ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Support Vector Ordinal Regression using Privileged Information

Fengzhen Tang[1], Peter Tiňo[2], Pedro Antonio Gutiérrez[3] and Huanhuan Chen[4] *

1,2,4- The University of Birmingham, School of Computer Science,
Birmingham B15 2TT, United Kingdom,
Email: {fxt126, P.Tino H.Chen}@cs.bham.ac.uk

3- University of Córdoba, Department of Computer Science and Numerical Analysis,
Córdoba 14071, Spain,
Email: pagutierrez@uco.es

Abstract. We introduce a new methodology, called SVORIM+, for utilizing privileged information of the training examples, unavailable in the test regime, to improve generalization performance in ordinal regression. The privileged information is incorporated during the training by modelling the slacks through correcting functions for each of the parallel hyperplanes separating the ordered classes. The experimental results on several benchmark and time series datasets show that inclusion of the privileged information during training can boost the generalization performance significantly.

1 Introduction

Machine learning algorithms for classification problems map inputs into categorical target values (class labels) [1]. In many practical applications, a natural order may exist on the class labels. A variety of algorithms (referred to as ordinal regression) have been developed that explicitly use the class order information, e.g. [2, 3, 4, 5, 6]. A direct generalization of support vector machine approach for ordinal regression has been proposed by finding $r-1$ parallel class separation hyperplanes such that the input/feature space is partitioned into r ranked regions corresponding to the classes [4]. This approach has been further extended in support vector ordinal regression (SVOR) with explicit and implicit constraints [6].

For some ordinal regression problem, along with the training inputs $\boldsymbol{x}$, we may have access to some additional information $\boldsymbol{x}^*$ about training examples, but this privileged information $\boldsymbol{x}^*$ will not be available for inputs $\boldsymbol{x}$ at the test stage. For instance, in predicting a financial indicator, during training we have access to both the past and future contexts of time instances within the training set. Such information, if used appropriately, may significantly enhance generalization performance of ordinal regression methods. Obviously, the information about

*The research work of P.A. Gutiérrez is partially funded by FEDER funds, the TIN2011-22794 project of the MINECO and the P11-TIC-7508 project of the "Junta de Andalucía" (Spain). The work of H. Chen was supported in part by the National Natural Science Foundation of China under Grants 61203292, 61311130140 and the One Thousand Young Talents Program.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

the future is privileged and will not be available in the test phase. Motivated by SVM+ [7], which incorporates privileged information by modelling the slack variables of training inputs through so-called correcting functions, in this paper, we propose to exploit privileged information in Support Vector Ordinal Regression with Implicit constraints (SVORIM) by constructing slack variable models for each parallel separation hyperplane, which will be referred to as SVORIM+.

2 SVORIM+ Approach

We chose the implicit SVOR formulation (SVORIM) instead of the explicit one [6], because in the explicit SVOR the $j-$th hyperplane ($j = 1, 2, ...r-1$) is constrained only by the slacks of patterns from adjacent classes, whereas in SVORIM it is constrained by the slacks of patterns from all classes. Since the key aspect of incorporating privileged information into SVOR is modelling of slacks via models operating on the privileged space, the SVORIM framework can provide more flexibility in using the privileged information through greater number of correcting functions. Also, SVORIM was empirically found to outperform the explicit constraint approach in terms of Mean Absolute Error (MAE) [6].

Suppose we have observations classified into r ordered categories and there are n^k examples $\boldsymbol{x}_i^k \in X$, $i = 1, ..., n^k$, in the k^{th} category, $k = 1, 2, ..., r$. To simplify the presentation we assume that each example $\boldsymbol{x}_i^k$ has an associated privileged information[1], $\boldsymbol{x}_i^{*k} \in X^*$. For $\boldsymbol{x}_i^k \in X$ with $\boldsymbol{x}_i^{*k} \in X^*$, the slack values corresponding to each separating hyperplane j are obtained through models ("correcting functions") of the form $\xi_{ki}^j = \boldsymbol{w}_j^* \Phi^*(\boldsymbol{x}_i^{*k}) + b_j^*$, operating on the privileged space X^*. The primal problem of the proposed SVORIM+ can be formulated as:

$$\begin{aligned}
&\min_{\boldsymbol{w},b,\boldsymbol{w}^*,b^*} \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \frac{\gamma}{2} \sum_{j=1}^{r-1} (\parallel \boldsymbol{w}_j^* \parallel^2) + C \sum_{j=1}^{r-1} \sum_{k=1}^{r} \sum_{i=1}^{n^k} (\boldsymbol{w}_j^* \cdot \Phi^*(\boldsymbol{x}_i^{*k}) + b_j^*), \\
&s.t. \text{ for every } j = 1, ...r-1, \qquad (1) \\
&\boldsymbol{w} \cdot \Phi(\boldsymbol{x}_i^k) - b_j \leq -1 + (\boldsymbol{w}_j^* \cdot \Phi^*(\boldsymbol{x}_i^{*k}) + b_j^*), \text{for } k = 1, ..., j; i = 1, ..., n^k, \\
&\boldsymbol{w} \cdot \Phi(\boldsymbol{x}_i^k) - b_j \geq +1 - (\boldsymbol{w}_j^* \cdot \Phi^*(\boldsymbol{x}_i^{*k}) + b_j^*), \text{for } k = j+1, ..., r; i = 1, ..., n^k, \\
&\boldsymbol{w}_j^* \cdot \Phi^*(\boldsymbol{x}_i^{*k}) + b_j^* \geq 0.
\end{aligned}$$

where Φ and Φ^* are feature maps induced by kernels operating in X and X^* spaces, respectively. The term $\sum_{j=1}^{r-1} (\parallel \boldsymbol{w}_j^* \parallel^2)$ corresponds the capacity of the correcting functions and is controlled by the parameter $\gamma \geq 0$, tuned via cross-validation.

Note that unlike in SVORIM [6], here the slack variables are reduced to one set per threshold and are replaced by correcting functions defined in the privileged information space.

[1] Extension to the case where only a subset of training examples has privileged information is straightforward.

Following the standard SVM practice, Lagrangian will be constructed:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\parallel \boldsymbol{w} \parallel^2 + \frac{\gamma}{2}\parallel \boldsymbol{w}_j^* \parallel^2 + C\sum_{j=1}^{r-1}\sum_{k=1}^{r}\sum_{i=1}^{n^k}(\boldsymbol{w}_j^*\cdot\Phi^*(\boldsymbol{x}_i^{*k})+b_j^*)\\ &- \sum_{j=1}^{r-1}\sum_{k=1}^{j}\sum_{i=1}^{n^k}\alpha_{ki}^j(-1+\boldsymbol{w}_j^*\cdot\Phi^*(\boldsymbol{x}_i^{*k})+b_j^*-\boldsymbol{w}\cdot\Phi(\boldsymbol{x}_i^k)+b_j) \qquad (2)\\ &- \sum_{j=1}^{r-1}\sum_{k=j+1}^{r}\sum_{i=1}^{n^k}\alpha_{ki}^j(-1+\boldsymbol{w}_j^*\cdot\Phi^*(\boldsymbol{x}_i^{*k})+b_j^*+\boldsymbol{w}\cdot\Phi(\boldsymbol{x}_i^k)-b_j)\\ &- \sum_{j=1}^{r-1}\sum_{k=1}^{r}\sum_{i=1}^{n^k}\beta_{ki}^j(\boldsymbol{w}_j^*\cdot\Phi^*(\boldsymbol{x}_i^{*k})+b_j^*),\end{aligned}$$

where $\alpha_{ki}^j, \beta_{ki}^j \geq 0$ are the Lagrange multipliers. The primal problem is then transformed into its (more manageable) dual formulation using the KKT Conditions:

$$\begin{aligned}&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\sum_{k,i}(\sum_{j=1}^{r-1}\alpha_{ki}^j)\\ &-\frac{1}{2}\sum_{k,i}\sum_{k',i'}\{(\sum_{j=1}^{k-1}\alpha_{ki}^j-\sum_{j=k}^{r-1}\alpha_{ki}^j)(\sum_{j=1}^{k'-1}\alpha_{k'i'}^j-\sum_{j=k'}^{r-1}\alpha_{k'i'}^j)\mathcal{K}(\boldsymbol{x}_i^k,\boldsymbol{x}_{i'}^{k'})\}\\ &-\frac{1}{2\gamma}\sum_{j=1}^{r-1}\sum_{k,i}\sum_{k',i}\{(\alpha_{ki}^j+\beta_{ki}^j-C)(\alpha_{k'i'}^j+\beta_{k'i'}^j-C)\mathcal{K}^*(\boldsymbol{x}_i^{*k},\boldsymbol{x}_{i'}^{*k'})\} \qquad (3)\\ &s.t.\\ &\sum_{k=1}^{j}\sum_{i=1}^{n^k}\alpha_{ki}^j=\sum_{k=j+1}^{r}\sum_{i=1}^{n^k}\alpha_{ki}^j,\ \sum_{k=1}^{r}\sum_{i=1}^{n^k}(\alpha_{ki}^j+\beta_{ki}^j-C)=0,\ \forall j\\ &\alpha_{ki}^j\geq 0,\beta_{ki}^j\geq 0,\forall i,\forall j\end{aligned}$$

where $\mathcal{K}(\cdot,\cdot)$ and $\mathcal{K}^*(\cdot,\cdot)$ are kernels in X and X^* spaces, respectively. Once the solution of the dual problem is found, the value of the discriminant function at (a new) input $\boldsymbol{x}$ is $F(\boldsymbol{x})=\sum_{k,i}(\sum_{j=1}^{k-1}\alpha_{ki}^j-\sum_{j=k}^{r-1}\alpha_{ki}^j)\mathcal{K}(\boldsymbol{x}_i^k,\boldsymbol{x})$ and the predictive ordinal decision function is defined as $\arg\min_i F(\boldsymbol{x}) < b_i$.

3 Experiments

We tested our methodology on 7 data sets of different nature and origin. The input vectors were normalized to zero mean and unit variance. RBF kernels were used in both X and X^* spaces with kernel widths σ and σ^*, respectively. In all experiments the parameter ranges were as follows: $\log_{10} C \in \{-2,-1,\ldots,2\}$, $\log_{10}\sigma \in \{-2,-1,\ldots,2\}$, $\log_{10}\sigma^* \in \{-2,-1,\ldots,2\}$ and $\log_{10}\gamma \in \{-2,-1,\ldots,2\}$.

Hyper-parameters were tuned via grid search based on 5-fold cross validation over the training set. The **cvx** matlab tool[2] was used as optimization routine.

3.1 Benchmark datasets

We employed two benchmark ordinal datasets, *Pyrimidines* and *MachineCPU* used in [6]. Following [6], the continuous targets were discretized to 5 ordinal categories (equal-frequency binning). Each data set was randomly independently partitioned into training/test splits 10 times, yielding 10 re-sampled training/test sets of size 50/24 and 150/59 for *Pyrimidines* and *MachineCPU*, respectively. In order to demonstrate the advantage of the proposed method for incorporating the privileged information, an initial experiment is conducted which categorizes the input dimensions into 'original' and 'privileged' features in spaces X and X^*, respectively. For each data set, we sort the input features in terms of their relevance for the ordinal classifier (in our case SVORIM). The first most relevant half of the features will form the privileged information, while the remaining half will constitute the original space X. Privileged features will only be incorporated in training of SVORIM+ and will be absent during testing.

The average results over 10 randomized data set splits (trials), along with standard deviations are shown in Table 1. Exploiting the privileged information slightly decreases the Mean zero-one error (MZE) (by roughly 1%), decreases MAE (roughly by 7%) and decreases Macroaveraged mean absolute error ($MMAE$) (with about 13% of improvement for *Pyrimidines* and 7% for *MachineCPU*). We used the non-parametric Wilcoxon signed-rank test [8] to assess significance of the performance differences. The corresponding p-values (included in Table 1) reveal that the differences in MAE and MMAE are significant. We stress that both models are using the same set of features during the test phase, the privileged information was used only during training.

Table 1: Performance comparison of SVORIM and SVORIM+ on Benchmark data sets.

Dataset	Criteria	SVORIM	SVORIM+	$p-$value
Pyrimidines	MZE	0.5834±0.0651	0.5750±0.0756	0.5000
	MAE	0.7875±0.1249	0.7250±0.1306	0.0156•
	MMAE	0.9627±0.1609	0.8373±0.1421	0.0039•
MachineCPU	MZE	0.4390±0.0611	0.4356±0.0480	0.7656
	MAE	0.5220±0.0984	0.4848±0.0599	0.0547◦
	MMAE	0.5197±0.0991	0.4808±0.0601	0.0488•

◦: Statistically significant differences with a level of significance of $\alpha = 0.15$.
•: Statistically significant differences with a level of significance of $\alpha = 0.05$.

3.2 Time series datasets

In time series data sets (see Table 2), during the training, information about the immediate future can be used as privileged information, i. e. if we predict

[2] http://cvxr.com/cvx

the value at time t, privileged information is a vector of future values at time $t+1, \cdots, t+d$, while original information is a vector of historical observations at time $t-d, \cdots, t-1$, where the time window length d for both original and privileged information is given in Table 2. As in [9], the time series were quantized into 4 ordinal categories. *FTSE100* is a series of daily price (index) spreads within 1 October 2008 - 31 September 2012 (downloaded from Yahoo Finance). *Wine* data set contains Australian red wine sales in the period of 1980-1991. *SOI* is monthly values of the Southern Oscillation Index (SOI), which indicates the sea surface temperature of the Pacific, in the period between January 1999 and October 2012. *Laser* data represents a cross-cut through periodic to chaotic intensity pulses of a real laser in a chaotic regime. *Birth* data set contains births per 10,000 of 23 year old women in U.S. in the period of 1917-1975.

Table 2: Description of the time series datasets, d is the time window length for both original and privileged information.

Dataset	training/test	d	# trails
FTSE100	1 year/1 month	5	36
wine	118/13	5	10
SOI	300/200	5	7
Laser	500×10/4874	10	-
Birth	39/9	5	10

For *FTSE100* and *SOI* we used the rolling window methodology with window size of test set size. For smaller data sets, *Wine* and *Birth*, we used 10-fold cross-validation. For the rather long Laser dataset, we trained an ensemble model consisting of 10 models independently trained on 10 non-overlapping folds of the training set (500 points in each fold). The results are given in Table 3, including the results of the Wilcoxon test. For *Laser* dataset, the statistical test is applied to the results obtained for each of the 10 models of the ensembles. Exploiting the privileged information decreases the classification error approximately by 3%, MAE approximately by 5% and $MMAE$ by about 2%. The improvement on *Laser* data is significant, both the MZE and MAE decreased by up to about 17% and $MMAE$ decreased by up to about 46%, the differences being found statistically significant.

4 Conclusion

How to utilize all available information during training to improve generalization performance of a learning algorithm is one of the main research questions in machine learning. This paper presents a new methodology called SVORIM+, for utilizing privileged information of the training examples, unavailable in the test regime, to improve generalization performance in ordinal regression. The proposed approach incorporates the privileged information into support vector ordinal regression by constructing correcting functions for each hyperplane. The

Table 3: Test Results of the two algorithms on time series datasets

Datasets	Criteria	SVORIM	SVORIM+	$p-$value
FSTE 100	MZE	0.5884± 0.1470	0.5640±0.1311	0.0263•
	MAE	0.6799±0.2672	0.6425±0.2291	0.1493◦
	MMAE	0.7528±0.2409	0.7294±0.2610	0.0745◦
Wine	MZE	0.6428±0.1348	0.6040±0.1155	0.1250◦
	MAE	0.7804±0.1336	0.7194±0.1557	0.0156•
	MMAE	0.9871±0.1061	0.9258±0.1368	0.0625◦
SOI	MZE	0.5724±0.0497	0.5502±0.0344	0.2187
	MAE	0.6354±0.0649	0.6017±0.0470	0.2187
	MMAE	0.8549±0.0919	0.8380±0.0548	0.2187
Laser	MZE	0.0554	0.0460	0.0098•
	MAE	0.0558	0.0460	0.0040•
	MMAE	0.0852	0.0454	0.0019•
Birth	MZE	0.6333± 0.1076	0.6017± 0.1531	0.5000
	MAE	0.7778± 0.1834	0.6972± 0.1708	0.1406◦
	MMAE	1.0083± 0.2362	0.8696± 0.2001	0.0156•

experimental results on several benchmark and time series datasets confirmed that the generalization performance of SVORIM+ can indeed be superior to that of SVORIM. This is so even though the test inputs for both approaches are exactly the same – the only difference is that during the training SVORIM+ is able to model the slack variable values through correcting functions operating in the privileged space. However, compared with SVORIM, SVORIM+ requires longer training time because there are more hyper-parameters to tune. Making SVORIM+ training more efficient is a matter for our future work.

References

[1] V. Vapnik. *Statistical Learning Theory.* New York ; Chichester : Wiley, 1998.

[2] R. Herbrich, T. Graepel, and K. Obermayer. *Large Margin Rank Boundaries for Ordinal regression*, chapter 7, pages 115–132. MIT Press, Cambridge, MA, 2000.

[3] E. Frank and M. Hall. A simple approach to ordinal classification. In *Proceedings of the European Conference on Machine Learning*, volume 2167, pages 145–165, 2001.

[4] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing System 15*, pages 937–944, 2003.

[5] L. Li and H. Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing System*, pages 865–872, 2006.

[6] W. Chu and S. S. Keerthi. Support vector ordinal regression. *Neural Computation*, 19(3):792–815, 2007.

[7] V. Vapnik and A. Vashist. A new paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[8] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[9] P. Tino, C. Schittenkopf, and G. Dorffner. Volatility trading via temporal pattern recognition in quantized financial time series. *Pattern Analysis and Application*, 4(4):283–299, 2001.