Neural Networks 98 (2018) 114–121

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A regularization path algorithm for support vector ordinal regression

Bin Gu *

Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, PR China
School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, PR China

ARTICLE INFO

Article history:
Received 12 January 2017
Received in revised form 18 September 2017
Accepted 7 November 2017
Available online 16 November 2017

Keywords:
Solution path
Support vector ordinal regression
QR decomposition
Singularity

ABSTRACT

Support vector ordinal regression (SVOR) is a popular method for tackling ordinal regression problems. Solution path provides a compact representation of optimal solutions for all values of regularization parameter, which is extremely useful for model selection. However, due to the complicated formulation of SVOR (including multiple equalities and extra variables), there is still no solution path algorithm proposed for SVOR. In this paper, we propose a regularization path algorithm for SVOR which can track the two sets of variables of SVOR w.r.t. the regularization parameter. Technically, we use the *QR* decomposition to handle the singular matrices in the regularization path. Experiment results on a variety of datasets not only confirm the effectiveness of our regularization path algorithm, but also show the superiority of our regularization path algorithm on model selection.

1. Introduction

Ordinal regression (OR) is an important learning task in machine learning which widely exists in real-world applications, such as collaborative filtering (Shashua & Levin, 2002), information retrieval (Herbrich, Graepel, & Obermayer, 1999), and flight delays forecasting (McCrea, Sherali, & Trani, 2008). In OR problems, training samples are labeled by a set of ranks which exhibit an ordering among different categories. Take collaborative filtering as an example, the rating that a customer assesses a movie might be one of do-not-bother, only-if-you-must, good, very-good, and run-to-see. The ratings have a natural order. Thus, OR is distinguished from traditional multiple classification problems.

Support vector ordinal regression (SVOR) is a popular method for tackling OR problems (Shashua & Levin, 2002). There are several versions of SVOR. All of them are to find multiple parallel discrimination hyperplanes. Specifically, Shashua and Levin (2002) first proposed two versions of SVOR based on the fixed-margin principle and the sum-of-margins principle, respectively. Later, Chu and Sathiya Keerthi (2007) improved the fixed-margin based SVOR by explicitly and implicitly keeping the partial order of the multiple parallel discrimination hyperplanes. Gu, Sheng, Tay, Romano, and Li (2015) proposed a modified version of the sum-of-margins based SVOR for the incremental learning. In this paper, we mainly focus on the most popular one (the fixed-margin based SVOR by the explicit constraints), and call it SVOR for the sake of convenience.

Assume that the samples in OR problem are labeled by a set of ranks $Y = \{1, 2, \ldots, r\}$. The number of training samples in the jth rank ($j \in Y$) is denoted as n^j, and the ith training sample is denoted as x_i^j ($x_i^j \in X$, where X is the input space with $X \subset \mathbb{R}^d$). Chu and Sathiya Keerthi (2007) proposed the primal problem of SVOR as follows.

$$\begin{aligned}
\min_{w,b,\epsilon,\epsilon^*} \quad & \frac{1}{2}\langle w, w\rangle + C\sum_{j=1}^{r-1}\left(\sum_{i=1}^{n^j}\epsilon_i^j + \sum_{i=1}^{n^{j+1}}\epsilon_i^{*j+1}\right) \\
s.t. \quad & \langle w, \phi(x_i^j)\rangle - b_j \le -1 + \epsilon_i^j, \\
& \epsilon_i^j \ge 0, \quad i = 1, \ldots, n^j, \\
& \langle w, \phi(x_i^{j+1}) - b_j\rangle \ge 1 - \epsilon_i^{*j+1}, \\
& \epsilon_i^{*j+1} \ge 0, \quad i = 1, \ldots, n^{j+1}, \\
& b_j \le b_{j+1}, \quad j = 1, \ldots, r-1
\end{aligned} \tag{1}$$

where training samples x_i^j are mapped into a high dimensional reproducing kernel Hilbert space (RKHS) (Schölkopf & Smola, 2001) by the transformation function ϕ. We have the kernel function $K(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$ with $\langle\cdot,\cdot\rangle$ denoting inner product in RKHS. Furthermore, ϵ_i^j (ϵ_i^{*j+1}) is a non-negative slack variable measuring the degree of misclassification of the data x_i^j (x_i^{j+1}). The regularization parameter $C \in [0, \infty)$ controls the trade-off between the errors in the training samples and the margin, which is usually tuned by model selection.

Solution path provides a compact representation of optimal solutions for all values of regularization parameter, which is extremely useful for model selection (Gu & Ling, 2015). There

* Correspondence to: Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, PR China.
E-mail address: jsgubin@nuist.edu.cn.

https://doi.org/10.1016/j.neunet.2017.11.008
0893-6080/

Table 1
Representative solution path algorithms. (BC, R and OR are the abbreviations of binary classification, regression and ordinal regression, respectively. EV denotes extra variables.)

Problem	Task	Reference	Parameter	Robust	Equalities	EV
C-SVC	BC	Hastie et al. (2004)	Regularization parameter C	No	One	No
$2C$-SVC	BC	Bach et al. (2006) and Gu et al. (2017)	Regularization parameters C_+, C_-	No	One	No
ε-SVR	R	Gunter and Zhu (2007)	Regularization parameter	No	One	No
ε-SVR	R	Wang et al. (2008)	Regularization parameter and ε	No	One	No
Lasso	R	Rosset and Zhu (2007)	Regularization parameter	No	Zero	No
KQR	R	Takeuchi et al. (2009)	Quantile order $\tau \in (0, 1)$	No	Zero	No
KQR	R	Li et al. (2007)	Regularization parameter	No	Zero	No
C-SVC	BC	Karasuyama and Takeuchi (2011)	Regularization parameter C	No	One	No
ν-SVC	BC	Gu et al. (2012)	Regularization parameter ν	No	Two	No
C-SVC	BC	Dai et al. (2013) and Ong et al. (2010)	Regularization parameter C	Yes	One	No
C-SVC	BC	Sentelle et al. (2016)	Regularization parameter C	Yes	One	No
ν-SVC	BC	Gu and Sheng (2017)	Regularization parameter ν	Yes	Two	No
SVOR	OR	Our	Regularization parameter C	Yes	Multiple	Yes

have been a lot of solution path algorithms proposed for several learning algorithms. For example, Hastie, Rosset, Tibshirani, and Zhu (2004) proposed a solution path approach for C-SVC. Bach, Heckerman, and Horvitz (2006) and Gu, Sheng, Tay, Romano, and Li (2017) proposed a solution path algorithm and a solution surface algorithm, respectively, for $2C$-SVC. Gunter and Zhu (2007) and Wang, Yeung, and Lochovsky (2008) proposed the solution path algorithms for ε-SVR to trace the solutions with respect to ε and the regularization parameter, respectively. Rosset and Zhu (2007) proposed a solution path for Lasso. Gu, Wang, Zheng, and Yu (2012) proposed a robust solution path algorithm for ν-SVC. Li, Liu, and Zhu (2007) and Takeuchi, Nomura, and Kanamori (2009) proposed the solution path algorithms for kernel quantile regression (KQR) to trace the solutions with respect to the regularization parameter and the quantile order $\tau \in (0, 1)$, respectively. Karasuyama and Takeuchi (2011) proposed an approximate solution path for C-SVC. Because computing an inverse matrix is needed for each iteration, the solution path algorithm will interrupt when the key matrix is singular. To address this issue, several robust solution path algorithms were proposed. For example, Dai, Chang, Mai, Zhao, and Xu (2013), Ong, Shao, and Yang (2010) and Sentelle, Anagnostopoulos, and Georgiopoulos (2016) proposed improved solution path algorithms to handle the singular matrices encountered in the method of Hastie et al. (2004). Gu and Sheng (2017) proposed a robust solution path algorithm for ν-SVC. We summarize the above algorithms in Table 1.

From Table 1, we find that existing solution path algorithms are mainly designed for binary classification and regression problems. However, SVOR essentially solves multiple binary classifications problems as mentioned in (1). More importantly, SVOR has a more complicated formulation than C-SVC, Lasso, KQR. Specifically, the existing solution path algorithms for binary classification and regression problems solve the convex optimization problem with zero, one or two equalities. However, the dual formulation of SVOR has multiple equalities and extra variables as shown in (2). Due to the complications in the formulation of SVOR as discussed above, there is still no solution path algorithm proposed for SVOR. To address this issue, in this paper, we propose a regularization path algorithm for SVOR (called RP-SVOR) which can track the two sets of variables of SVOR w.r.t. the regularization parameter C. Particularly, the robust solution path algorithm is implemented by using the QR decomposition to handle the singular matrices. Experiment results on a variety of datasets not only confirm the effectiveness of RP-SVOR, but also show the advantage of RP-SVOR for model selection.

The remainder of the paper is organized as follows. Section 2 presents the dual formulation of SVOR and its Karush–Kuhn–Tucker (KKT) conditions. Our RP-SVOR is presented in Section 3. In Section 4, we present the experiment results. We draw conclusions in Section 4.

Notations: To make the notations easier to follow, we give a summary of the notations in the following list.

- α_i, g_i The ith element of vector α and g.
- x_i^j The ith sample of the training samples for the jth rank.
- Δ The amount of the change of each variable.
- $H_{S_S S_S}$ The submatrix of H with the rows and columns indexed by S_S.
- $\boxed{\mu^j}$ If $j \in J$, $\boxed{\mu^j}$ stands for μ^j, where the active set $J \subseteq \{2, \ldots, r-1\}$ is defined in Section 2.2. Otherwise, they will be ignored, i.e., $\boxed{\mu^j} = 0$.
- $u_{S_S^j}$ A $|S_S|$-dimensional column vector with all zeros except that the positions corresponding to the samples (x_i, y_i) of S_S^j are equal to $-y_i$, respectively.
- e_j A $(r-1)$-dimensional column vector with all zeros except that the jth and $(j+1)$th elements are equal to 1 and -1, respectively.

2. SVOR

In this section, we first present the dual formulation of SVOR, and then give the corresponding KKT conditions.

2.1. Dual formulation of SVOR

If directly solving the problem (1), we need to handle a complicated Lagrange problem with the primal variables w, b, ϵ and ϵ^*, and the Lagrange variables with the size of $2 \times \sum_{j=1}^{r} n^j - n^1 - n^r + r - 2$ corresponding to the inequality constraints in (1). To avoid the complicated Lagrange problem, we turn to solve a compact and equivalent problem, i.e., the dual problem of SVOR.

Before giving the dual problem, we first introduce an extended training sample set. Based on the reduction framework of Li and Lin (2007), SVOR can be regarded as $r-1$ binary classification problems. Thus, we define a two-class training sample set $S^j = \{(x_i^j, y_i^j = -1)\}_{i=1}^{n^j} \cup \{(x_i^{j+1}, y_i^{j+1} = +1)\}_{i=1}^{n^{j+1}}$, and an extended training sample set $S = \bigcup_{j=1}^{r-1} S^j = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, where $l = 2 \times \sum_{j=1}^{r} n^j - n^1 - n^r$.

Based on the extended training sample set S, the dual formulation of (1) can be formulated as follows Chu and Sathiya Keerthi (2007).

$$\min_{0 \le \alpha \le C, \mu \ge 0} \quad \frac{1}{2}\alpha H \alpha^T - \sum_{j=1}^{r-1} \sum_{i \in S^j} \alpha_i \tag{2}$$
$$s.t. \quad \sum_{i \in S^j} y_i \alpha_i = \mu^j - \mu^{j+1},$$