Application and Analysis of HodgeRank on Crowdsourced Datasets

Yuxuan Chen yx.chen@connect.ust.hk

Department of Mathematics, HKUST

Introduction

As crowdsourcing gains popularity as a data collection method, ensuring the quality of such data becomes increasingly critical. In this report, we introduce HodgeRank and its extended version HodgeRank-GLM and apply them on crowdsourced data World College dataset and Human Age dataset. We also analyse the performance of four different models of HodgeRank-GLM and the inconsistency. At last, we evaluate assessors' reliability through a simple analysis.

Data Description

- **World College Dataset** The dataset is collected from the crowdsourcing task on pairwise ranking on universities on ALL OUR IDEAS website up to Nov 26, 2017. After cleaning, the dataset is composed of 261 colleges and their corresponding scores (not ground-truth), 9408 pairwise comparisons given by 409 distinct annotators from various countries who are given a pair of universities and asked to choose which university is more attractive to attend. After combining the comparisons of all the annotators, there are 8175 distinct comparisons and 41402 triangles.
- **Human Age Dataset** The dataset contains 30 images from the FG-NET 1 dataset, which have been annotated by 94 volunteer users on ChinaCrowds platform. There are 12778 pairwise comparisons given by annotators who are presented with two images and asked to choose which one appears older. After combining the comparisons of all the annotators, there are 435 distinct comparisons and 4060 triangles, which means that every pair of the images has been compared. The ground-truth ages for the images are also included in the dataset.

Methodology

- **HodgeRank** is based on the Hodge Decomposition Theorem in [1] which mainly illustrates that the pairwise comparison data Y can be decomposed into three parts,

$$Y = \mathrm{im}(\mathrm{grad}) \oplus \ker(\Delta_1) \oplus \mathrm{im}(\mathrm{curl}^*).$$

  Hence HodgeRank in [1] obtains the global ranking component which is induced by global score s by solving

$$\min_{s \in C^0} \|\mathrm{grad}\, s - Y\|^2_{2,\omega} = \min_{s \in C^0} \sum_{i,j} W_{ij} \left(s_j - s_i - Y_{ij}\right)^2,$$

  and obtains the cyclic ranking component by solving

$$\min_{\Phi \in C^2} \|\mathrm{curl}^* \Phi - Y\|^2_{2,\omega},$$

  where $\|r\|^2_{2,\omega} = \sum_{i,j} W_{ij} r^2_{ij}$, $W_{ij} = \sum_u W^u_{ij}$, $Y_{ij} = \sum_u W^u_{ij} Y^u_{ij} / W_{ij}$. Here if the annotator u has made a pair comparison between item i and item j then $W^u_{ij} = 1$ and otherwise $W^u_{ij} = 0$; if the annotator u has made a pair comparison and think i-item is better than j-item, then $Y^u_{ij} = 1, Y^u_{ji} = -1$ and otherwise $Y^u_{ij} = Y^u_{ji} = 0$. We use Algorithm 1 in [2] to find the solutions s, Φ of the above two optimization problems and obtain global ranking component $\mathrm{grad}\, s$ and cyclic ranking component $\mathrm{curl}^* \Phi$.
- **HodgeRank-GLM** in [2] is an extended version of HodgeRank which uses several different way to formulate pairwise comparison data Y. First we define a_{ij} as the annotators have a preference on item i over item j and $n_{ij} = a_{ij} + a_{ji}$ represents the number of annotators who have made pair comparisons between item i and item j, then we can approximate the preference probability by $\hat{\pi}_{ij} = a_{ij}/n_{ij}$ and pairwise comparison data $Y_{ij} = F^{-1}(\hat{\pi}_{ij})$ is formulated through a function F:
  1. *Uniform* model: $F^{-1}(x) = 2x - 1$, which leads to HodgeRank.
  2. *Bradley-Terry* model: $F^{-1}(x) = \log \frac{x}{1-x}$.
  3. *Thurstone-Mosteller* model: F is essentially the Gauss error function (we use standard Gauss error function here).
  4. *Angular transform* model: $F^{-1}(x) = \arcsin(2x - 1)$.

Results

	Uniform	Bradley-Terry	Thurstone-Mosteller	Angular transform
Kendall τ	0.899027	0.897318	0.897436	0.898143
Inc.Total	0.698909	0.699572	0.699433	0.699117

Table 1. The Kendall value and Inc.Total of HodgeRank-GLM on World College dataset.

	Uniform	Bradley-Terry	Thurstone-Mosteller	Angular transform
Kendall τ	0.701149	0.678161	0.701149	0.705747
Inc.Curl	0.194847	0.385546	0.267967	0.198434
Inc.Harm	0	0	0	0

Table 2. The Kendall value, Inc.Curl and Inc.Harm of HodgeRank-GLM on Human Age dataset.

Triangle	Cr
(18, 23, 27)	1961.402557
(23, 27, 68)	1943.379683
(23, 27, 92)	1942.611733
(23, 27, 250)	1942.121511
(23, 27, 201)	1294.866261

Table 3. Top-5 Cr values triangles in the World College dataset.

Triangle	Cr
(5, 15, 16)	184.886978
(15, 16, 24)	143.838794
(3, 15, 16)	128.961495
(9, 15, 16)	125.495994
(7, 15, 16)	111.588494

Table 4. Top-5 Cr values triangles in the Human Age dataset.

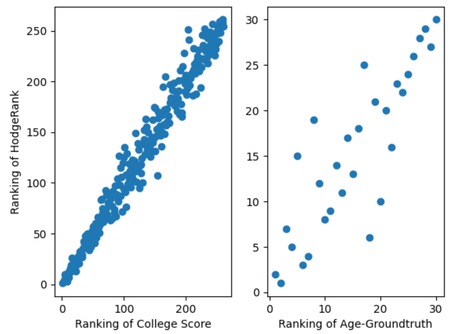

Figure 1. The predicted ranking of HodgeRank versus the ground-truth ranking for World College (left) and Human Age (right) dataset.

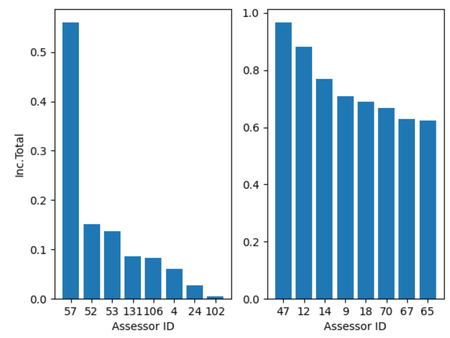

Figure 2. Top-8 Inc.Total values assessors in the World College (left) and Human Age (right) dataset.

Analysis

- **Performance** First we apply HodgeRank-GLM of four different models on the two datasets and obtain the global rankings which induced by global score s, respectively. In order to evaluate the performance of four models of HodgeRank-GLM, we use metrics Kendall τ, total inconsistency (Inc.Total), curl inconsistency (Inc.Curl) and harmonic inconsistency (Inc.Harm). Kendall τ measures the similarity of the predicted ranking and the ground-truth ranking (or the approximation of ground-truth ranking on World College data), which equals to 1 when rankings are exactly the same and equals to -1 when they're completely different. Inc.Curl, Inc.Harm and Inc.Total are the ratios of cyclic ranking component, harmonic ranking component and the sum of the above two components to pairwise comparison data Y in the sense of norm $\|\cdot\|^2_{2,\omega}$. Note that there are 41402 triangles which leads to an expensive process to calculate the Inc.Curl and Inc.Harm for World College data, so we only calculate Inc.Total; for Human Age data the calculation of Inc.Curl and Inc.Harm is possible but from Hodge Theory in [1] we know that Inc.Harm equals to 0 and Inc.Curl equals to Inc.Total since every pair of the images in Human Age data has been compared. As shown in Table 1 and Table 2, the performance of Uniform model is almost optimal in both datasets. Therefore we use Uniform model version of HodgeRank-GLM, which is actually HodgeRank, in the following experiments. In Figure 1 we visualize the predicted ranking of HodgeRank versus the ground-truth ranking (or the approximation of ground-truth ranking) by scatter plots for both datasets. From a global view, we can find the ranking result of the World College data is better than Human Age data where the scatter plot has a linear shape. This illustrates that making a decision on which person is older from images maybe a harder task than giving a decision on which college is better.
- **Inconsistency** Since there still exists local inconsistency in the data, we want to know which triangles make great contribution to the local inconsistency. In specific, we use average of relative curl in [1] to quantifies the local inconsistency of triangle (i, j, k):

$$Cr(i, j, k) = \left| \frac{Y_{ij} + Y_{jk} + Y_{ki}}{9} \left[(s_i - s_j)^{-1} + (s_j - s_k)^{-1} + (s_k - s_i)^{-1} \right] \right|.$$

  Table 3 and Table 4 show the top-5 Cr values triangles in the World College and Human Age dataset. It can be seen that the edge $(23, 27)$ in World College data and edge $(15, 16)$ in Human Age data are most significant inconsistent edges. College ID 23 and 27 correspond to 'University of Michigan, USA' and 'University of Washington, USA', it seems hard for assessors making decision on which of the two universities is better; ID 15 and 16 in Human Age data correspond to images of 15-year-old 17-year-old teenagers, which is natural as it is difficult to make a decision on which teenager is older.
- **Assessors' Reliability** In the last part, we evaluate assessors' reliability following the analysis in [2]. In specific, we apply HodgeRank on pairwise comparison data of each assessor and calculate Inc.Total. Figure 2 shows the top-8 Inc.Total values assessors in the World College and Human Age dataset.It can be seen from this figure that assessors 57 in World College data and assessors 47 in Human Age data have extraordinarily large total inconsistency, which means that the two assessors maybe careless. Note that this example is preliminary and a systematic treatment of this topic needs a distribution model of harmonic or curl flows under various experimental conditions.

References

[1] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

[2] Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.