Non-linear Label Ranking for Large-scale Prediction of Long-Term User Interests

Nemanja Djuric, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, Slobodan Vucetic

Personalization

Personalization of online content may lead to improved user experience and directly translate into financial gains for online businesses. In addition, personalization fosters stronger bond between users and companies, and can help in increasing user loyalty and retention.

We consider content personalization from the viewpoint of targeted advertising. Here, for each individual user the task is to find the best matching ads to be displayed, which improves user's online experience (as only relevant and interesting ads are shown to the user) and can lead to increased revenue for the advertisers (as users are likely to click on the ad and make a purchase).

Ad Targeting

Popular approach in present-day targeting is to assign categories to display ads, such as "*sports*" or "*finance*", and separately learn a predictive model for each, able of estimating the probability of an ad click for the entire user population. Then, *N* users with the highest click probability are selected for ad exposure. Known issues with the approach include **overexposure**, where a single user may be among the top *N* users for many categories, and **starvation**, where some users do not qualify for any of the categories.

The user-interest model

An alternative avenue is to sort for each user outputs of the predictive models, and qualify users based on their top K categories. The approach guarantees that a user is qualified into several categories, eliminating overexposure and starvation issues. However, this method may still be suboptimal, as the predictive models are trained in isolation and do not consider relationships between different categories. We explore methods capable of capturing more complex class dependencies, and consider the user-interest model from a label ranking standpoint.

Current work highlights

We propose a novel label ranking algorithm with the following characteristics:

- Highly efficient, suitable for large-scale settings;
- Based on the state-of-the-art non-linear AMM classifiers;
- Learning accurate, non-linear models on very limited resources;
- Complexity of the model adapts to the complexity of the training data.

The results show that the algorithm significantly outperformed the existing methods, indicating the benefits of the proposed approach to label ranking tasks.

Adaptive Multi-hyperplane Machine

The proposed large-scale, non-linear method is based on Adaptive Multi-hyperplane Machines (AMM, Wang et al., 2011). A highly-optimized C++ code is available from SourceForge, as a part of ***BudgetedSVM*** software toolbox). Assume we are given a data set $D = \{(\mathbf{x}_t, y_t), t = 1..T\}$, where $\mathbf{x}_t$ is a *d*-dim feature vector and y_t is a label from set *Y*. AMM is a multi-class model, represented by a weight matrix $\mathbf{W}$,

$$\mathbf{W} = [\mathbf{w}_{1,1}, \dots, \mathbf{w}_{1,b_1} \mid \mathbf{w}_{2,1}, \dots, \mathbf{w}_{2,b_2} \mid \dots \mid \mathbf{w}_{M,1}, \dots, \mathbf{w}_{M,b_M}],$$

M is number of labels, b_i is number of weights for the i^{th} label, and for the i^{th} label the score is computed as,

$$g(i, \mathbf{x}_t) = \max_j \mathbf{w}_{i,j}^T \mathbf{x}_t.$$

At each Stochastic Gradient Descent (SGD) training iteration we minimize the following loss until convergence,

$$loss(\mathbf{W}, \mathbf{x}_t y_t) = \max(0, 1 + \max_{i \in Y \setminus y_t} g(i, \mathbf{x}_t) - g(y_t, \mathbf{x}_t)).$$

Interestingly, by assuming each class has *infinite* number of zero weights prior to training, during SGD the model complexity will *adapt* to the complexity of the data set.

The proposed method: AMM-rank

We extend AMM to the label ranking domain, retaining its useful properties. During inference, label ranking for $\mathbf{x}_{new}$ is obtained by sorting its label scores,

$$\hat{\pi}_{new} = \text{sort}([g(1, \mathbf{x}_{new}), g(2, \mathbf{x}_{new}), \dots, g(L, \mathbf{x}_{new})]).$$

During training, assuming example $\mathbf{x}_t$ has label ranking π_t, we optimize the objective function at each training step *t*,

$$L_{rank}^{(t)}(\mathbf{W}) = \frac{\lambda}{2} \|\mathbf{W}\|_F^2 + loss_{rank}(\mathbf{W}; (\mathbf{x}_t, \pi_t)).$$

If by *L* we denote the total number of labels and L_t available number of labels at the t^{th} training iteration, instantaneous rank loss is defined as

$$loss_{rank}(\mathbf{W}; (\mathbf{x}_t, \pi_t)) = \sum_{i=1}^{L_t} \nu(i) \sum_{j=1}^{L} I(\pi_{ti} > j) \cdot \max(0, 1 + g(j, \mathbf{x}_t) - w_{\pi_i, z_{ti}}^T \mathbf{x}_t).$$

higher ranks incur higher costs

penalty when higher-ranked label is misranked

enforce margin between label predictions

We need to compute the gradient for SGD training. By taking the derivative of L_{rank} with respect to weight $\mathbf{w}_{ij}$, we obtain

$$\nabla_{ij}^{(t)} = \lambda \mathbf{w}_{ij}^{(t)} - \mathbf{x}_t I(j = z_{ti})\, \nu(\pi_i^{-1}) \sum_{k=1}^{L} \Big(I(i > k) I(1 + g(k, \mathbf{x}_t) > \mathbf{w}_{ij}^{(t)} \mathbf{x}_t) \Big) +$$
$$\mathbf{x}_t I(j = z_{ti}) \sum_{k=1}^{L} \Big(\nu(k) I(k > i) I(1 + \mathbf{w}_{ij}^{(t)} \mathbf{x}_t > \mathbf{w}_{iz_{tk}}^{(t)} \mathbf{x}_t) \Big),$$

where z_{ti} determines which weight belonging to label *i* is used to compute the i^{th} class score for the t^{th} example. Then, training data points are observed one-by-one, and the training is stopped upon convergence to the local optimum.

Data set

The data set was generated using the information about online activities of 3.3 million users collected on Yahoo servers. Activities are temporal sequences of events (page view, search query, ad view, ad click, search link click, sponsored search click), categorized into one of 50 categories. For each activity type, we generate per-category features as recency and intensity before time *t*:

$$\text{intensity}_{cat} = \sum_{\text{all actions}} \alpha^{t - t_i} \qquad \text{recency}_{cat} = \min_{\text{all actions}} (t - t_i),$$

and ground-truth label rankings by sorting category intensities for ad click event collected during one month after time *t*. Example of label ranks are given below:

Females, aged 21-25	Female, aged 65+
01. Retail/Apparel	01. Consumer Goods/Beauty
02. Technology/Internet Services	02. Retail/Apparel
03. Telecommunications/Wireless	03. Life Stages/Education
04. Travel/Destinations	04. Finance/Loans
05. Consumer Goods/Beauty	05. Finance/Insurance
06. Technology/Consumer Electronics	06. Finance/Investment
07. Consumer Goods/Lottery	07. Technology/Internet Services
08. Travel/Vacations	08. Entertainment/Television
09. Travel/Non US	09. Retail/Home
10. Life Stages/Education	10. Telecommunications/Wireless

Empirical evaluation

We compared our method to a number of state-of-the-art, large-scale methods: 1) original AMM; 2) Central Mallows; 3) age/gender Mallows; 4) Instance-based Mallows; 5) logistic regression; 6) pairwise approach. We compared methods when we used ad view events, and when we excluded ad view events. In the following table we report the disagreement error:

Table 1. Disagreement error of various label ranking methods

Method	No adv	With adv
AMM	0.3446	0.2611
Central-Mal	0.2957	0.2957
AG-Mal	0.2820	0.2820
IB-Mal	0.2694	0.1899
LR	0.2110	0.1419
PW-LR	0.2091	0.1226
AMM-rank	**0.1996**	**0.1083**

We also report precision and recall by observing the top-K performance of the algorithms:

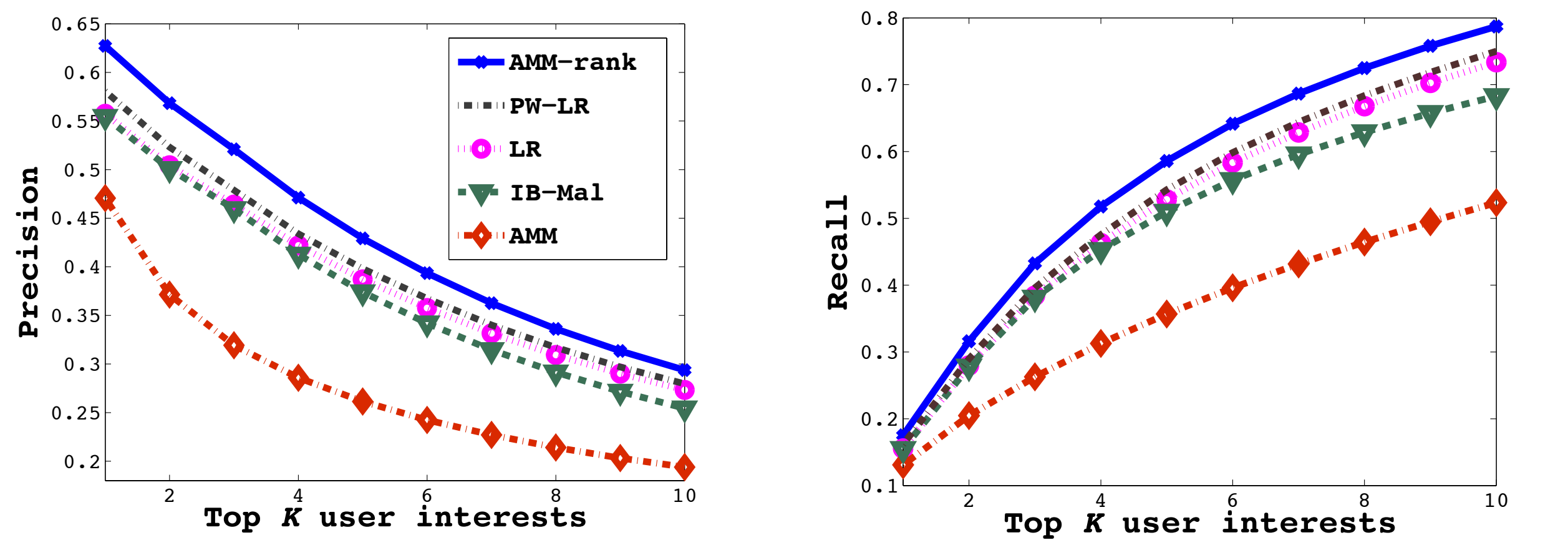

Figure 1. Comparison of retrieval performance of different methods

If printing, cut here below the line.