A Novel Ordinal Regression Method with Minimum Class Variance Support Vector Machine

Jinrong Hu[1, 2, a], Xiaoming Wang [1] and Zengxi Huang [1]

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

[2]Key Laboratory of Pattern Recognition and Intelligent Information Processing, Chengdu University, Chengdu 610106, China

[a]dewhjr@hotmail.com

Keywords: Machine learning, Ordinal regression, Support vector machine, Support vector ordinal regression.

Abstract. In the paper, we propose a novel ordinal regression method called minimum class variance support vector ordinal regression (MCVSVOR). MCVSVOR is derived from minimum class variance support vector machine (MCVSVM) which is a variant of SVM, and so inherits the latter's characteristics such as taking the distribution of the categories into consideration and good generalization performance. Finally, the experimental results validate the effectiveness of MCVSVOR and indicate its superior generalization performance over SVOR.

1. Introduction

In the practical applications of machine learning, a situation is frequently involved, i.e. exhibiting an order among the different categories. This type of supervised learning problems is referred to as ordinal regression which predicts categories of ordinal scale [2-4]. Different from traditional metric regression problems, its grades are usually discrete and finite. Also, it differs from traditional classification problems in that there is an ordinal relationship among different classes. In fact, ordinal regression shows resemblance to both regression and classification because labels are discrete and ordinal [12].

In the past decade, many methods have been proposed to deal with the ordinal regression problems [1, 9, 13]. Support vector ordinal regression (SVOR) is a powerful method which is designed to tackle the ordinal regression problems and originated in support vector machine (SVM). However, SVM is actually a local method in the sense that solution is exclusively determined by support vectors whereas all other data points are irrelevant to the decision hyperplane, i.e., the SVM solution does not take into consideration the distribution of the categories and may result in a non-robust solution [16]. In order to overcome the drawback of SVM, a modified class of SVM called minimum class variance support vector machine (MCVSVM) was presented in [16]. This method is inspired from the optimization of Fisher's discriminant ratio [5]. Similar to SVM, MCVSVM implements the large margin principle [15]. However, unlike SVM, the solution of MCVSVM takes into consideration both the samples in the boundaries and the distribution of the categories and gives a robust solution.

In this paper, we propose a novel ordinal regression learning method called minimum class variance ordinal regression (MCVSVOR) in which the distribution of the categories is explicitly considered and the large margin principle is embodied. Following the basic idea of SVOR, we define the MCVSVOR optimization problem. Since MCVSVOR is derived from MCVSVM, it inherits the latter's characteristics such as taking fully the distribution of the categories into consideration and embodying the large margin principle. At the same time, we also develop the linear and nonlinear cases of MCVSVOR and analyze the relationship between MCVSVOR and SVR. The relationship shows that MCVSVOR can be solved using the existing SVOR software, which makes the solution easy to be computed. Finally, the experimental results indicate that MCVSVOR is effective and can get superior generalization performance over SVOR.

2. Related work

In this paper, we consider an ordinal regression problem with r ordered categories which are denoted by consecutive integers $Y=\{1,\cdots,r\}$ to keep the known rank information. The training dataset is represented by $D=\{(\mathbf{x}_i^j, \mathrm{y}^j) \mid \mathbf{x}_i^j \in R^d, \mathrm{y}^j \in Y\}$, where $\mathbf{x}_i^j$ refers to the *i*th sample in the *j*-th category, and y^j represents the corresponding rank of the input data point $\mathbf{x}_i^j$. Here d refers to the dimensionality of sample vector. The dataset composed of $N=\sum_{j=1}^{r} N^j$ sample points. Here N^j is the number of training samples in the *j*-th category. And, we set $\mathbf{X}=[\mathbf{x}_1^1,\cdots,\mathbf{x}_{N^1}^1,\mathbf{x}_1^2,\cdots,\mathbf{x}_{N^2}^2,\cdots,\mathbf{x}_1^r,\cdots,\mathbf{x}_{N^r}^r]=[\mathbf{x}_1,\cdots,\mathbf{x}_N]$.

2.1 Support vector ordinal regression

The task of ordinal regression is to compute a function $f: R \to \{1,\cdots,r\}$ such that $f(\mathbf{x}_i^j)=y^j$ [10, 11]. Moreover, SVOR aims at finding $r-1$ parallel discriminant hyperplanes $\mathbf{w}^T\mathbf{x}-b_j=0\ (j=1,\cdots,r)$ that separate the data points of different ranks. So, the following optimization problem is defined [3,4]

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

Where $j=1\cdots,r$ and $i=1,\cdots,N^j$. Here, two auxiliary variables $b_0=-\infty$ and $b_r=+\infty$ are introduced. Note, SVM implements the large margin principle [14]. So, SVOR also embodies the principle since it is derived from SVM.

2.2 Kernel discriminant learning for ordinal regression

For the above given training dataset, the within-class scatter matrix $\boldsymbol{S}_W$ is defined as [5, 12]

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

Where $\mathbf{X}^j=\{\mathbf{x}_i^j \mid y^j=j, i=1,\cdots,N^j\}$, $\mathbf{u}^j=\frac{1}{N^j}\sum_{\mathbf{x}\in\mathbf{X}^j}\mathbf{x}$ is the mean sample vector of $\mathbf{X}^j$, and T denotes vector transpose. Here, N^j is the cardinality of $\mathbf{X}^j$. KDLOR defines the following optimization [12]

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

3. Minimum class variance support vector ordinal regression

Following the idea of SVOR, in the linear case we define the primal optimization problem of MCVSVOR as follows

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

Where $j=1\cdots,r$, $i=1,\cdots,N^j$, and $\mathbf{S}_W$ is the within-class scatter matrix which is defined as (2). Similar to MCVSVM, by this way, the distribution of the categories is taken fully into consideration.

Besides, the proposed method embodies the large margin principle since it is derived from MCVSVM which implements large margin principle [15]. So, it is different from KDLOR although they both take the distribution of the categories into consideration.

Similar to SVOR, the primal optimization problem of MCVSVOR can be efficiently solved by its dual optimization problem. Obviously, (4) is a quadratic programming problem. The primal Lagrangian (4) is

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

Where the vectors $\boldsymbol{\alpha} = [\alpha_1^1, \cdots, \alpha_{N^r}^r]^T$, $\boldsymbol{\alpha}^* = [\alpha_1^{*1}, \cdots, \alpha_{N^r}^{*r}]^T$, $\boldsymbol{\beta} = [\beta_1^1, \cdots, \beta_{N^r}^r]^T$, $\boldsymbol{\beta}^* = [\beta_1^{*1}, \cdots, \beta_{N^r}^{*r}]^T$ and $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_r]^T$ are the Lagrangian multipliers for the constraints of (4). By differentiating with respect to $\mathbf{w}$, $\boldsymbol{\xi}$, $\boldsymbol{\xi}^*$ and $\mathbf{b}$ and using the Karush-Kuhn-Tucker (KKT) conditions, the following holds

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

If the matrix $\mathbf{S}_W$ is nonsingular or invertible, we have

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

As in MCVSVM and KDLOR, MCVSVOR may encounter the singularity problem of $\mathbf{S}_W$ since its inverse matrix is necessary, which often occurs in the case where the number of samples is smaller than the dimensionality of the samples. To solve this singularity problem, similar to KDLOR, we can employ the regularization method [5, 6, 7] which is to add a constant $\rho > 0$ to the diagonal elements of $\mathbf{S}_W$ as $\mathbf{S}_W = \mathbf{S}_W + \rho\mathbf{I}$, where $\mathbf{I}$ is an identity matrix. The optimum value of ρ can be estimated through a cross validation method.

By replacing (6) into (5) and using the KKT conditions, the constraint optimization problem (4) is reformulated to the Wolf dual problem

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

Where j runs over $1, \cdots, r-1$. This is a convex quadratic programming problem and similar to the dual optimization problem of SVOR. Suppose $\{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\gamma}\}$ is the solution of the above optimization problem, $\mathbf{w}$ is obtained from (7), and so the discriminant function value for a new input vector $\mathbf{x}$ is

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

Thus, the predictive ordinal decision function is given by

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

4. Experiments

4.1 Synthetic dataset

As is shown in Fig.1, the synthetic dataset includes three ordinal categories and each category consists in 100 samples. In this experiment, the kernel function $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \| \mathbf{x} - \mathbf{y} \|^2)$ is adopted. The experimental result is illustrated in Fig.1. It is can be found that the samples can be arranged orderly by the hyperplane generated by MCVSVOR, i.e., the samples with the same rank is classified in same bin by MCVSVOR. The experimental result validates the effectiveness of the proposed method.

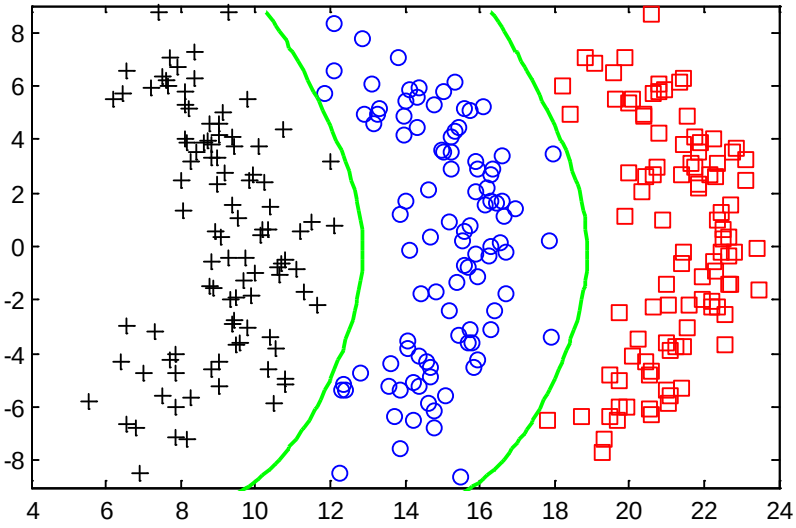

Illustration of the decision hyperplane generated by MCVSVOR

4.2 Benchmark datasets

In order to evaluate the performance of the proposed method, in this section the experimental results on several benchmark datasets, which were used in [4] and [12], will be reported. A summary of the characteristics of the selected datasets are presented in Table 1. For each dataset, the target values were discretized into ten ordinal quantities using equal-frequency binning. Each dataset was randomly partitioned into training/test splits as specified in Table 1. The partitioning was repeated 20 times independently. The input vectors were normalized to zero mean and unit variance, coordinate-wise.

Table 1 Characteristics of the selected datasets.

Datasets	No. of Attributes	No. of Training Samples	No. of Test Samples
Pyridimines	27	50	24
Machine CPU	6	150	59
Boston Housing	13	300	206
Abalone	8	1000	3177
Bank	32	3000	5192
Computer	21	4000	4192
California	8	5000	15640
Census	16	16784	16784

5. Conclusion

In this paper, we propose a novel ordinal regression method called MCVSVOR. Different from traditional SVOR which is obtained by extending SVM to tackle the ordinal regression problems, the proposed method is derived from MCVSVM and inherits its merits such as good robustness and generalization ability. The experimental results indicate the effectiveness of MCVSVOR by comparing it with the traditional regression methods SVOR and KDLOR.

References

[1] J. S. Cardoso, R. Sousa, Classification models with global constraints for ordinal data, in: ICMLA, 2010, pp.71-77.

[2] W. Chu, Z. Ghahramani, Gaussian processes for ordinal regression, Journal of Machine Learning Research 6 (2005) 1019-1041.

[3] W. Chu, S. S. Keerthi, New approaches to support vector ordinal regression, in: Proceeding of International Conference on Machine Learning (ICML-22), 2005, pp. 145-152.

[4] W. Chu, S. S. Keerthi, Support vector ordinal regression, Neural Computation 19 (3) (2007) 792-815.

[5] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification (Second Edition), New York: Wiley, 2001.

[6] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, Biostatistics 8 (1) (2007) 86-100.

[7] R. Herbrich, T. Graepel, K. Obermayer, Support vector learning for ordinal regression, in: International Conference on Artificial Neural Networks, 1999, pp. 97-102.

[8] M. Kadziński, S. Greco, R. Słowiński, Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting, Information Sciences 283 (1) (2014) 211-228.

[9] Y. Liu, Y. Liu, K. C. Chan, Ordinal regression via manifold learning, in: AAAI, 2011, pp. 398-403.

[10] A. Shashua, A. Levin. Ranking with large margin principle: two approaches, In: Advances in Neural Information Processing Systems 15, 2003, pp. 961-968.

[11] S. K. Shevade, W. Chu, Minimum enclosing spheres formulations for support vector ordinal regression, in: Sixth International Conference on Data Mining, 2006, pp: 1054-1058.

[12] B. Y. Sun, J. Li, D. D. Wu, X. M. Zhang, W. B. Li, Kernel discriminant learning for ordinal regression, IEEE Transactions on Knowledge and Data Engineering 22(6) (2010) 906-910.

[13] V. Torra, J. Domingo-Ferrer, J. M. Mateo-Sanz, M. Ng, Regression for ordinal variables without underlying continuous variables, Information Sciences 176 (4) (2006) 465-474.

[14] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer Verlag, 1995.

[15] M. Wang, F. L. Chung, S. T. Wang, On minimum class locality preserving variance support vector machine, Pattern Recognition 43 (8) (2010) 2753-2762.

[16] S. Zafeiriou, A. Tefas, I. Pitas, Minimum class variance support vector machines, IEEE Transanctions on Image Processing 16 (10) (2007) 2551-2564.