LPRankBoost and Column Generation

Kristiaan Pelckmans*, Johan A.K. Suykens
Kristiaan.Pelckmans@esat.kuleuven.be

ESAT - SCD/SISTA, KULeuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Abstract. We investigate the use of LPboost for combining a set of weak learning functions into a global ranking function for predicting the order of a new subject. The notion of risk is translated as an appropriate concordance score (related to AUC and Kendall's tau), while the regularization mechanism results in a sparse solution useful for discovering structure in the specific task at hand. The result can be analyzed as a global linear programming problem, while a column generation approach yields an time- and space-efficient implementation.

1 Introduction

Ordinal regression amounts to the task of building a predictive model from a finite number n of observations consisting of m covariates, trying to explain the ordinal relations apparent in the output responses. The ranking problem here assumes no finite set of ordered outputs, but can handle a infinite set of distinct outputs which possess no known metric but are only structured by the relation '$<$'. This problem is also referred to as the preference learning setting. This problem setting is found to have many direct applications in machine learning, information retrieval, image processing and collaborative filtering, amongst many others, see e.g. [8, 7, 2] for citations. Recent developments in survival modeling also put the ranking algorithms forward in the analysis of failure time data [13]. One way of formalizing the result is to reformulate the task as getting the pairwise ordering right, and as such reducing the problem to a binary classification task [8]. The problem with such an approach would be that the number of slack-variables needed for taking mis-orderings into account grows as the square of n. In general, the practical applicability of many ranking methods is prohibited by computational issues, especially when n grows large.

A problem occurs when considering a model with a large set of parameters. Regularization - or the implicit restriction of the hypothesis space - is a main tool for tackling the increased variance due to such a large amount of parameters. More specific, the idea of maximal margin was underlying main advances in the machine learning literature such as the perceptron algorithm and the support vector machine, see e.g. [11, 12] for an introduction. Optimizing the margin was also found to be an effective tool in ordinal regression, see e.g. [10] and references.

* Research performed during a visit of CSML at CS, UCL, London, UK, supported by a FWO visiting grant

This paper considers the case where one wants to build a global ranking model consisting of a weighted sum of so-called *base learners* - simple models which can account for part of the desired ranking scheme, but not all of it. An example of a class of base-learners are the so-called stubs, i.e. thresholded versions of the different covariates. This research line was prompted by the RankBoost approach described in [7], the perceptron based ranking procedure ("Pranking") as described in [5], as well as the linear programming column generation approach for binary classification described in [6].

The contribution of this short paper is threefold. (i). An ordinal model which takes into account the mis-orderings using only n slack variables. (ii). A formulation of an optimal strategy to combine the base learners which can be solved by a convex linear program. (iii). A column generation approach is proposed to solve this linear programming problem in the case where one has a large number of base learners. The paper is organized as follows. Section 2 reviews the LPboostRank approach, while section 3 reports on the column generation strategy. Section 4 gives some preliminar numerical results.

The following formal setting is considered. Let the data $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be sampled i.i.d. from a distribution P_{XY}. Suppose we have a set of weak ranking functions $\mathcal{S} = \left\{ f_l : \mathbb{R}^d \to \mathbb{R}, \;\; l = 1, \ldots, L \right\}$, we aim at designing a global ordinal model $f : \mathbb{R}^d \to \mathbb{R}$ which is taken from

$$\mathcal{H}_L = \left\{ f : \mathbb{R}^d \to \mathbb{R} : \;\; f = \sum_{l=1}^{L} \alpha_l f_l, \;\; a_l \geq 0 \right\}. \tag{1}$$

A most appropriate criterion to compare to univariate variables Z and Y with respect to ordering is given by the concordance index

$$C(Z, Y) = P\left((Z - Z')(Y - Y') < 0\right). \tag{2}$$

where (Z, Y) and (Z', Y') are two i.i.d. samples from P_{ZY}. Its sample version is given as

$$C_n(Z, Y) = \frac{2}{n(n-1)} \sum_{i<j} I\left((Z_i - Z_j)(Y_i - Y_j) < 0\right). \tag{3}$$

Its relation with Kendall's tau ($2C(Z, Y) - 1 = \tau$) and the AUC is direct. Now the learning task can be formalized as

$$\hat{f} = \arg\min_{f \in \mathcal{H}_L} C_n(f(X), Y)), \tag{4}$$

and the analysis concerns the difference $(\min_{f \in \mathcal{H}_L} C_n(f(X), Y) - \min_{f \in \mathcal{H}_L} C(f(X), Y))$ as in [4, 1].

2 The LPrankBoost Algorithm

There exists a number of ways in which one can formalize a ranking problems. Here we take the perspective that the rank of the outputs $\{Y_i\}$ is the *true* one,

and the covariates represent only part of the information giving rise to this ranking. The part which is not contained in the covariates is modeled as noise terms.

Let the ranked permutation of the outputs be denoted as $\{Y_{(i)}\}$ such that $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ with ties broken arbitrarily, and let $X_{(1)}, \ldots, X_{(n)}$ be the corresponding input datapoints. The following formulation of the learning problem is studied.

$$\min_{a,e} \sum_{i=1}^{n} |e_{(i)}| + \gamma \sum_{l=1}^{L} a_l$$
$$\text{s.t.} \quad \begin{cases} \sum_{l=1}^{L} a_l f_l(X_{(i)}) + e_{(i)} \\ \quad \geq 1 + \sum_{l=1}^{L} a_l f_l(X_{(i-1)}) + e_{(i-1)} & \forall i = 2, \ldots, n \\ \alpha_l \geq 0 & \forall l = 1, \ldots, L. \end{cases} \tag{5}$$

This formalization can be generalized towards other learning schemes as follows. Let $D \in \{-1, 0, 1\}^{N \times n}$ be the difference matrix expressing the orders of the samples as they have to be compared. It turns out that this matrix will play a critical role into the tuning of the ranking machine. If the optimality is only based on the difference between any two consecutive points $(X_{(i)}, X_{(i+1)})$, the matrix D takes the following form

$$D = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}, \tag{6}$$

and $N = n-1$. Let $F \in \mathbb{R}^{n \times L}$ be the design matrix containing the outputs of the weak ranking models, or $F_{i,l} = f_l(X_i)$. Then the linear programming problem (5) can be generalized as

$$\min_{a,e_+,e_-} 1_L^T a + \gamma 1_n^T (e_- + e_+)$$
$$\text{s.t.} \quad DFa + D(e_+ - e_-) \geq 1_N, a, e_-, e_+ \geq 0, \tag{7}$$

using the expansion $e = e_+ - e_-$ with $e_-, e_+ \geq 0$ as classical. Remark that LPboost for binary classification is recovered by setting $D = \text{diag}(Y_{(1)}, \ldots, Y_{(n)})$. The dual problem of (7) becomes

$$\max_{\lambda} 1_N^T \lambda \quad \text{s.t.} \quad \begin{cases} (DF)^T \lambda \leq 1_L \\ -\gamma 1_n \leq D^T \lambda \leq \gamma 1_n \\ \lambda \geq 0_N. \end{cases} \tag{8}$$

Let for ease of notation let $H \in \mathbb{R}^{N \times L}$ denote the modified design matrix where $H_{il} = \left(\sum_{i=1}^{n} D_{ij} f_l(X_{(i)})\right)$ for all $i = 1, \ldots, N$ and $l = 1, \ldots, L$. Then problem

(8) is equal to

$$\max_{\lambda} \sum_{j=1}^{N} \lambda_j \quad \text{s.t.} \quad \begin{cases} \sum_{j=1}^{N} \lambda_j H_{jl} \leq 1 & \forall l = 1, \ldots, L \\ \left| \sum_{i=1}^{N} D_{ij} \lambda_i \right| \leq \gamma & \forall j = 1, \ldots, N \\ 0 \leq \lambda_j & \forall j = 1, \ldots, N. \end{cases} \tag{9}$$

The complementary slackness conditions become

$$\begin{cases} \lambda_i \left(\sum_{l=1}^{L} a_l \left(\sum_{j=1}^{N} D_{ij} f_l(X_i) + e_i \right) - 1 \right) = 0 & \forall i = 1, \ldots, N \\ a_l \left(\sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} f_l(x_i) \right) = 0 & \forall l = 1, \ldots, L. \end{cases} \tag{10}$$

3 A Column Generation Strategy

Now, we have set the stage for deriving a column generation approach for solving this large linear system as in [6] and references - i.e. in many cases of interest, both n, N and L can be large, prohibiting the straightforward application of standard solvers. Primal strategies like column generation maintain primal feasibility while working towards dual feasibility implying optimality of the result. Consider that in iteration t of the algorithm one has an active set $S_t \subset \{1, \ldots, L\}$ which denotes the columns of F (or functions f_l) which are used to solve the restricted problem, i.e. the primal problem (7) with only $\{a_i\}_{i \in S}$ being allowed to differ from zero. Let $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N) \in \mathbb{R}_+^N$ denote the Lagrange multipliers corresponding to the optimum to this restricted linear problem.

The so-called *pricing strategy* amounts to the choice of which index $l^* \in \{1, \ldots, n\}/S$ to add to the set in order to obtain S_{t+1} in each iteration step. The chosen approach is

$$l^* = \max_{l \notin S_t} \left(\sum_{j=1}^{N} \sum_{i=1}^{n} D_{ij} f_l(x_i) \hat{\lambda}_j - 1 \right). \tag{11}$$

For ease of notation, let $h_l \in \mathbb{R}^N$ defined as $h_{lj} = \sum_{i=1}^{n} D_{ij} f_l(x_i)$ such that $H_{jl} = h_{lj}$ for all $l = 1, \ldots, L$ and $j = 1, \ldots, N$. This problem can be solved in $O(mn \log(n))$ computational steps.

4 Illustrating Examples

This section illustrates the approach on a set of real-world examples. Table 1 shows preliminary results of experiments on data sets decscibed in [2], with the second column indicating the size of training set and test set. Hyperparameters including the regularization constant, the number of iterations in the RankBoost algorithm or the width of the RBF kernel were determined using a crossvalidation strategy. The proposed LPrankBoost approach and the RankBoost aproach of

[7] are constrasted to classical strategies. The weak learners are in this case build from thresholded versions of any covariate. More classical strategies include the $O(n)$ simple estimator proposed in [9], a LS-SVM regressor of $O(n^2) - O(n^3)$ on the rank-transformed responses $\{(X_i, r_Y(i))\}$, the $O(n^4) - O(n^6)$ SVM approach as proposed in [3] and the Gaussian Process approach of $O(n^4) - O(n^6)$ given in [2]. The performance of the different algorithms is expressed in terms of Kendall's τ computed on the test data. Computation time was restricted to 5 min. on a standard laptop PC, Matlab, in order to simulate feasibility issues often arising in practice.

Data (train/test)	LPRankBoost	RankBoost	oMAM	LS-SVM	oSVM	oGP
Bank(1) (100/8.092)	0.43	0.42	0.37	0.43	0.46	0.41
Bank(1) (500/7.629)	0.54	0.49	0.49	0.51	0.55	0.50
Bank(1) (5.000/3.192)	0.56	0.56	0.56	0.56	-	-
Bank(1) (7.500/692)	**0.58**	0.57	0.57	-	-	-
Bank(2) (100/8.092)	0.84	0.83	0.81	0.84	0.87	0.80
Bank(2) (500/7.629)	0.85	0.85	0.83	0.86	0.87	0.81
Bank(2) (5.000/3.192)	0.86	0.86	0.86	**0.88**	-	-
Bank(2) (7.500/692)	0.87	0.86	**0.88**	-	-	-
Cpu(1) (100/20.540)	0.62	0.60	0.44	0.62	0.64	0.63
Cpu(1) (500/20.140)	0.62	0.62	0.50	0.66	0.66	0.65
Cpu(1) (5.000/15.640)	0.68	0.67	0.57	0.68	-	-
Cpu(1) (7.500/13.140)	0.68	0.86	0.60	-	-	-
Cpu(1) (15.000/5.640)	**0.69**	0.68	**0.69**	-	-	-

Fig. 1. *Numerical performances on large scaled datasets described in [2], where the computation was terminated if it took more than 5 minutes. Results on ordinal regression tasks using the proposed LPrankBoost approach, the RankBoost approach as described in [7], ordinal Maximal Average Margin Classifiers (oMAMC) ([9]), a regression on the rank-transformed responses using LS-SVMs [12], ordinal SVMs and ordinal Gaussian Processes for preferential learning. The results are expressed as Kendall's τ (with $-1 \leq \tau \leq 1$) computed on a test set.*

5 Conclusions

This abstract considers a boosting algorithm for ranking based on a linear program which can be solved by a column generation strategy. While only the contour of the approach is sketched in this paper, one can already see from the preliminary results that the approach can be highly effective in practice.

A deeper theoretical as well as practical understanding of the formalization belongs certainly to the near future, while the application to analysis of failure time data as in [13] opens up a fascinating application field.

Acknowledgements

K. Pelckmans is supported by an FWO PDM. J.A.K. Suykens and B. De Moor are a (full) professor at the Katholieke Universiteit Leuven, Belgium.

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006 OPTEC, IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0302.07, (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+ Belgian Federal Science Policy Office: IUAP P6/04, EU: ERNSI;

References

1. S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.
2. W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2006.
3. W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *in Proc. of International Conference on Machine Learning*, pages 145–152. 2005.
4. S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Scoring Using Empirical Risk Minimization. *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30: Proceedings*, 2005.
5. K. Crammer and Y. Singer. Pranking with ranking. *Advances in Neural Information Processing Systems*, 14:641–647, 2002.
6. A. Demiriz, K.P. Bennett, and J. Shawe-Taylor. Linear Programming Boosting via Column Generation. *Machine Learning*, 46(1):225–254, 2002.
7. Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4(6):933–969, 2004.
8. R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000. MIT Press, Cambridge, MA.
9. K. Pelckmans, J.A.K. Suykens, and B. De moor. A risk minimization principle for a class of parzen estimators. *Advances in Neural Information Processing Systems*, 20:1–8, 2007.
10. A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. *Advances in Neural Information Processing Systems*, 15:961–968, 2003.
11. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
12. J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.
13. Suykens J.A.K. Van Huffel S. Van Belle V., Pelckmans K. Support vector machines for survival analysis. *in Proc. of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, 2007.