A Margin-based MLE for Crowdsourced Partial Ranking

Qianqian Xu[1], Jiechao Xiong[2], Xinwei Sun[3,4], Zhiyong Yang[5], Xiaochun Cao[5], Qingming Huang[1,6,7]*, Yuan Yao[8]*

[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China [2] Tencent AI Lab, Shenzhen, 518057, China

[3] School of Mathematical Sciences, Peking University, Beijing, 100871, China [4] DeepWise AI Lab, Beijing, 100085, China

[5] State Key Laboratory of Info. Security (SKLOIS), Inst. of Info. Engin., CAS, Beijing, 100093, China

[6] University of Chinese Academy of Sciences, Beijing, 100049, China

[7] Key Lab of Big Data Mining and Knowledge Management, CAS, Beijing, 100190, China

[8] Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Introduction

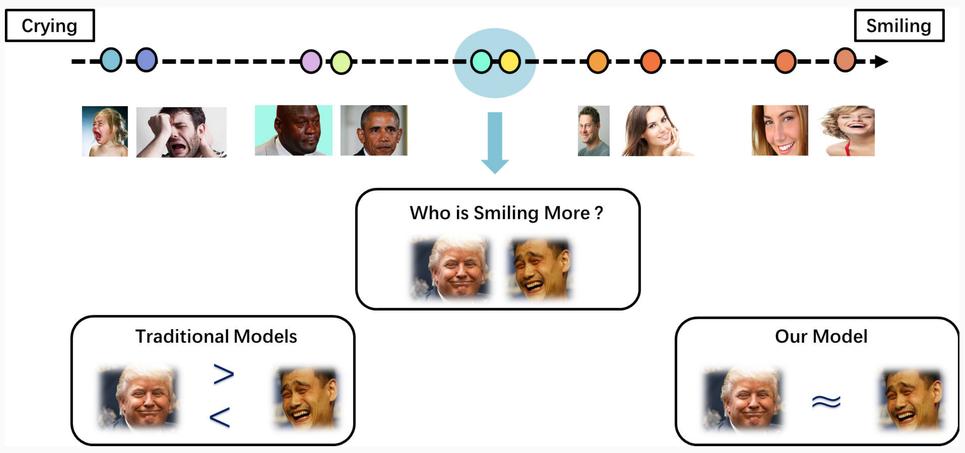

Fig. 1: Illustration of our motivation

A preference order or ranking aggregated from pairwise comparison data is commonly understood as a strict total order. However, in real world scenarios, some items are intrinsically ambiguous in comparisons, which may very well be an inherent uncertainty of the data. In this case, the conventional total order ranking can not capture such uncertainty with mere global ranking or utility scores. In this paper, we are specifcally interested in the recent surge in crowdsourcing applications to predict partial but more accurate (i.e., making less incorrect statements) orders rather than complete ones. To do so, we propose a novel framework to learn some probabilistic models of partial orders as a marginbased Maximum Likelihood Estimate (MLE) method.

Model Formulation

- **PairWise Ranking On Graph**: Suppose there are n alternatives or items to be ranked. The pairwise comparison labels collected from users can be naturally represented as a directed comparison graph $G = (V; E)$.
  - → **Vertexes**. $V = \{1, 2, \ldots, n\}$ be the vertex set of n items
  - → **Edges** $E = \{(u, i, j) : i, j \in V, u \in U\}$ be the set of edges, where U is the set of all users who compared items.
  - → y_{ij}^u :**User Annotations on the Edge** User u provides his/her preference between choice i and j, such that $y_{ij}^u > 0$ means u prefers i to j and $y_{ij}^u \leq 0$ otherwise. Hence we may assume $y : E \to R$ with skew-symmetry (orientation) $y_{ij}^u = -y_{ji}^u$. The magnitude of y_{ij}^u can represent the degree of preference and it varies in applications. The simplest setting is the binary choice, where $y_{ij}^u = 1$ if u prefers i to j and $y_{ij}^u = -1$ otherwise.
- **Partial Ranking** Define $i \succ j$ as $(i > j) \wedge (\neg(j > i))$. If, for two alternatives i and j, neither $i \succ j$ nor $j \succ i$, then these alternatives are considered as incomparable, we then denote $i \perp j$ or equivalently $j \perp i$. In other words, if i and j are too similar such that we think neither i precedes j nor j precedes i, we then claim that i and j are incomparable. Suppose $P(i, j)$ is a measure of support for the order (preference) relation $i \succ j$ with property $P(j, i) = 1 - P(i, j)$. Then a POR is defined as

$$\mathcal{R}_\alpha = \{i \succ j : P(i, j) \geq \alpha\}$$

  by setting α big enough.
- **A Probabilitis Model for The Generation of The Labels** Suppose that the true scaling scores for n items are $\boldsymbol{s} = [s_1, \cdots, s_n]$ and we collect N pairwise comparison samples $\{(i_k, j_k, y_k)\}_{k=1}^N$ in total. Here (i_k, j_k) is a pair of items, and y_k is the corresponding comparison label. Suppose that, for the kth observation, y_k is generated by:

$$y_k = \begin{cases} 1, & s_{i_k} - s_{j_k} + \epsilon_k > \lambda; \\ -1, & s_{i_k} - s_{j_k} + \epsilon_k < -\lambda; \\ 0, & \text{else.} \end{cases} \tag{1}$$

  Where ϵ_k is the noise effect which induces the randomness of y_k. To model ϵ_k, we assume that $\epsilon_k \overset{i.i.d}{\sim} \Phi(\cdot)$, where Φ is the c.d.f. function of the corresponding distribution. Practically, we adopt three commonly used distribution: the basic uniform model, Bradley-Terry model, and Thurstone-Mosteller model
- **A MLE Framework**: denoting ζ_k^+ as $[1, x_k^\top]^\top \theta$ and ζ_k^- as $[-1, x_k^\top]^\top \theta$, following the MLE framework, the objective function could be defined as the negative log-likelihood function over the observed samples:

$$\begin{aligned} \ell(y|\boldsymbol{s}, \lambda) = -\sum_k \Big(& 1\{y_k = 1\} log\big[1 - \Phi(\zeta_k^+)\big] \\ & + 1\{y_k = 0\} log\big[\Phi(\zeta_k^+) - \Phi(\zeta_k^-)\big] + 1\{y_k = -1\} log\big[\Phi(\zeta_k^-)\big] \Big). \end{aligned} \tag{2}$$

Theoretical Analysis

we construct the set of all incomparable pairs as $\mathcal{M}$, a conservative set as $\widehat{\mathcal{M}}$ and the aggressive set as $\widetilde{\mathcal{M}}$:

$$\mathcal{M} = \{(i, j) : |s_i^\star - s_j^\star| \leq \lambda^*\}, \tag{3}$$

$$\widehat{\mathcal{M}} = \{(i, j) : |\hat{s}_i - \hat{s}_j| \leq \hat{\lambda} - 3\Delta\}, \tag{4}$$

$$\widetilde{\mathcal{M}} = \{(i, j) : |\hat{s}_i - \hat{s}_j| \leq \hat{\lambda} + 3\Delta\}, \tag{5}$$

where N is the number of samples. Now we first propose a theorem which shows that with high probability, $\widehat{\mathcal{M}} \subseteq \mathcal{M} \subseteq \widetilde{\mathcal{M}}$, followed by a practical interpretation via the remark that comes right after the theorem.

Theorem 1. *Let $\theta = (\lambda, \boldsymbol{s})$. Then with probability at least $1 - 2(n+1)^{\frac{\delta - 2\delta}{\delta}}$, we will have that $\widehat{\mathcal{M}} \subseteq \mathcal{M} \subseteq \widetilde{\mathcal{M}}$.*

Remark 1. *If $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ occurs, we set the threshold λ as $\underline{\lambda} = \hat{\lambda} - 3\Delta$. Then all the detected incomparable pairs are truly incomparable, thus FDR = 0 is guaranteed. Likewise, if $\mathcal{M} \subseteq \widetilde{\mathcal{M}}$ i.e. $\widetilde{\mathcal{M}}^c \subseteq \mathcal{M}^c$ occurs, we have $|\hat{s}_i - \hat{s}_j| > \hat{\lambda} + 3\Delta$ indicating $|s_i^\star - s_j^\star| > \lambda^\star$. Consequently, if we set the threshold as $\overline{\lambda} = \hat{\lambda} + 3\Delta$, then all the comparable pairs will be detected as comparable and thus Power = 1 is guaranteed.*

Experiments

Table 1: Experimental results on IQA dataset.

types	algorithms	correctness		completeness		geomean	
		median	std	median	std	median	std
α-cut [?]	LRLASSO	0.9137	0.0173	0.8309	0.0325	0.8760	0.0200
	LRRidge	0.9227	0.0150	0.8044	0.0301	0.8582	0.0148
	SVMLASSO	0.9158	0.0137	0.8310	0.0297	0.8721	0.0166
	SVMRidge	0.9184	0.0099	0.8083	0.0484	0.8594	0.0246
	LSLASSO	0.9154	0.0117	0.8095	0.0285	0.8623	0.0146
	LSRidge	0.9139	0.0126	0.8218	0.0336	0.8668	0.0182
	SVRLASSO	0.9236	0.0119	0.7405	0.0291	0.8311	0.0167
	SVRRidge	0.9191	0.0145	0.7594	0.0386	0.8378	0.0187
ours	**Uniform**	**0.9137**	0.0107	**0.8623**	0.0142	**0.8867**	0.0081
	Bradley-Terry	**0.9113**	0.0124	**0.9254**	0.0141	**0.9064**	0.0082
	Thurstone-Mosteller	**0.9146**	0.0122	**0.9077**	0.0122	**0.9084**	0.0075

Table 2: Experimental results on human age dataset.

type	algorithms	correctness		completemess		geomean	
		median	std	median	std	median	std
α-cut[?]	LRLASSO	0.8640	0.0095	0.8352	0.0974	0.8511	0.0562
	LRRidge	0.8693	0.0070	0.8467	0.0186	0.8584	0.0090
	SVMLASSO	0.8674	0.0084	0.8565	0.0315	0.8619	0.0144
	SVMRidge	0.8660	0.0076	0.8447	0.1049	0.8542	0.0597
	LSLASSO	0.8688	0.0072	0.8583	0.0265	0.8617	0.0128
	LSRidge	0.8681	0.0072	0.8513	0.0193	0.8556	0.0096
	SVRLASSO	0.8732	0.0087	0.7687	0.0380	0.8177	0.0188
	SVRRidge	0.8732	0.0082	0.7750	0.0229	0.8237	0.0118
ours	**Uniform**	**0.8655**	0.0056	**0.8523**	0.0098	**0.8591**	0.0056
	Bradley-Terry	**0.8671**	0.0061	**0.8990**	0.0070	**0.8826**	0.0042
	Thurstone-Mosteller	**0.8682**	0.0062	**0.8949**	0.0067	**0.8816**	0.0044

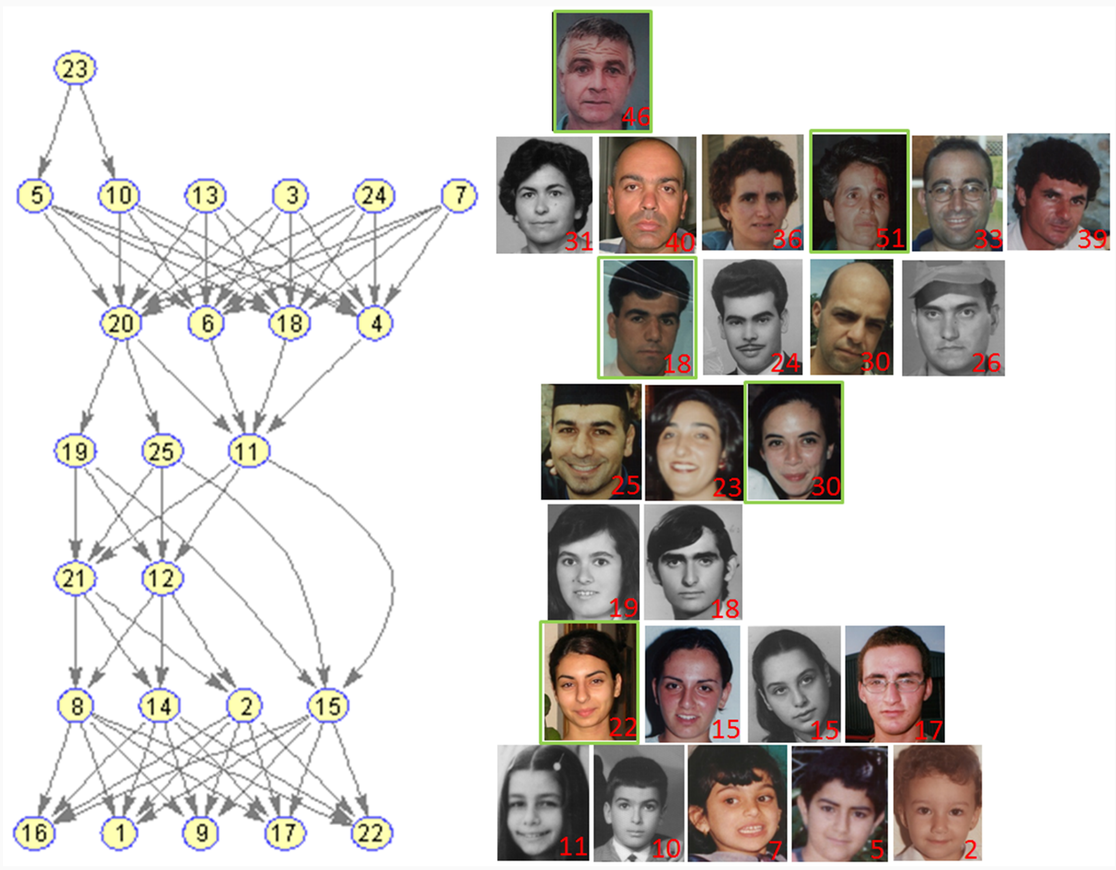

Fig. 2: Partial Ranking of human age dataset