation
Indian Institute of Science
Bangalore - 560 012, INDIA
shirish@csa.iisc.ernet.in

Wei Chu
Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA
chuwei@cs.columbia.edu

Abstract

We present two new support vector approaches for ordinal regression. These approaches find the concentric spheres with minimum volume that contain most of the training samples. Both approaches guarantee that the radii of the spheres are properly ordered at the optimal solution. The size of the optimization problem is linear in the number of training samples. The popular SMO algorithm is adapted to solve the resulting optimization problem. Numerical experiments on some real-world data sets verify the usefulness of our approaches for data mining.

1 Introduction

We consider the supervised learning problem of predicting variables of ordinal scale, a setting referred to as *ranking learning* or *ordinal regression*. Here, the training samples are labelled by a set of ranks, which exhibits an order among different categories. This problem arises frequently in information retrieval where a user grades the documents based on their importance. For example, a user can grade every document in a set of retrieved documents into one of the following categories: *highly relevant*, *relevant*, *average*, *irrelevant* and *highly irrelevant*. These grades are different from the class labels in classification problem as they represent some ranking information. Standard classification problems cannot make use of this ranking information as they treat the class labels as unordered categories.

There are various approaches to solve the ordinal regression problem [3, 2]. The main difficulty of these approaches is that the problem size of these formulations is a quadratic function of the training data size. Shashua and Levin [6] generalized the support vector formulation for ordinal regression by finding $r-1$ separating hyperplanes which would separate the training data into r ordered classes. This is done by modeling the ranks as the intervals on the real line. The problem with this approach is that the ordinal inequalities on the thresholds, $b_1 \leq b_2 \leq \ldots \leq b_{r-1}$ are not included in the formulation. This might result in disordered thresholds at the solution. This problem can be handled by introducing explicit constraints in the problem formulation that enforce the inequalities on the thresholds [1]. In this case, the size of the optimization problem is linear in the number of training samples and the popular SMO algorithm [5, 4] for SVMs can be easily adapted. Chu and Keerthi [1] also proposed a new formulation which considers the training samples from all the ranks to determine each threshold and gave the SMO algorithm for finding the solution of this formulation.

In this paper, we propose two approaches which use minimum enclosing sphere formulations to solve the ordinal regression problem. For both approaches, the size of the optimization problems is linear in the number of training samples and the SMO algorithm can be easily adapted. Comparison with the approaches in [1] on several benchmark datasets shows that the proposed approaches are competitive and scale well.

The paper is organized as follows. In Section 2 we present the minimum enclosing sphere formulation with explicit inequality constraints on the radii. The second approach with implicit constraints on the radii is discussed in Section 3. Numerical experiments comparing the two approaches with the other support vector approaches on various benchmark datasets are given in Section 4. Section 5 concludes the paper.

A word about our notation. We will use x to denote the input vector of the ordinal regression problem. Let $z = \phi(x)$ denote the feature space representation in a high dimensional reproducing kernel Hilbert space (RKHS) related to x by a transformation. All the computations are done using the reproducing kernel function, which is defined as $K(x_p, x_q) = \phi(x_p) \cdot \phi(x_q) = z_p \cdot z_q$ where $\|\cdot\|$ denotes inner product in the RKHS. We consider the ordinal regres-

sion problem with r ordered categories. These categories will be denoted by consecutive integers, $\mathbf{Y} = \{1, \ldots, r\}$. Assume that n_j training examples exist in the j-th category where $j \in \mathbf{Y}$ and let the i-th training sample in this category be denoted by x_i^j, where $x_i^j \in \mathbb{R}^d$. The total number of training samples is n ($= \sum_{j=1}^{r} n_j$).

2 Approach 1: Explicit Constraints on Radii

Support Vector Machines (SVM) map the input vectors into feature vectors in a high dimensional RKHS [8], where a linear machine is constructed by minimizing a regularized functional. For ordinal regression problem, the support vector formulation determines $r-1$ separating hyperplanes, $\mathbf{w} \cdot \phi(x) - b_j,\ j = 1, \ldots, r-1$, which would separate the training data into r ordered bins [6, 1]. For each threshold b_j, the samples from the two adjacent categories, j and $j+1$ are considered for empirical errors.

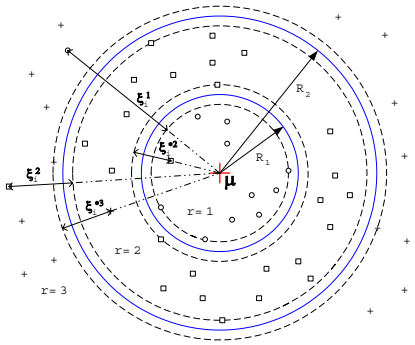
We formulate the ordinal regression problem as finding $r-1$ minimum volume concentric spheres with centre μ and radii, $R_j, j = 1, \ldots, r-1$ such that all (or most of) the data points in j-th category lie in the annular region formed by the spheres of radii R_j and R_{j-1}. Here, we introduce two auxiliary variables, $R_0 = 0$ and $R_r = \infty$. One or a few very remote training samples could result in very large spheres which will not represent the data well. Therefore, we allow for some data points in every category to lie outside their designated region. More specifically, the squared distance of each sample in the j-th category from the centre μ should be at most $R_j^2 - 1$, otherwise $\|z_i^j - \mu\|^2 - (R_j^2 - 1)$ is the error (denoted as ξ_i^j). Similarly, the squared distance of each sample in the j-th category from the centre μ should be at least $R_{j-1}^2 + 1$, otherwise $R_{j-1}^2 + 1 - \|z_i^j - \mu\|^2$ is the error (denoted as ξ_i^{*j}). See Figure 1 for the illustration. Then the ordinal regression problem can be formulated as the following *primal* problem:

$$
\begin{array}{ll}
\min & \sum_{j=1}^{r-1} R_j^2 + C \sum_{j=1}^{r} \sum_{i=1}^{n_j} (\xi_i^j + \xi_i^{*j}) \\
\text{s.t.} & \|z_i^j - \mu\|^2 \leq R_j^2 + \xi_i^j - 1,\ \xi_i^j \geq 0, \\
& j = 1, \ldots, r-1, i = 1, \ldots, n_j \\
& \|z_i^j - \mu\|^2 \geq R_{j-1}^2 - \xi_i^{*j} + 1,\ \xi_i^{*j} \geq 0, \\
& j = 2, \ldots, r, i = 1, \ldots, n_j \\
& R_{j-1}^2 \leq R_j^2\ \forall j = 2, \ldots, r-1
\end{array}
$$

where C gives the trade-off between the volume of the spheres and the number of errors. Note that $\xi_i^r = \xi_i^{*1} = 0\ \forall\ i$.

The *primal* problem is a convex programming problem. Using Wolfe duality theory, by introducing the KKT conditions into the Lagrangian and applying the kernel trick, the

Figure 1. An illustration of the definition of slack variables ξ and ξ^*.The samples from rank 1, 2 and 3 are shown respectively by circles, squares and crosses. ξ and ξ^* indicate the squared distances between the points under consideration.



dual problem becomes

$$
\begin{array}{ll}
\min & \frac{1}{r-1} \sum_{i,j} \sum_{i',j'} (\alpha_i^j - \alpha_i^{*j})(\alpha_{i'}^{j'} - \alpha_{i'}^{*j'}) K(x_i^j, x_{i'}^{j'}) \\
& - \sum_{j,i} (\alpha_i^j - \alpha_i^{*j}) K(x_i^j, x_i^j) - \sum_{j,i} (\alpha_i^j + \alpha_i^{*j}) \\
\text{s.t.} & 0 \leq \alpha_i^j, \alpha_i^{*j} \leq C,\ \ \forall\ i,\ j \\
& \sum_{i=1}^{n_j} \alpha_i^j - \sum_{i=1}^{n_{j+1}} \alpha_i^{*j+1} = \eta^{j+1} - \eta^j + 1 \\
& \forall\ j = 1, \ldots, r-1 \\
& \eta^j \geq 0\ \ \forall\ j = 2, \ldots, r-1
\end{array}
$$

The *dual* problem is a convex quadratic programming problem. One can solve this problem with respect to α_i^j, α_i^{*j} and η^j using SMO type algorithm. The training set samples with non-zero values of α_i^j or α_i^{*j} are called *support vectors*. Once the optimal values of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^*$ and $\boldsymbol{\eta}$ are obtained by solving this dual problem, the different radii, R_j, of the spheres can be obtained by calculating the distance between the centre of the sphere and a support vector whose Lagrange multiplier lies in the range $(0, C)$.

The rank of a test sample x can be determined by finding the smallest sphere in the RKHS (out of r spheres) which encompasses the given sample. For this purpose, the distance between $\phi(x)$ and the centre of sphere, μ, needs to be calculated. The rank of any point x is then given by

$$\arg\min_k \{k : \|\phi(x) - \mu\|^2 < R_k^2\}. \tag{1}$$

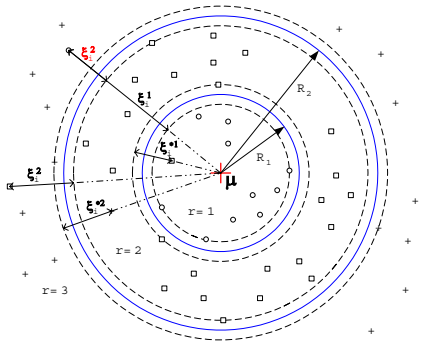
where μ is derived from the KKT conditions as

$$\mu = \frac{\sum_{j=1}^{r} \sum_{i=1}^{n_j} (\alpha_i^j - \alpha_i^{*j}) z_i^j}{r-1}. \tag{2}$$

Let us define, $f(x) = K(x,x) - \frac{2}{r-1} \sum_{j,i} (\alpha_i^j - \alpha_i^{*j}) K(x_i^j, x)$. Then, the rank of x is given by

$$\arg\min_k \{k : f(x) < R_k^2 - \|\mu\|^2\}. \tag{3}$$

Figure 2. An illustration of the definition of slack variables ξ and ξ^*. The samples from rank 1, 2 and 3 are shown respectively by circles, squares and crosses. ξ and ξ^* indicate the squared distances between the points under consideration.



3 Approach 2: Implicit Constraints on Radii

In this section we present a different formulation for support vector ordinal regression. As in the case of Approach 1, this formulation also finds minimum volume concentric spheres with centre μ such that all (or most of) the data points in the j-th category lie in the annular region formed by spheres of radii R_j and R_{j-1}. However, instead of considering the empirical errors from the samples of adjacent categories to determine a radius, we allow the samples in all the categories to contribute errors for each radius. More specifically, the squared distance of a sample z_i in the category y_i from the centre μ should be at most $R_j^2 - 1$ for every category $j \geq y_i$ and should be at least $R_j^2 + 1$ for every category $j < y_i$. Otherwise, $\|z_i - \mu\|^2 - (R_j^2 - 1)$ is the error denoted as $\xi_i^j, j \geq y_i$ and $(R_j^2 + 1) - \|z_i - \mu\|^2$ is the error denoted as $\xi_i^{*j}, j < y_i$. See Figure 2 for the illustration.

We can thus write the new support vector ordinal regression problem as the following *primal problem*:

$$\min \sum_{j=1}^{r-1} R_j^2 + C \sum_{j=1}^{r-1} \left(\sum_{i:y_i \leq j} \xi_i^j + \sum_{i:y_i > j} \xi_i^{*j} \right) \qquad (4)$$

s.t.

$$\begin{aligned} \|z_i - \mu\|^2 &\leq R_j^2 - 1 + \xi_i^j, \xi_i^j \geq 0 \text{ for } i : y_i \leq j \\ -\|z_i - \mu\|^2 &\leq -R_j^2 - 1 + \xi_i^{*j}, \xi_i^{*j} \geq 0 \text{ for } i : y_i > j \end{aligned} \qquad (5)$$

where j runs over $1, \ldots, r-1$. Thus, there are $r-1$ constraints associated with every sample rather than only two constraints as in the case of the formulation discussed in Section 2. A very nice property of this approach is that the inequality constraints $R_{j+1}^2 \geq R_j^2, j = 1, \ldots, r-1$ always hold at the optimal solution in spite of the fact that they are not explicitly included in the new formulation. The proof of this property is given below.

To prove the inequalities on the R_j's at the optimal solution, let us consider the situation where the center μ is fixed and only the R_j's are optimized. In this case, the empirical errors ξ_i^j and ξ_i^{*j} are automatically determined once the R_j is given. For each rank k let us define, $I_k^{\text{low}} \stackrel{\text{def}}{=} \{i : y_i = k \text{ and } \|z_i - \mu\|^2 - R_k^2 \geq -1\}$ and $I_k^{\text{up}} \stackrel{\text{def}}{=} \{i : y_i = k \text{ and } \|z_i - \mu\|^2 - R_k^2 \leq +1\}$. It is easy to see that R_j is optimal iff it minimizes the functional, $e_j(R^2) = \frac{1}{C} R^2 + \sum_{k=1}^{j} \sum_{i \in I_k^{\text{low}}} (\|z_i - \mu\|^2 - R^2 + 1) + \sum_{k=j+1}^{r} \sum_{i \in I_k^{\text{up}}} (R^2 + 1 - \|z_i - \mu\|^2)$, Due to strict convexity, there is a unique solution to this problem. Let $\tilde{R}_j^2$ denote the minimizer of $e_j(R^2)$. Now we need to prove the following lemma.

Lemma 1. $\tilde{R}_1^2 \leq \tilde{R}_2^2 \leq \cdots \leq \tilde{R}_{r-1}^2$

Proof. The "right side derivative" of e_j with respect to R^2 is

$$g_j(R^2) = \tfrac{1}{C} - \sum_{k=1}^{j} |I_k^{\text{low}}(R^2)| + \sum_{k=j+1}^{r} |I_k^{\text{up}}(R^2)| \qquad (6)$$

where $|I_k^{\text{low}}(R^2)|$ denotes the size of the set $I_k^{\text{low}}(R^2)$.[1] Take any one j and suppose $\tilde{R}_j^2 > \tilde{R}_{j+1}^2$. Since $\tilde{R}_{j+1}^2$ is strictly to the left of $\tilde{R}_j^2$ that minimizes e_j, we have $g_j(\tilde{R}_{j+1}^2) < 0$.

Since $\tilde{R}_{j+1}^2$ is a minimizer of e_{j+1} we also have $g_{j+1}(\tilde{R}_{j+1}^2) \geq 0$. Thus we have $g_{j+1}(\tilde{R}_{j+1}^2) - g_j(\tilde{R}_{j+1}^2) > 0$; also, by (6) we get

$$0 < g_{j+1}(\tilde{R}_{j+1}^2) - g_j(\tilde{R}_{j+1}^2) = -|I_{j+1}^{\text{low}}(\tilde{R}_{j+1}^2)| - |I_{j+1}^{\text{up}}(\tilde{R}_{j+1}^2)|$$

which is impossible. This proves the lemma.

It is worth noting that Lemma 1 holds even for an extended problem formulation that allows the use of different costs (different C values) for different misclassifications (class k misclassified as class j can have the cost C_k^j). This problem formulation is more appropriate for applications such as collaborative filtering.

Using Wolfe duality theory, the *dual* problem of (4) - (5) can be written as

$$\begin{aligned} &\max \sum_i \left(\sum_{j=y_i}^{r-1} \alpha_i^j + \sum_{j=1}^{y_i-1} \alpha_i^{*j} \right) \\ &-\tfrac{1}{r-1} \sum_i \sum_k \Psi_i(\alpha, \alpha^*) \Psi_k(\alpha, \alpha^*) K(x_i, x_k) \\ &+ \sum_i \left(\sum_{j=y_i}^{r-1} \alpha_i^j - \sum_{j=1}^{y_i-1} \alpha_i^{*j} \right) K(x_i, x_i) \end{aligned}$$

subject to the constraints, $\sum_{i:y_i \leq j} \alpha_i^j - \sum_{i:y_i > j} \alpha_i^{*j} = 1 \, \forall \, j$ and $0 \leq \alpha_i^j, \alpha_i^{*j} \leq C \, \forall \, i, j$, where $\Psi_p(\alpha, \alpha^*) =$

[1] The curve of g_j is of slope 1 almost anywhere and having finite vertical jumps.

$\sum_{j=y_p}^{r-1} \alpha_p^j - \sum_{j=1}^{y_p-1} \alpha_p^{*j}$. One can solve the dual problem with respect to α_i^j and α_i^{*j} using SMO type algorithm. Once the optimal α are obtained, the different radii, R_j of the spheres can be obtained by calculating the distance between the centre μ of the sphere and the training set sample whose Lagrange multiplier is in $(0, C)$.

The rank of any point x is then given by

$$\arg\min_k \{k : \|\phi(x) - \mu\|^2 < R_k^2\}. \quad (7)$$

Let us define $f(x) = K(x,x) - \frac{2}{r-1}\sum_i(\sum_{j=y_i}^{r-1} \alpha_i^j - \sum_{j=1}^{y_i-1} \alpha_i^{*j})K(x, x_i)$. Then, the rank of x is given by

$$\arg\min_k \{k : f(x) < R_k^2 - \|\mu^2\|\}. \quad (8)$$

4 Numerical Experiments

The SMO algorithms for the two proposed support vector ordinal regression formulations were evaluated on some benchmark data sets. We collected eight benchmark datasets that were used for metric regression problems.[2] The target values were discretized into ordinal quantities using equal length binning. These bins divide the range of target values into a given number of intervals that are of same length. The resulting rank values are ordered, representing these intervals of the original metric quantities. For each dataset, we generated two versions by discretizing the target values into five or ten ordinal scales respectively. Each dataset was randomly partitioned into training/test splits as specified in Table 1. The partitioning was repeated 20 times independently.[3] The input vectors were normalized to zero mean and unit-variance, coordinate-wise. The Gaussian kernel, $K(x, x') = \exp(-\frac{\kappa}{2}\sum_{p=1}^{d}(x_p - x'_p)^2)$ where $\kappa > 0$ and x_p denotes the p-th element of the input vector x, was used in these experiments. Five-fold cross validation was used to determine the optimal values of model parameters (the parameter κ in the Gaussian kernel and the regularization factor C) involved in the problem formulations, and the test error was obtained using the optimal model parameters for each formulation. The following two metrics which quantify the accuracy of predicted ordinal scales $\{\hat{y}_1, \ldots, \hat{y}_t\}$ with respect to true targets $\{y_1, \ldots, y_t\}$ were used: 1. *Mean absolute error* is the average deviation of the prediction from the true target, i.e. $\frac{1}{t}\sum_{i=1}^{t}|\hat{y}_i - y_i|$, in which we treat the ordinal scales as consecutive integers. 2. *Mean zero-one error* is the fraction of incorrect predictions on individual samples. We compare the generalization capabilities of the proposed approaches with the Support Vector approaches for ordinal regression described in [1].

[2]These regression datasets are available at http://www.liacc.up.pt/~ltorgo/Regression/Datasets.html.

[3]The generated partitions can be accessed at http://www.gatsby.ucl.ac.uk/~chuwei/ordinalregression.html.

Table 1. Datasets and their characteristics

Datasets	#Attributes	Training Samples	Test Samples
Diabetes	2	30	13
Pyridimines	27	50	24
Triazines	60	100	86
Wisconsin	32	130	64
Machine CPU	6	150	59
Auto MPG	7	200	192
Boston Housing	13	300	206
Abalone	8	1000	3177

Table 3. CPU Time (in seconds) for the proposed algorithms on the California Housing dataset and the Bank Domain dataset. EXC (R) and IMC (R) indicate the proposed approaches with explicit and implicit constraints on the radii.

Training dataset size	CPU Time (sec)			
	California Housing		Bank Domain	
100	0.0105	0.055	0.007	0.0815
200	0.0355	.1315	0.03	.2315
500	0.371	0.6845	0.205	1.2345
1000	2.5685	4.417	1.7225	6.6355
2000	38.28	28.82	23.649	34.1325
5000	1279	254.01	320.691	163.6165
8000	-	-	2473.3	645.921

The test results of these algorithms are reported in Table 2. From this table, it is clear that the proposed approaches are competitive. In particular, the approach with implicit constraints on the radii (Approach 2) is comparable with the support vector approaches in [1] in terms of mean absolute error. Clearly, Approach 2 discussed in this paper is an excellent alternative to the Support Vector approaches [1] for ordinal regression.

4.1 Scaling

In this experiment, we empirically studied how the two SMO algorithms scale with respect to training data size. The California housing dataset and the Bank Domain dataset were used in the scaling experiments.[4] For the California housing dataset, twenty training datasets with sizes ranging from 100 to 5000 were generated by random selection from the original dataset. The continuous target variable of the California Housing dataset was discretized to ordinal scale by using 10 equal-frequency bins. The datasets were trained by the SMO algorithms for ordinal regression formulations (using explicit and implicit constraints on the radii) with a linear kernel and $C = 10$. For the Bank Domain dataset, a similar procedure was adapted to generate training datasets with sizes ranging from 100 to 8000. The details of the CPU times of the SMO algorithms for differ-

[4]The California Housing dataset and the Bank Domain dataset are available at http://www.liacc.up.pt/~ltorgo/Regression/.

Table 2. Comparison of the proposed approaches with the approaches reported in [1] on different benchmark datasets. The targets of these benchmark datasets were discretized by 10 and 5-equal length bins. The results are the averages over 20 trials, along with the standard deviation. EXC (R) and IMC (R) respectively represent the proposed approaches with explicit and implicit constraints on the radii. EXC (H) and IMC (H) respectively denote the approaches with explicit and implicit constraints on the thresholds [1]. We use bold face to indicate the best result in each row.

	10-equal length bins				5-equal length bins			
Data	Mean absolute error				Mean absolute error			
	EXC (R)	IMC(R)	EXC (H)	IMC (H)	EXC (R)	IMC(R)	EXC (H)	IMC (H)
Diabetes	1.50 ± 0.46	1.57 ± 0.47	$\mathbf{1.34 \pm 0.40}$	1.57 ± 0.48	0.68 ± 0.14	0.72 ± 0.16	$\mathbf{0.67 \pm 0.12}$	0.70 ± 0.12
Pyridimines	0.92 ± 0.12	$\mathbf{0.86 \pm 0.11}$	0.95 ± 0.21	0.88 ± 0.21	0.48 ± 0.11	0.46 ± 0.11	0.47 ± 0.13	$\mathbf{0.46 \pm 0.09}$
Triazines	1.33 ± 0.09	$\mathbf{1.25 \pm 0.08}$	1.32 ± 0.07	1.30 ± 0.08	0.70 ± 0.05	$\mathbf{0.70 \pm 0.04}$	0.71 ± 0.03	0.71 ± 0.03
Wisconsin	2.67 ± 0.23	$\mathbf{2.06 \pm 0.15}$	2.62 ± 0.28	2.14 ± 0.23	1.14 ± 0.14	$\mathbf{1.00 \pm 0.09}$	1.18 ± 0.15	1.02 ± 0.08
MachineCPU	0.46 ± 0.06	$\mathbf{0.45 \pm 0.08}$	0.47 ± 0.06	0.46 ± 0.06	0.19 ± 0.04	$\mathbf{0.18 \pm 0.04}$	0.19 ± 0.04	0.19 ± 0.04
Auto MPG	0.52 ± 0.04	0.52 ± 0.03	0.52 ± 0.03	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.26 \pm 0.02}$	0.27 ± 0.02	$\mathbf{0.26 \pm 0.02}$	$\mathbf{0.26 \pm 0.02}$
Boston	$\mathbf{0.50 \pm 0.04}$	0.51 ± 0.04	0.51 ± 0.04	0.51 ± 0.04	0.28 ± 0.02	$\mathbf{0.27 \pm 0.02}$	0.28 ± 0.02	$\mathbf{0.27 \pm 0.02}$
Abalone	0.54 ± 0.01	$\mathbf{0.52 \pm 0.01}$	0.53 ± 0.01	$\mathbf{0.52 \pm 0.01}$	$\mathbf{0.24 \pm 0.01}$	0.23 ± 0.01	0.25 ± 0.01	$\mathbf{0.24 \pm 0.01}$
	Mean zero-one error (%)				Mean zero-one error (%)			
Diabetes	77.30 ± 8.08	76.54 ± 9.82	$\mathbf{75.77 \pm 8.74}$	78.85 ± 11.13	54.23 ± 12.60	58.84 ± 13.27	$\mathbf{53.47 \pm 11.57}$	58.08 ± 12.09
Pyridimines	61.67 ± 6.28	58.33 ± 8.76	63.13 ± 7.80	$\mathbf{57.29 \pm 10.20}$	43.54 ± 8.71	41.88 ± 9.01	43.13 ± 9.10	$\mathbf{41.67 \pm 6.48}$
Triazines	71.67 ± 3.02	$\mathbf{69.77 \pm 2.99}$	71.28 ± 2.54	71.92 ± 3.52	53.66 ± 2.78	$\mathbf{53.60 \pm 2.77}$	53.95 ± 1.57	54.19 ± 2.18
Wisconsin	$\mathbf{80.78 \pm 3.87}$	84.30 ± 3.88	82.16 ± 3.74	85.94 ± 3.89	66.64 ± 4.84	69.69 ± 4.77	$\mathbf{66.41 \pm 5.66}$	72.03 ± 3.65
MachineCPU	33.39 ± 4.04	$\mathbf{32.71 \pm 4.57}$	34.57 ± 3.27	32.97 ± 3.63	17.12 ± 3.06	$\mathbf{16.52 \pm 3.10}$	17.03 ± 3.28	17.37 ± 3.15
Auto MPG	44.67 ± 2.76	$\mathbf{42.29 \pm 3.11}$	44.19 ± 2.75	44.22 ± 2.86	25.96 ± 1.53	26.43 ± 1.71	25.73 ± 2.19	$\mathbf{25.57 \pm 2.33}$
Boston	40.99 ± 2.46	42.18 ± 2.81	$\mathbf{40.90 \pm 2.82}$	42.23 ± 3.1	25.44 ± 1.87	$\mathbf{25.36 \pm 1.91}$	25.44 ± 1.99	25.36 ± 2.19
Abalone	$\mathbf{42.77 \pm 0.62}$	42.99 ± 0.88	42.98 ± 0.79	42.95 ± 0.77	21.88 ± 0.47	$\mathbf{21.80 \pm 0.42}$	22.28 ± 0.37	21.99 ± 0.40

ent dataset sizes are reported in Table 3. The SMO algorithm for the formulation with implicit constraints has scaling exponents of 2.4 and 2.0 respectively on the California Housing dataset and the Bank Domain dataset. On the other hand, the scaling exponents of the SMO algorithm for the formulation with explicit constraints on these datasets were close to 3. Thus the SMO algorithm for the support vector ordinal regression formulation with implicit constraints on the radii has better scaling properties and is suitable for data mining applications.

5 Conclusion

Ordinal regression is an important supervised learning problem with properties of both metric regression and classification. In this paper, we proposed two new approaches to support vector ordinal regression which find $r-1$ concentric spheres for r ranks using $r-1$ radii. We also adapted the SMO algorithms for these formulations. The results of numerical experiments suggested that the generalization capabilities of the proposed approaches are competitive. Also, the algorithm for implicit constraints formulation was found to scale well. Thus, the proposed approaches are useful alternatives to the existing support vector approaches and are suitable for applications such as collaborative filtering. Further, they can also be used for data with ordinal inputs.

Recently, the idea of Core Vector Machines (CVM) was proposed where the two-category classification problem was formulated as minimum enclosing ball (MEB) problem in computational geometry [7]. The resulting algorithm is very fast and is especially useful for very large datasets. The idea of CVM can be extended to the formulations proposed in this paper. We are currently investigating these details and the results will be reported elsewhere.

References

[1] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the International Conference on Machine Learning*, 2005.

[2] S. Har-Peled, D. Roth, and D. Zimac. Constraint classification: A new approach to multiclass classification and ranking. In *Advances in Neural Information Processing Systems 15*, 2002.

[3] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

[4] S. S. Keerthi, S. K. Shevade, C. Bhattachryya, and K. R. K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649, March 2001.

[5] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press, 1999.

[6] A. Shashua and A. Levin. Ranking with large margin principle: two approaches. In *Advances in Neural Information Processing Systems 15*, 2002.

[7] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[8] V. N. Vapnik. *The nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.