Int. J. Mach. Learn. & Cyber. (2017) 8:2025–2034
DOI 10.1007/s13042-016-0582-3

ORIGINAL ARTICLE

Minimum class variance support vector ordinal regression

Xiaoming Wang[1] · Jinrong Hu[1,2] · Zengxi Huang[1]

Received: 7 November 2015 / Accepted: 8 August 2016 / Published online: 18 August 2016

Abstract The support vector ordinal regression (SVOR) method is derived from support vector machine and developed to tackle the ordinal regression problems. However, it ignores the distribution characteristics of the data. In this paper, we propose a novel method to handle the ordinal regression problems. This method is referred to as minimum class variance support vector ordinal regression (MCVSVOR). In contrast with SVOR, MCVSVOR explicitly takes into account the distribution of the categories and achieves better generalization performance. Moreover, the problem of MCVSVOR can be transformed into one of SVOR. Thus, the existing software of SVOR can be used to solve the problem of MCVSVOR. In the paper, we first discuss the linear case of MCVSVOR and then develop the nonlinear MCVSVOR through using the kernelization trick. The comprehensive experiment results show that the proposed method is effective and can achieve better generalization performance in contrast with SVOR.

Keywords Machine learning · Ordinal regression · Support vector machine · Support vector ordinal regression

1 Introduction

Over the past decades, a lot of attention has been drawn to machine learning of which many methods have been successfully applied in substantial practical applications including protein–protein interactions prediction [1–3], image retrieval [4], and text data analysis [5]. Ordinal regression is a type of supervised machine learning problems and can be applied in a wide range of fields such as medical image analysis [6], image ranking [7], facial age estimation [8], and so on. It involves how to find an order among the different categories according to the learned rules which are used to predict order of ordinal scale [9–13]. Therefore, the key issue in ordinal regression is how to learn the prediction rules from the training data. Ordinal regression differs from traditional regression and traditional classification [14–19]. The reason is that the labels of different categories in ordinal regression are discrete and simultaneously have an ordinal relationship.

Recently, many efforts have been directed toward tackling the ordinal regression problems [20–23]. In [20], Herbrich et al., based on a threshold model in which the threshold values of each ordinal category are estimated, explored the use of support vector (SV) learning in ordinal regression. In [21], the authors employed the classification and regression trees to tackle the ordinal regression problems. This method actually first maps the ordinal variables into numeric values and then employs the traditional classification and regression methods to solve the problems. However, it is difficult to devise an appropriate mapping function in that ones can not know the true metric distances between the ordinal scales in most cases. Thus, the application of this method may be limited. In [12], based on Gaussian processes, the authors presented a probabilistic kernel method to deal with ordinal regression. In [22], Sun et al. extended the kernel discriminant analysis (kDa) [16] algorithm to handle the ordinal regression problems. In [23], through making use of the abundance of unlabeled patterns, the authors proposed a transductive learning paradigm for the ordinal regression problems.

✉ Xiaoming Wang
wxmwm@aliyun.com

[1] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

[2] School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China

The basic idea of support vector machine (SVM) [17], which is one of the best-known kernel learning methods [24], is also developed to deal with the ordinal regression problems. By extending the formulation of SVM to ordinal regression, Shashua and Levin [25] introduced two methods for ordinal regression. These methods embody the large margin principle, which is the intuitive idea of SVM. In [13], Chu and Keerthi improved the SVM formulation for ordinal regression by including the ordinal inequalities on the thresholds. This method is referred to as support vector ordinal regression (SVOR). It can be solved by a sequential minimal optimization (SMO)-type algorithm and achieves the encouraging experimental results. In [26], Shevade and Chu further employed minimum enclosing sphere (MEB) to handle ordinal regression. In [27], the authors proposed a called block-quantized support vector ordinal regression (BQSVOR) to tackle the large-scale ordinal regression problems. All of these methods are derived from the traditional SVM and share a common property: borrowing the idea from SVM and generalizing the SVM formulation to solve the ordinal regression problems.

However, SVM is essentially a local classifier. This is because only the so-called support vectors determine the decision hyperplane of SVM, whereas all other data points have no impact on it. Therefore, the traditional SVM does not take into consideration the distribution characteristics of the data and may receive a non-robust solution. In order to overcome the limitation of SVM, in [18] the authors proposed a modified class of SVM. This method is called minimum class variance support vector machine (MCVSVM) and motivated by Fisher discriminant analysis [16]. Similarly to SVM, MCVSVM embodies the large margin principle [19]. However, different from SVM, MCVSVM can give a more robust solution in that it further considers the distribution of the categories in its model.

In this paper, we propose a novel method to handle the ordinal regression problems. This method is referred to as minimum class variance ordinal regression (MCVSVOR). The key difference between MCVSVOR and the traditional SVOR is that the former generalizes MCVSVM to deal with the ordinal regression problems and explicitly incorporates the distribution information of the categories, whereas the latter extends SVM to tackle ordinal regression and ignores the distribution characteristics of the data. Simultaneously, as SVOR, MCVSVOR also embodies the large margin principle in that it stems from MCVSVM. Following the basic idea of SVOR, we define the primal optimization model of MCVSVOR and develop the linear and nonlinear cases of MCVSVOR. We further analyze the relationship between MCVSVOR and SVOR. The analysis shows that MCVSVOR can be solved using the existing SVOR software, which makes the solution easy to be computed. The comprehensive experimental results suggest that the proposed method is effective and can achieve superior generalization performance in contrast with SVOR.

The rest of this paper is organized as follows. The related work is reviewed in Sect. 2. In Sect. 3, the linear case of MCVSVOR is first presented, and the relationship between MCVSVOR and SVOR is then analyzed. In Sect. 4, the nonlinear case of MCVSVOR is discussed. The experimental results are reported in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2 Related work

In this paper, we will address an ordinal regression problem with r ranked categories. The training dataset contains N sample points and is represented by $\{(\mathbf{x}_i^j, \mathrm{y}^j) | \mathbf{x}_i^j \in R^d, \mathrm{y}^j \in Y, i = 1, \ldots, N\}$, where $\mathbf{x}_i^j$ refers to the ith sample in the j-th category and y^j is its corresponding rank. Here, d is the dimension of the sample space and $Y = \{1, \ldots, r\}$ are consecutive integers and used to keep the known rank information of the training samples.

Besides, let N^j be the number of the samples in the j-th category and $\mathbf{X} = [\mathbf{x}_1^1, \ldots, \mathbf{x}_{N^1}^1, \mathbf{x}_1^2, \ldots, \mathbf{x}_{N^2}^2, \ldots, \mathbf{x}_1^r, \ldots, \mathbf{x}_{N^r}^r] = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ represents the sample matrix. Note, here $N = \sum_{j=1}^{r} N^j$ holds.

2.1 Support vector ordinal regression

Generally, the key step of solving the ordinal regression problems is to learn a function $f : R \to \{1, \ldots, r\}$ such that $f(\mathbf{x}_i^j) = y^j$ from the training samples [25, 26]. Therefore, SVOR directs at constructing $r - 1$ hyperplanes $\mathbf{w}^T\mathbf{x} - b_j = 0$ $(j = 1, \ldots, r)$ which can separate the samples of different categories. In the linear case, SVOR defines its primal optimization problem as [13]

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi_i^j + \xi_i^{*j}) \\ &s.t.\ \mathbf{w}^T\mathbf{x}_i^j - b_j \le -1 + \xi_i^j,\ \xi_i^j \ge 0,\ \forall i,j \\ &\quad \mathbf{w}^T\mathbf{x}_i^j - b_{j-1} \ge 1 - \xi_i^{*j},\ \xi_i^{*j} \ge 0,\ \forall i,j \\ &\quad b_{j-1} \le b_j,\ \forall i,j \end{aligned} \tag{1}$$

where $j = 1 \ldots, r$ and $i = 1, \ldots, N^j$. Note, SVOR introduces two auxiliary variables b_0 and b_r which are respectively set $b_0 = -\infty$ and $b_r = +\infty$. This type of SVOR is with the explicit constraints on the thresholds. In [13], the authors simultaneously presented another type of SVOR with the implicit constraints on the thresholds. More details

can be found in [13]. Note, SVOR is derived from SVM in which the large margin principle [17] is implements. Thus, the principle is also embodied in SVOR.

2.2 Kernel discriminant learning for ordinal regression

For the given training dataset, the within-class scatter matrix $\mathbf{S}_W$ is defined as [16]

$$\mathbf{S}_W = \sum_{j=1}^{r} \sum_{\mathbf{x}\in\mathbf{X}^j} (\mathbf{x}-\mathbf{u}^j)(\mathbf{x}-\mathbf{u}^j)^T \tag{2}$$

where $\mathbf{X}^j = \{\mathbf{x}_i^j | y^j = j,\ i = 1,\ldots,N^j\}$ is the j-th category samples, $\mathbf{u}^j = \frac{1}{N^j}\sum_{\mathbf{x}\in\mathbf{X}^j}\mathbf{x}$ is the mean sample vector of $\mathbf{X}^j$ and T denotes vector transpose. Here, N^j is the number of the j-th category samples $\mathbf{X}^j$. KDLOR defines the following optimization [22]

$$\begin{aligned} &\min_{\mathbf{w}} \quad \mathbf{w}^T\mathbf{S}_W\mathbf{w} - C\rho \\ &s.t.\ \mathbf{w}^T(\mathbf{u}^{j+1}-\mathbf{u}^j) \geq \rho,\ j = 1,2,\ldots,r-1 \end{aligned} \tag{3}$$

Obviously, KDLOR takes into account the distribution of the categories by introducing the within-class scatter matrix $\mathbf{S}_W$ in its objective function. However, KDLOR does not directly embody the large margin principle as SVM.

3 Minimum class variance support vector ordinal regression

In this section, we will first present the formulation of the linear SVOR and discuss how to solve it. Then, we will analyze the relationship between MCVSVOR and SVOR.

3.1 The linear case of the proposed method

By following the basic idea of SVOR and using the within-class scatter matrix $\mathbf{S}_W$ defined as (2), in the linear case, the primal optimization problem of the proposed method MCVSVOR is defined as follows

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b},\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{S}_W\mathbf{w} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi_i^j + \xi_i^{*j}) \\ &s.t.\ \mathbf{w}^T\mathbf{x}_i^j - b_j \leq -1+\xi_i^j,\ \xi_i^j \geq 0,\ \forall i,j \\ &\quad \mathbf{w}^T\mathbf{x}_i^j - b_{j-1} \geq 1-\xi_i^{*j},\ \xi_i^{*j} \geq 0,\ \forall i,j \\ &\quad b_{j-1} \leq b_j,\ \forall i,j \end{aligned} \tag{4}$$

where $j = 1\ldots,r$ and $i = 1,\ldots,N^j$. As SVOR, MCVSVOR actually tries to find $r-1$ binary classifiers with a shared mapping direction $\mathbf{w}$ and the additional constraint thresholds. Here, as in SVOR with the explicit threshold constraints, we explicitly include the natural ordinal inequalities on the thresholds constraints. However, in contrast with SOVR, MCVSVOR introduces the matrix $\mathbf{S}_W$ in the objective function of its primal optimization problem. In this way, MCVSVOR takes fully into account the distribution of the categories. Besides, MCVSVOR implements the large margin principle because it modifies MCVSVM, which embodies the principle, to handle the ordinal regression tasks.

Figure 1 illustrates the difference between SOVR and MCVSVOR. Here, we consider a synthetic ordinal regression task with 3 ranked categories, each category of which consists of 100 samples. Figure 1a describes the decision hyperplanes of SOVR on the artificial dataset. As can be seen in Fig. 1a, the decision hyperplanes of SOVR actually do not reflect the characteristic of the data although it can separate or rank each category. Figure 1b shows the decision hyperplanes of MCVSVOR. Obviously, they reflect the distribution characteristic of the data and shows more reasonable in contrast with ones of SOVR. This example clearly demonstrates the limitation of SOVR and the advantage of MCVSVOR in contrast with SVOR.

It is easy to find that the primal optimization problem (4) of MCVSVOR is a quadratic programming (QP) problem and similar to one (1) of SVOR. This problem can be efficiently solved by transforming it into its dual optimization problem [28]. First, we can formulate the primal Lagrangian of (4) as follows

$$\begin{aligned} L = &\frac{1}{2}\mathbf{w}^T\mathbf{S}_W\mathbf{w} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi_i^j+\xi_i^{*j}) - \sum_{j=1}^{r}\sum_{i=1}^{N^j}\alpha_i^j(-1+\xi_i^j-\mathbf{w}^T\mathbf{x}_i^j+b_j) \\ &-\sum_{j=1}^{r}\sum_{i=1}^{N^j}\alpha_i^{*j}(-1+\xi_i^{*j}+\mathbf{w}^T\mathbf{x}_i^{j+1}-b_{j-1}) - \sum_{j=1}^{r}\sum_{i=1}^{N^j}\beta_i^j\xi_i^j - \sum_{j=1}^{r}\sum_{i=1}^{N^j}\beta_i^{*j}\xi_i^{*j} \\ &-\sum_{j=1}^{r}\gamma^j(b_j-b_{j-1}) \end{aligned} \tag{5}$$

where the vectors $\boldsymbol{\alpha} = [\alpha_1^1,\ldots,\alpha_{N^r}^r]^T$, $\boldsymbol{\alpha}^* = [\alpha_1^{*1},\ldots,\alpha_{N^r}^{*r}]^T$, $\boldsymbol{\beta} = [\beta_1^1,\ldots,\beta_{N^r}^r]^T$, $\boldsymbol{\beta}^* = [\beta_1^{*1},\ldots,\beta_{N^r}^{*r}]^T$ and $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_r]^T$ are the Lagrangian multipliers for the constraints of (4). By differentiating with respect to $\mathbf{w}$, $\boldsymbol{\xi}$, $\boldsymbol{\xi}^*$ and $\mathbf{b}$, we have the following formulas

$$\begin{aligned} &\frac{\partial L}{\partial \mathbf{w}} = \mathbf{S}_W\mathbf{w} - \sum_{j=1}^{r}\sum_{i=1}^{N^j}(\alpha_i^{*j}-\alpha_i^j)\mathbf{x}_i^j = 0 \\ &\frac{\partial L}{\partial \xi_i^j} = C-\alpha_i^j-\beta_i^j = 0,\ \forall i,j \\ &\frac{\partial L}{\partial \xi_i^{*j}} = C-\alpha_i^{*j}-\beta_i^{*j} = 0,\ \forall i,j \\ &\frac{\partial L}{\partial b_j} = -\sum_{i=1}^{N^j}(\alpha_i^j+\gamma^j) + \sum_{i=1}^{N^{j+1}}(\alpha_i^{*j+1}+\gamma^{j+1}) = 0,\ \forall j \end{aligned} \tag{6}$$

Note, we implement the Karush–Kuhn–Tucker (KKT) conditions [28] of (4) in the above formulas. If $\mathbf{S}_W$ is

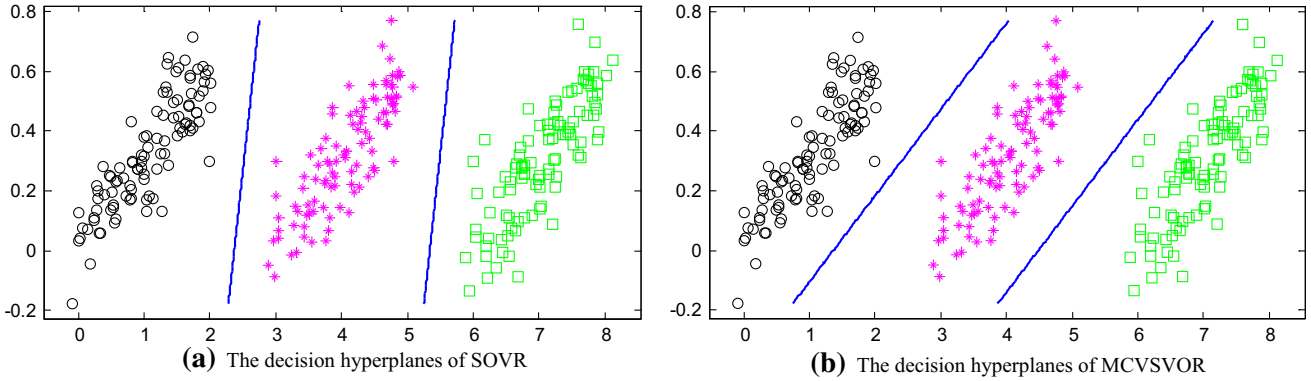


Fig. 1 Illustration of the decision hyperplanes generated by SOVR and MCVSVOR on a synthetic dataset

nonsingular, according to (6), the mapping direction $\mathbf{w}$ can be formulated as

$$\mathbf{w} = \mathbf{S}_W^{-1} \sum_{j=1}^{r} \sum_{i=1}^{N^j} (\alpha_i^{*j} - \alpha_i^j)\mathbf{x}_i^j \tag{7}$$

In the practical applications, as in MCVSVM, the matrix singularity problem may be encountered in MCVSVOR in that the inverse matrix of $\mathbf{S}_W$ is needed during the process of solving the problem (4). Once this singularity problem occurs, as in [22], ones can add a diagonal matrix to $\mathbf{S}_W$, i.e., $\mathbf{S}_W = \mathbf{S}_W + \rho\mathbf{I}$. Here $\rho > 0$ and $\mathbf{I}$ is an identity matrix. This technique that deals with the matrix singularity problem is referred to as regularization method [29]. It is difficult to directly obtain the optimum value of ρ. As determining other parameters in SVOR and MCVSVOR, we can estimate a suitable value of ρ through using the cross validation technique.

According to the KKT conditions of (4) and the formula (7), the Wolf dual problem of the primal problem (4) of MCVSVOR can be formulated as

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \sum_{j,i} \sum_{j',i'} (\alpha_i^{*j} - \alpha_i^j)(\alpha_{i'}^{*j'} - \alpha_{i'}^{j'})(\mathbf{x}_i^j)^T \mathbf{S}_W^{-1} \mathbf{x}_{i'}^{j'} - \sum_{j,i} (\alpha_i^{*j} + \alpha_i^j) \\
&s.t.\ 0 \le \alpha_i^j \le C,\ \forall i,j \\
&\quad 0 \le \alpha_i^{*j+1} \le C,\ \forall i,j \\
&\quad \sum_{i=1}^{N^j} \alpha_i^j + \gamma^j = \sum_{i=1}^{N^{j+1}} \alpha_i^{*j+1} + \gamma^{j+1},\ \gamma^j \ge 0, \forall j
\end{aligned} \tag{8}$$

where j runs over $1,\ldots,r-1$. This problem is similar to the dual optimization problem of SVOR and can be solved by using the same technique as in SVOR. Suppose that $\{\boldsymbol{\alpha},\boldsymbol{\alpha}^*,\boldsymbol{\gamma}\}$ solves the above optimization problem (8), then the mapping direction $\mathbf{w}$ is obtained with (7) and the discriminant function value of a new input vector $\mathbf{x}$ in MCVSVOR is given by

$$\begin{aligned}
f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} &= \left(\mathbf{S}_W^{-1} \sum_{j=1}^{r} \sum_{i=1}^{N^j} (\alpha_i^{*j} - \alpha_i^j)\mathbf{x}_i^j \right)^T \mathbf{x} \\
&= \sum_{j=1}^{r} \sum_{i=1}^{N^j} (\alpha_i^{*j} - \alpha_i^j)(\mathbf{x}_i^j)^T \mathbf{S}_W^{-1} \mathbf{x}
\end{aligned} \tag{9}$$

Further, the predictive ordinal rank of the input vector $\mathbf{x}$ can be determined by the following decision function

$$\min_i \arg\{i : f(\mathbf{x}) < b_i\} \tag{10}$$

Here, as in SVOR, b_j's are the thresholds and can be determined by the optimality conditions for the dual problem, which is discussed in detail in [13]. For the sake of completeness, here we offer the computation method of b_j's, and more details can be found in [13]. First, we set

$$\begin{aligned}
I_{0a}^j &= \{i \in \{1,\ldots,N^j\} : 0 < \alpha_i^j < C\} \\
I_{0b}^j &= \{i \in \{1,\ldots,N^{j+1}\} : 0 < \alpha_i^{*j+1} < C\} \\
I_1^j &= \{i \in \{1,\ldots,N^{j+1}\} : \alpha_i^{*j+1} = 0\} \\
I_2^j &= \{i \in \{1,\ldots,N^j\} : \alpha_i^j = 0\} \\
I_3^j &= \{i \in \{1,\ldots,N^j\} : \alpha_i^j = C\} \\
I_4^j &= \{i \in \{1,\ldots,N^{j+1}\} : \alpha_i^{*j+1} = C\} \\
I_0^j &= I_{0a}^j \cup I_{0b}^j, I_{up}^j = I_0^j \cup I_1^j \cup I_3^j, I_{low}^j = I_0^j \cup I_2^j \cup I_4^j
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
F_{up}^j(\mu_j) &= \begin{cases} f(\mathbf{x}_i^j) + 1 & if\ i \in I_{0a}^j \cup I_3^j \\ f(\mathbf{x}_i^{j+1}) - 1 & if\ i \in I_{0b}^j \cup I_1^j \end{cases} \\
F_{low}^j(\mu_j) &= \begin{cases} f(\mathbf{x}_i^j) + 1 & if\ i \in I_{0a}^j \cup I_2^j \\ f(\mathbf{x}_i^{j+1}) - 1 & if\ i \in I_{0b}^j \cup I_4^j \end{cases} \\
b_{low}^j &= \max\{F_{low}^i(\mu_j) : i \in I_{low}^j\} \\
b_{up}^j &= min\{F_{up}^i(\mu_j) : i \in I_{up}^j\}
\end{aligned} \tag{12}$$

Therefore, any value from the interval $b_j \in [B^j_{low}, B^j_{up}]$ can be viewed as the feasible value of the threshold b_j. Under the circumstances, the final value of b_j might encounter the non-uniqueness problem. In order to handle this problem, ones can determine the final value of b_j by simply taking

$$b_j = \frac{1}{2}(B^j_{low} + B^j_{up}) \tag{13}$$

where

$$B^j_{low} = \begin{cases} \tilde{B}^{j+1}_{low} & if\ \gamma^{j+1} > 0 \\ \tilde{B}^{j}_{low} & otherwise \end{cases} \quad and \quad B^j_{up} = \begin{cases} \tilde{B}^{j-1}_{up} & if\ \gamma^{j} > 0 \\ \tilde{B}^{j}_{up} & otherwise \end{cases} \tag{14}$$

Here $\tilde{B}^j_{low} = \max\{b^k_{low} : k = 1, \ldots, j\}$ and $\tilde{B}^j_{up} = \min\{b^k_{up} : k = j, \ldots, r-1\}$.

3.2 Connection to SVOR

By observing the optimization problems of MCVSVOR and SVOR, it is easy to find that they actually have a close relationship. Suppose the within-class scatter matrix $\mathbf{S}_W$ is nonsingular and let $\mathbf{P} = \mathbf{S}_W^{-\frac{1}{2}}$, we have $\mathbf{P}^T = (\mathbf{S}_W^{-\frac{1}{2}})^T = \mathbf{S}_W^{-\frac{1}{2}} = \mathbf{P}$ since $\mathbf{S}_W$ is invertible and symmetric. Therefore, the optimization problem (4) of MCVSVOR can be transformed into

$$\begin{aligned} &\min_{\mathbf{v},\mathbf{b},\xi,\xi^*} \frac{1}{2}\mathbf{v}^T\mathbf{v} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi^j_i + \xi^{*j}_i) \\ &s.t.\ \mathbf{v}^T\mathbf{y}^j_i - b_j \le -1 + \xi^j_i,\ \xi^j_i \ge 0,\ \forall i,j \\ &\quad \mathbf{v}^T\mathbf{y}^j_i - b_{j-1} \ge 1 - \xi^{*j}_i, \xi^{*j}_i \ge 0, \forall i,j \\ &\quad b_{j-1} \le b_j,\ \forall i,j \end{aligned} \tag{15}$$

where $\mathbf{y}^j_i = \mathbf{P}^T\mathbf{x}^j_i$ and $\mathbf{v} = \mathbf{P}^{-1}\mathbf{w}$. The above optimization problem is actually the same as the primal optimization problem of SVOR. This reveals the close relationship between SVOR and MCVSVOR. Further, it can be concluded that the MCVSVOR problem can be solved by using the existing SVOR software. Thus, the solution of MCVSVOR is easy to be computed.

4 The nonlinear case

In the nonlinear case, ones generally use the kernelization trick [24] to map the d-dimensional sample space into a high-dimensional feature space. In this way, a linear hyperplane in the feature space corresponds to a nonlinear hyperplane in the original sample space. Without loss of generality, the optimization problem of MCVSVOR in the feature space is defined as

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b},\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{S}^{\phi}_W\mathbf{w} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi^j_i + \xi^{*j}_i) \\ &s.t.\ \mathbf{w}^T\varphi(\mathbf{x}^j_i) - b_j \le -1 + \xi^j_i,\ \xi^j_i \ge 0,\ \forall i,j \\ &\quad \mathbf{w}^T\varphi(\mathbf{x}^j_i) - b_{j-1} \ge 1 - \xi^{*j}_i,\ \xi^{*j}_i \ge 0,\ \forall i,j \\ &\quad b_{j-1} \le b_j,\ \forall i,j \end{aligned} \tag{16}$$

where $\varphi(\mathbf{x}^j_i)$ denotes the sample in the feature space and $\mathbf{S}^{\varphi}_W$ is the corresponding within-class scatter matrix. Assume

$$\begin{aligned} \mathbf{X}^{\varphi} &= [\varphi(\mathbf{x}^1_1), \ldots, \varphi(\mathbf{x}^1_{N^1}), \varphi(\mathbf{x}^2_1), \ldots, \varphi(\mathbf{x}^2_{N^2}), \varphi(\mathbf{x}^r_1), \ldots, \varphi(\mathbf{x}^r_{N^r})] \\ &= [\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_N)] \end{aligned} \tag{17}$$

according to [30], $\mathbf{S}^{\varphi}_W$ can be further rewritten as

$$\mathbf{S}^{\varphi}_W = \mathbf{X}^{\varphi}\mathbf{L}(\mathbf{X}^{\varphi})^T \tag{18}$$

where $\mathbf{L} = \mathbf{I} - \mathbf{W}$. Here $\mathbf{I}$ is a identity matrix and $\mathbf{W}$ is defined as

$$\mathbf{W}_{ij} = \begin{cases} 1/N^k, & if\ \mathbf{x}_i\ and\ \mathbf{x}_j\ belong\ to\ the\ kth\ category \\ 0, & otherwise. \end{cases} \tag{19}$$

On the other hand, according to the representation theorem for Reproducing Kernel Hilbert Spaces [28], of which the vector $\mathbf{w}$ can be formulated as

$$\mathbf{w} = \sum_{i=1}^{N} a_i\varphi(\mathbf{x}_i) \tag{20}$$

where $a_i \in \mathbf{R}$.

Thus, according to the above discussion and by using (18) and (20), the optimization problem (16) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{a},\mathbf{b},\xi,\xi^*} \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{KLKa} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi^j_i + \xi^{*j}_i) \\ &s.t.\ \mathbf{a}^{\mathrm{T}}\mathbf{k}^j_i - b_j \le -1 + \xi^j_i,\ \xi^j_i \ge 0,\ \forall i,j \\ &\quad \mathbf{a}^{\mathrm{T}}\mathbf{k}^j_i - b_{j-1} \ge 1 - \xi^{*j}_i,\ \xi^{*j}_i \ge 0,\ \forall i,j \\ &\quad b_{j-1} \le b_j,\ \forall i,j \end{aligned} \tag{21}$$

where $\mathbf{K} = \left(k(\mathbf{x}_i, \mathbf{x}_j)\right)_{N\times N}$ is the kernel matrix, the vectors $\mathbf{k}^j_i$ and $\mathbf{a}$ are defined as $\mathbf{k}^j_i = [k(\mathbf{x}^j_i, \mathbf{x}_1), k(\mathbf{x}^j_i, \mathbf{x}_2), \ldots, k(\mathbf{x}^j_i, \mathbf{x}_N)]^T$ and $\mathbf{a} = [a_1, \ldots, a_N]^T$, respectively. Here $k(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j)$ is a predefined kernel function. Let $\mathbf{M} = \mathbf{KLK}$, the above optimization problem (21) can be further written as

$$\begin{aligned}
&\min_{\mathbf{a},\mathbf{b},\xi,\xi^*} \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{M}\mathbf{a} + C\sum_{j=1}^{r}\sum_{i=1}^{N^j}(\xi_i^j + \xi_i^{*j}) \\
&s.t.\ \mathbf{a}^{\mathrm{T}}\mathbf{k}_i^j - b_j \le -1 + \xi_i^j,\ \xi_i^j \ge 0,\ \forall i,j \\
&\quad \mathbf{a}^{\mathrm{T}}\mathbf{k}_i^j - b_{j-1} \ge 1 - \xi_i^{*j},\ \xi_i^{*j} \ge 0,\ \forall i,j \\
&\quad b_{j-1} \le b_j,\ \forall i,j
\end{aligned} \tag{22}$$

This is the final formulation of the optimization problem of the nonlinear MCVSVOR. It should be noted that the above optimization problem (22) actually is a optimization problem defined by linear MCVSVOR since $\mathbf{M} = \mathbf{KLK}$ is the within-class scatter matrix of the dataset which consists of $\mathbf{k}_i (i = 1, \ldots, N)$. So, according to the previous discussion about the linear MCVSVOR, it can be efficiently solved. Here, it is worthwhile to note that our method uses $\mathbf{a}^{\mathrm{T}}\mathbf{KLKa}$ as the regularization term in the nonlinear case. Actually, here $\mathbf{a}^{\mathrm{T}}\mathbf{KLKa}$ is employed in that it is derived form $\mathbf{w}^T\mathbf{S}_W^{\phi}\mathbf{w}$ in (16) by using (18) and (20). If ones used $\mathbf{a}^{\mathrm{T}}\mathbf{Ka}$ in (21), then it is reduced to the optimization problem of the kernelized SVOR. Besides, the matrix $\mathbf{M} = \mathbf{KLK}$ may be singular. This singularity problem can be tackled as the linear case in Sect. 3.1 since the problem (22) essentially formulates a linear MCVSVOR in which the training data is represented by $\mathbf{k}_i (i = 1, \ldots, N)$ and the matrix $\mathbf{M} = \mathbf{KLK}$ is its within-class scatter matrix.

Suppose $\{\mathbf{a}, \mathbf{b}, \xi, \xi^*\}$ is the solution of the above optimization problem, the discriminant function value of a new input vector $\mathbf{x}$ is

$$f(\mathbf{x}) = \mathbf{a}^T\mathbf{k} \tag{23}$$

where $\mathbf{k} = [k(\mathbf{x}_1, \mathbf{x}), k(\mathbf{x}_2, \mathbf{x}), \ldots, k(\mathbf{x}_N, \mathbf{x})]^T$. Thus, the predictive ordinal decision function is given by

$$\min_i \arg\{i : f(\mathbf{x}) < b_i\} \tag{24}$$

Here thresholds b_j's can be determined by the same strategy in Sect. 3.1.

5 Experiments

In the experiments, we first demonstrate the effectiveness of the proposed method on a synthetic dataset. Then, we conduct the experiments on a synthetic dataset with noise in which the data is non-separable. This experiments intuitionisticly illustrates the difference between SOVR and the proposed method MCVSVOR in the situation where a noisy data is encountered. After that, we conduct experiments on several benchmark datasets to evaluate its performance by comparing it with KDLOR and SVOR-EXC. Finally, we report the experimental results on several real datasets. Note, for the sake of fairness, SVOR with explicit threshold constraints (called SVOR-EXC) is considered since our method is with explicit threshold constraints.

5.1 Synthetic dataset

In order to evaluate the effectiveness of the proposed method, in this subsection we report the experimental result on a synthetic dataset. As is shown in Fig. 2, the dataset contains three ordinal categories, each ordinal category of which has 100 samples.

In the experiment, we used the Gaussian kernel, i.e. $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2)$. The experimental result is showed in Fig. 2. It can be found that the samples of different category can be ordinally arranged by the hyperplanes of the proposed method, i.e., the samples that have the same rank are arranged in same bin by the proposed method. The experimental result shows that the proposed method MCVSVOR is effective to handle the ordinal regression task.

5.2 Noisy dataset

In the above experiment, we validate the effectiveness of the proposed method on a synthetic dataset where the data is obviously separable. In order to further intuitionisticly demonstrate its ability of dealing with the data with noise, we create a synthetic dataset which contains noise and so the categories are not completely separable. As is shown in Fig. 3, the synthetic dataset also includes three ordinal categories and each category consists of 100 samples.

In this experiment, we adopted the liner kernel function, i.e. $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$. Figures 3a and b respectively show the decision hyperplanes generated by SOVR and MCVSVOR. It is easy to be observed that the decision hyperplanes of MCVSVOR reflect the characteristic of the data and show

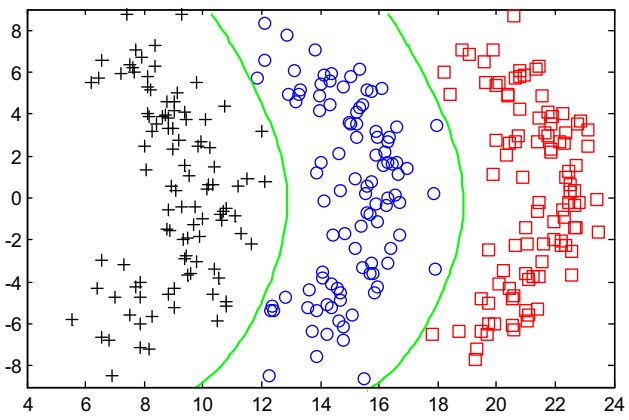
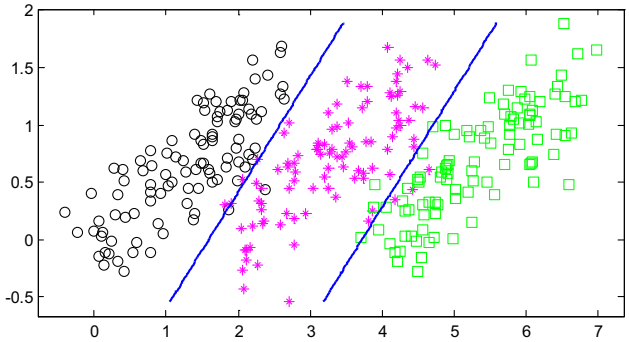
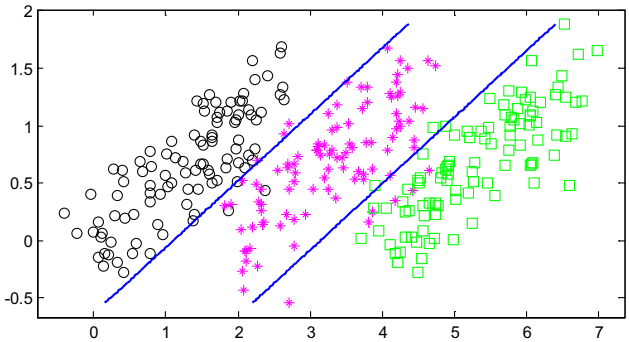


Fig. 2 Illustration of the decision hyperplanes generated by MCVSVOR



(a) Illustration of the decision hyperplanes generated by SVOR



(b) Illustration of the decision hyperplanes generated by MCVSVOR

Fig. 3 Illustration of the decision hyperplanes of SVOR and MCVSVOR on a synthetic data with noise

more reasonable in contrast with SOVR. This experimental results further verify the fact that incorporating the distribution information of the data in SOVR can improve its performance. The proposed method embodies this principle by using the within-class scatter matrix in SOVR.

5.3 Benchmark datasets

In order to assess the performance of MCVSVOR, we conducted the experiments on several benchmark datasets. These datasets were frequently used to test the ordinal regression methods, for example, in [13, 22]. In this section we report the experimental results. Table 1 shows a summary of the characteristics of the selected datasets. For each dataset, the target values were discretized into ten ordinal quantities through using equal-frequency binning. In the experiments, each dataset was randomly partitioned into training/test splits as specified in Table 1. Each attribute of the samples were scaled to 0 mean and 1 variance.

On evaluating the ordinal regression methods, there are generally two metrics to be used to quantify the accuracy of predicted values with respect to true targets [13, 22]. One of which is mean-absolute-error (MAE) which measures how far the predicted ordinal scales of the samples differ from their true targets and is formulated as $\frac{1}{N}\sum_{i=1}^{N}|y_i - \tilde{y}_i|$,

Table 1 Characteristics of the selected benchmark datasets

Datasets	No. of attributes	No. of training samples	No. of test samples
Pyrimidines	27	50	24
Machine CPU	6	150	59
Boston housing	13	300	206
Abalone	8	1000	3177
Bank	32	3000	5182
Computer	21	200	192
California	8	5000	15,640
Census	16	6000	16,784

where y_i and $\tilde{y}_i$ are respectively the predicted ordinal scales and the true targets, and $|\cdot|$ denotes the absolute operation. The other is mean-zero–one-error (MZE). It measures the classification error of the samples and is defined as $\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \tilde{y}_i)$. Here y_i and $\tilde{y}_i$ are respectively the predicted rank of the respective method and the true rank, and $I(\cdot)$ is an indicator function that takes 1 if $y_i \neq \tilde{y}$ and returns 0 otherwise.

In the experiments, we employed the Gaussian kernel, which is formulated as $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2)$. We employed fivefold cross validation to choose the appropriate values for the relevant parameters (i.e. the Gaussian

Table 2 The experimental results on the selected benchmark datasets

Datasets	Mean-zero–one-error			Mean-absolute-error		
	KDLOR	SVOR-EXC	MCVSVOR	KDLOR	SVOR-EXC	MCVSVOR
Pyrimidines	0.739 ± 0.050	0.752 ± 0.063	**0.713 ± 0.052**	1.100 ± 0.100	1.331 ± 0.193	**1.003 ± 0.089**
MachineCPU	**0.480 ± 0.010**	0.661 ± 0.056	0.491 ± 0.027	**0.690 ± 0.015**	0.986 ± 0.127	0.694 ± 0.071
Boston	0.560 ± 0.020	0.569 ± 0.025	**0.548 ± 0.032**	0.700 ± 0.035	0.773 ± 0.049	**0.692 ± 0.014**
Abalone	0.740 ± 0.020	**0.736 ± 0.011**	0.738 ± 0.031	1.400 ± 0.050	**1.391 ± 0.021**	1.397 ± 0.005
Bank	0.745 ± 0.002	0.744 ± 0.005	**0.739 ± 0.016**	1.450 ± 0.020	1.512 ± 0.017	**1.445 ± 0.014**
Computer	0.472 ± 0.020	0.462 ± 0.005	**0.458 ± 0.025**	0.601 ± 0.025	0.602 ± 0.009	**0.597 ± 0.031**
California	0.643 ± 0.005	0.640 ± 0.003	**0.639 ± 0.001**	0.907 ± 0.004	1.068 ± 0.005	**0.898 ± 0.003**
Census	0.711 ± 0.020	0.699 ± 0.002	**0.688 ± 0.013**	1.213 ± 0.003	1.270 ± 0.007	**1.206 ± 0.005**

Bold values in each row denote the best performance

kernel parameter γ and the regularization factor C, which were chose from the sets $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}\} \times \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$) involved in the problem formulation of each method. The final test error of each method was obtained by using the selected parameters for it. The experiment on each dataset was repeated 20 times independently.

The experimental results are reported in Table 2. In comparison with KDLOR and SVOR-EXC, the proposed method has the lowest MZE and MAE on the whole. These indicate that it is competitive with the other two methods in generalization ability. The reason is that in MCVSVOR not only the distribution of the categories is explicitly considered but also the large margin principle is embodied. Actually, KDLOR also takes the distribution characteristic of the data into consideration, and so it performs better than SVOR-EXC on the whole. However, it does not explicitly embody the large margin principle but MCVSVOR does.

In order to further investigate the statistical significance, we performed the paired two-tailed t-tests [31] according to MZE of these methods. The smaller the p value of a t test, the more significant the difference of the two average values is. T test usually takes a p value of 0.05 as a typical threshold which is considered statistically significant. Table 3 reports the experimental results of the t-tests. For example, the p value of the t test is 0.0207 (<0.05) when comparing KDLOR and SVOR-EXC on the Pyrimidines dataset. This means that KDLOR performs significantly better than SVOR-EXC on this dataset at the 0.05 significant level. From Table 3, although the proposed method MCVSVOR has on the whole better MZE, compared with KDLOR, its improvement is not significant. However, MCVSVOR significantly outperforms SVOR-EXC on five of eight datasets.

We further investigated the computational cost of several methods. Table 4 reports the results. It can be observed that KDLOR has the least time consumption on the whole. The proposed method need more time consumption in contrast with the other two methods. The reason is that the inversion of $\mathbf{S}_W$ is necessary when solving the solving the optimization (8) of the proposed method. So, we need to further research more efficient method to solve the optimization problem of the proposed method in the future.

Table 3 P-value of t-test on the selected benchmark datasets

Datasets	KDLOR/ SVOR-EXC	MCVSVOR/ SVOR-EXC	MCVSVOR/ KDLOR
Pyrimidines	**0.0207**	**0.0076**	**0.0271**
MachineCPU	**0.0161**	**0.0391**	0.0942
Boston	0.0611	**0.0113**	**0.0246**
Abalone	0.8913	0.1743	0.5314
Bank	0.0878	**0.0426**	0.2138
Computer	0.1632	0.6236	0.3492
California	0.3215	0.3218	0.0872
Census	**0.0139**	**0.0573**	**0.0164**

Bold values in each row mean that the p-values are smaller that 0.05

Table 4 The computational cost (in seconds) on the selected benchmark datasets

Datasets	SVOR-EXC	KDLOR	MCVSVOR
Pyrimidines	0.70	0.71	0.85
MachineCPU	0.92	1.02	1.26
Boston	2.09	2.25	2.41
Abalone	18.93	17.19	19.86
Bank	158.26	149.47	179.59
Computer	331.37	300.19	450.93
California	697.82	567.64	728.36
Census	5357.38	4939.41	6891.52

5.4 Real datasets

In order to further evaluate the performance of the proposed method, we conducted the experiments on several real datasets including USPS [32], UMIST [33], FG-NET

Table 5 The experimental results on the real datasets

Datasets	Mean-zero–one-error			Mean-absolute-error		
	KDLOR	SVOR-EXC	MCVSVOR	KDLOR	SVOR-EXC	MCVSVOR
USPS	0.093 ± 0.264	0.107 ± 0.463	**0.088** ± 0.476	1.718 ± 0.253	1.821 ± 0.193	**1.693** ± 0.217
UMIST	0.653 ± 0.225	0.699 ± 0.362	**0.572** ± 0.429	0.831 ± 0.352	1.013 ± 0.729	**0.729** ± 0.241
FG-NET	0.274 ± 0.428	**0.233** ± 0.211	0.238 ± 0.315	2.963 ± 0.375	**2.492** ± 0.293	2.502 ± 0.598
Mixed-gambles	0. 501 ± 0.327	0.491 ± 0.269	**0. 469** ± 0.412	2.451 ± 0.375	2.217 ± 0.417	**2.212** ± 0.342

Bold values in each row denote the best performance

[8] and Mixed-Gambles [34]. The USPS dataset comprises 11,000 grayscale handwritten digit images and each image has a resolution with 16 × 16. The dataset is divided into ten ranked categories (from 0 to 9) and each category consists of 1100 images. In this experiment, our aim is to rank the data. The UMIST dataset is a multi-view face dataset. It contains 564 images of 20 individuals and the images of each individual change from profile to frontal views. We select six consecutive ordinal interval angles for ordinal regression. The FG-NET dataset is widely used for evaluating the age estimation methods. It has 1002 face images of 82 individuals. In this dataset, the age varies from 0 to 69 years. Note, the age distributions are imbalanced in the dataset. So, we divided the ages into five ranges: 0–9, 10–19, 20–29, 30–49, and 50+. The age label is ordinally denoted by 1,…, 5, respectively. Here we use the AAM algorithm [35] to extract the used features. The Mixed-Gambles task is a functional MRI dataset. In this dataset, there are 16 different rank levels. Similar to [7, 36], we also only consider gain levels and adopt the GLM regression coefficients as the used features.

For each dataset, we randomly selected 40 % samples from each category for training and the rest for testing. The relevant parameters were determined by the strategy used in Sect. 5.2. Finally, we report the averaged results over 30 runs.

Table 5 shows the ranking results of several methods on the datasets. It can be found that KDLOR and SVOR-EXC achieve comparable performance on the whole. However, the proposed method performs better in comparison with KDLOR and SVOR-EXC on three of the four datasets. The reason is, in our opinion, that MCVSVOR embodies the large margin principle as SVOR-EXC and simultaneously incorporates the distribution information of the categories as well as KDLOR.

6 Conclusions

In this paper, we proposed a novel ordinal regression method. This method stems from MCVSVM and is called MCVSVOR. In contrast with SVOR, it takes full use of the distribution information of the categories. At the same time, MCVSVOR embodies the large margin principle as well as SVOR. Therefore, MCVSVOR achieves better generalization performance compared with SVOR and KDLOR. The comprehensive experimental results suggest that MCVSVOR is effective to handle the ordinal regression problems and can obtain better generalization performance over SVOR and KDLOR.

Additionally, similar to SVOR, the proposed method can also be extended to the case where the threshold constraints are implicit. Moreover, according to the discussion in Sect. 3.2, MCVSVOR can be efficiently solved using the existing SVOR software and so its solution is easy to be computed. However, compared with SVOR, the proposed method is time-consuming because the inverse matrix of the within-class scatter matrix $\mathbf{S}_W$ is necessary in solving the optimization problem (4). Therefore, how to accelerate the proposed method is another important research topic.

Acknowledgments This work is supported in part by the Scientific Research Project "Chunhui Plan" of Ministry of Education of China (Grant Nos. Z2015102, Z2015108), the Key Scientific Research Foundation of Sichuan Provincial Department of Education (No. 11ZA004), the Sichuan Province Science and Technology Support Program (No. 2016RZ0051), the Open Research Fund from Province Key Laboratory of Xihua University (No. szjj2013-022) and the National Science Foundation of China (Grant Nos. 61303126, 61103168).

Compliance with ethical standards

Conflict of interest The authors have declared that there is no conflict of interests regarding the publication of this article.

References

1. You ZH, Lei YK, Zhu L, Xia JF, Wang B (2013) Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinform 14(8):69–75
2. You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP, Ji Z (2014) Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. BMC Bioinform 15(Suppl 15):S9–S9
3. You ZH, Yu JZ, Zhu L, Li S, Wen ZK (2014) A mapreduce based parallel SVM for large-scale predicting protein-protein interactions. Neurocomputing 145(18):37–43

4. Kundu MK, Chowdhury M, Banerjee M (2012) Interactive image retrieval using M-band wavelet, earth mover's distance and fuzzy relevance feedback. J Mach Learn Cybern 3(4):285–296
5. Yan J, Liu N, Yan SC, Yang Q, Fan WG, Wei W, Chen Z (2011) Trace-oriented feature analysis for large-scale text data dimension reduction. IEEE Trans Knowl Data Eng 23(7):1103–1117
6. Pedregosa F, Gramfort A, Varoquaux G, Cauvet E, Pallier C, Thirion B (2012) Learning to rank from medical imaging data. In: Proceedings Int. Workshop Mach. Learn. Med. Imag, pp 234–241
7. Li C, Liu Q, Liu J, Lu H (2015) Ordinal distance metric learning for image ranking. IEEE Trans Neural Netw Learn Syst 26(7):1551–1559
8. Han H, Otto C, Liu X, Jain AK (2015) Demographic estimation from face images: human vs. machine performance. IEEE Trans Pattern Anal Mach 37(7):1148–1161
9. McCullagh P (1980) Regression models for ordinal data. J R Stat Soc B 42(2):109–142
10. McCullagh P, Nelder A (1983) Generalized linear models. Chapman & Hall, London
11. Johnson VE, Albert JH (1999) Ordinal data modeling (statistics for social science and public policy). Springer, New York
12. Chu W, Ghahramani Z (2005) Gaussian processes for ordinal regression. J Mach Learn Res 6:1019–1041
13. Chu W, Keerthi SS (2007) Support vector ordinal regression. Neural Comput 19(3):792–815
14. Wang XZ, Ashfaq RAR, Fu AM (2015) Fuzziness based sample categorization for classifier performance improvement. J Intell Fuzzy Syst 29(3):1185–1196
15. Wang XZ (2015) Learning from big data with uncertainty-editorial. J Intell Fuzzy Syst 28(5):2329–2330
16. Mika S (2002) Kernel fisher discriminants (PhD thesis). University of Technology, Berlin
17. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
18. Zafeiriou S, Tefas A, Pitas I (2007) Minimum class variance support vector machines. IEEE Trans Image Process 16(10):2551–2564
19. Wang M, Chung FL, Wang ST (2010) On minimum class locality preserving variance support vector machine. Pattern Recogn 43(8):2753–2762
20. Herbrich R, Graepel T, Obermayer K (1999) Support vector learning for ordinal regression. In: International conference on artificial neural networks, pp 97–102
21. Kramer S, Widmer G, Pfahringer B, DeGroeve M (2001) Prediction of ordinal classes using regression trees. Fundamenta Informaticae 47:1–13
22. Sun BY, Li J, Wu DD, Zhang XM, Li WB (2010) Kernel discriminant learning for ordinal regression. IEEE Trans Knowl Data Eng 22(6):906–910
23. Seah CW, Tsang IW, Ong YS (2012) Transductive ordinal regression. IEEE Trans Neural Netw Learn Syst 23(7):1074–1086
24. Scholkopf B, Smola A (2002) Learning with kernels. MIT, Cambridge
25. Shashua A, Levin A (2003) Ranking with large margin principle: two approaches. Adv Neural Inf Process Syst 15:961–968
26. Shevade SK, Chu W (2006) Minimum enclosing spheres formulations for support vector ordinal regression. In: Sixth international conference on data mining, pp 1054–1058
27. Zhao B, Wang F, Zhang CS (2009) Block-quantized support vector ordinal regression. IEEE Trans Neural Netw 20(5):882–890
28. Fletcher R (1987) Practical methods of optimization, 2nd edn. Wiley, New York
29. Guo Y, Hastie T, Tibshirani R (2007) Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8(1):86–100
30. He XF, Yan SC, Hu YX, Niyogi P, Zhang HJ (2005) Face recognition using laplacianfaces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
31. Alpaydin E (2004) Introduction to machine learning. The MIT, Cambridge
32. Hull JJ (1994) A database for handwritten text recognition research. IEEE Trans Pattern Anal Mach Intell 16(5):550–554
33. Graham DB, Allinson NM (1998) Characterizing virtual eigensignatures for general purpose face recognition. In: Face recognition: from theory to applications, pp 446–456
34. Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neuralbasis of loss aversion in decision-making under risk. Science 315(5811):515–518
35. Cootes T, Edwards G, Taylor C (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):68–685
36. Pedregosa F, Gramfort A, Varoquaux G, Cauvet E, Pallier C, Thirion B (2012) Learning to rank from medical imaging data. In: Proc. Int. Workshop Mach. Learn. Med. Imag, pp 234–241