Novel Distance Measures for Rank Aggregation

Farzad Farnoud
Joint work with O. Milenkovic
02/08/12

Rankings (permutations)

Permutations are arrangements of a set of objects.

Arrangements are ubiquitous combinatorial objects.

Example: [2431] – a permutation over the set $\{1, 2, 3, 4\}$

Used in almost every branch of mathematics, physics and biology:

- Coding and information theory
- Computer science
- Biology and bioinformatics
- Recommender systems
- Social sciences: competitions, voting
- Management and decision making

Rank aggregation I

Rank Aggregation: Combining a set of rankings (permutations) such that the result is "representative" of each individual ranking.

Title	**IMDB**	**FilmCrave**	**Aggregate**
The Shawshank Redemption	1	1	?
The Godfather	2	3	?
Fight Club	10	2	?
The Godfather: Part II	3	11	?
Pulp Fiction	4	4	?
Schindler's List	5	8	?
The Dark Knight	7	5	?
One Flew Over the Cuckoo's Nest	6	13	?
LoR: The Fellowship of the Ring	13	6	?
LoR: The Return of the King	8	7	?
SWV: The Empire Strikes Back	9	10	?
Goodfellas	11	9	?
Star Wars	12	12	?

Rank aggregation I

Rank Aggregation: Combining a set of rankings (permutations) such that the result is "representative" of each individual ranking.

Title	IMDB	FilmCrave	Aggregate
The Shawshank Redemption	1	1	1
The Godfather	2	3	?
Fight Club	10	2	?
The Godfather: Part II	3	11	?
Pulp Fiction	4	4	4
Schindler's List	5	8	?
The Dark Knight	7	5	5/6/7?
One Flew Over the Cuckoo's Nest	6	13	?
LoR: The Fellowship of the Ring	13	6	?
LoR: The Return of the King	8	7	?
SWV: The Empire Strikes Back	9	10	?
Goodfellas	11	9	?
Star Wars	12	12	?

Related work

Kemeny 59 Kemeny's axiomatic approach, use of Kendall's τ by Kemeny

Dwork 01 "Kemenization" is NP-hard, bipartite matching and Markov chain methods for aggregation by Dwork et al.

Sculley 07 Aggregation with similarity score by Sculley

Kumar 10 Generalizing Kendall's τ and Spearman's footrule by Kumar et al.

Kemeny's axioms

Formal description of rank aggregation: find a ranking that has smallest cumulative distance from a set of given rankings

Kemeny's axiomatic approach for determining good distance functions:

1. Distance function is a metric: $\mathrm{d}(\pi,\sigma)=\mathrm{d}(\sigma,\pi)\geq 0$, satisfies triangle inequality
2. $\mathrm{d}(\sigma,\pi)=\mathrm{d}(\pi,\omega)+d(\omega,\sigma)$ iff ω is "between" π and σ.
3. Relabeling of objects does not change distance: $\mathrm{d}(\pi,\sigma)=\mathrm{d}(\omega\pi,\omega\sigma)$
4. If two rankings agree except on a set S of k elements, d can be measured as if only elements of S were considered: $\mathrm{d}(abcde, acdbe)=\mathrm{d}(bcd, cdb)$.

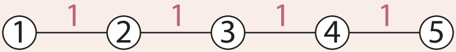
ω is between π and σ if for each pair of elements ω agrees with π or σ.

Kendall's τ

The unique distance function that satisfies Kemeny's axioms is Kendall's τ:

Kendall's τ distance between two permutations is the minimum number of swaps of adjacent elements needed to transform one permutation into the another

Example: $d(12345, 31245) = 2$: $\quad 12345 \xrightarrow{(23)} 13245 \xrightarrow{(12)} 31245$



Rank aggregation II

Need a new distance function that addresses shortcomings of Kendall's τ in terms of having following additional properties:

1. Non-uniform importance of rank values:
   - Top versus bottom

$$d(1234, 2134) > d(1234, 1243)$$

   - Similar items versus dissimilar items

$$d(ABA'B', B'BA'A) > d(ABA'B', A'BAB')$$

Clearly, ease of calculation or approximation is desirable for any such solution

Generalizing Kendall's distance

Relaxing Kemeny's axioms:

1. Distance function is a metric: $\mathrm{d}(\pi,\sigma)=\mathrm{d}(\sigma,\pi)\geq 0$, satisfies triangle inequality
2. $\mathrm{d}(\sigma,\pi)=\mathrm{d}(\pi,\omega)+d(\omega,\sigma)$ ~~iff ω is "between" π and σ~~ for some ω between π and σ if π and σ are not "adjacent"
3. Relabeling of objects does not change distance: $\mathrm{d}(\pi,\sigma)=\mathrm{d}(\omega\pi,\omega\sigma)$
4. ~~If two rankings agree except on a set S of k elements, d can be measured as if only elements of S are considered: $\mathrm{d}(abcde, acdbe)=\mathrm{d}(bcdae, cdbae)$~~

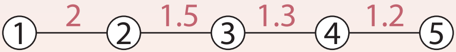
σ and π are adjacent if there is no permutation between them

Weighted Kendall distance

Weighted Kendall distance: a class of distance functions that satisfy relaxed Kemeny axioms

Generalized Kendall distance between two permutations: minimum cost of transforming one transposition into another other using swaps of adjacent element where each swap may have a cost

Example: $d(12345, 31245) = 3.5 :\quad 12345 \xrightarrow{(23)} 13245 \xrightarrow{(12)} 31245$



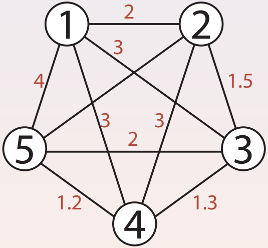
A new aggregation metric

We are interested in metrics based on swaps

A swap of two elements is called a transposition
Transposition of elements in positions i and j denoted by (ij)

Assign cost $\varphi(i,j) \geq 0$ to any transposition (ij)

Distance between two permutations π and σ is the minimum cost of transforming π to σ using transpositions



Cost functions and distance

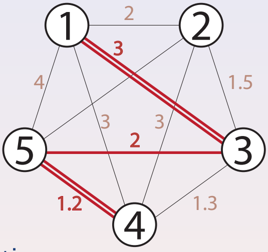
$M_\varphi(\pi, \sigma)$ = distance between π and σ based on φ
Fact: For permutations π, σ, η, proposed distance metric M_φ satisfies

1. $M_\varphi(\pi, \pi) = 0$
2. $M_\varphi(\pi, \sigma) = M_\varphi(\sigma, \pi) \geq 0$
3. $M_\varphi(\pi, \sigma) \leq M_\varphi(\pi, \eta) + M_\varphi(\eta, \sigma)$ (triangular inequality)
4. $M_\varphi(\pi, \sigma) = M_\varphi(\eta\pi, \eta\sigma)$ (left-invariance)

Because of left-invariance, it suffices to be able to compute distances between given permutations and the identity:
$M_\varphi(\pi, \sigma) = M_\varphi(\sigma^{-1}\pi, e)$

Three-stage **polynomial time** algorithm for finding distance

1. Distance of a single transposition from identity
   - A Viterbi-style algorithm on a trellis, or
   - Bellman-Ford type algorithm over graphs



2. Distance of a single cycle from identity
   - A dynamic program finds an approximation,
   - Proof of approximation using "h-transpositions"
3. Distance of a general permutation from identity
   - By extending results for single cycle permutations

See F.Farnoud and O.Milenkovic, Sorting of permutations by cost-constrained transpositions. IT Transaction, 58(1):3–23, Jan. 2012.

Back to rank aggregation I

Formally, given $\sigma_1, \cdots, \sigma_m$ and a distance function d, find

$$\arg\min_{\pi} \sum_{i=1}^{m} d(\pi, \sigma_i)$$

Several distance functions used so far [Diaconis and Graham, 80s]:

- Kendall's τ: $K(\pi, \sigma) = \#$ of transpositions of adjacent ranks. Equivalent to $\varphi_K(i, i+1) = 1$.
- Spearman's Footrule: $F(\pi, \sigma) = \sum_i |\pi(i) - \sigma(i)|$. Equivalent to the path cost function $\varphi_F(i, j) = |i - j|$.
- Cayley's distance: $T(\pi, \sigma) = \#$ of transpositions. Equivalent to $\varphi_T(i, j) = 1$.
- Spearman's rank correlation: $S^2(\pi, \sigma) = \sum_i (\pi(i) - \sigma(i))^2$

Rank aggregation II

Cost-constrained transposition distance generalizes

- Kendall's τ to model the significance of top vs bottom of a ranking, e.g.,

$$\varphi(1,2) > \varphi(9,10)$$

- Cayley's distance to model similarities and dissimilarities of elements, e.g.,

$$\varphi(\text{God Father I}, \text{God Father II}) < \varphi(\text{God Father I}, \text{Goodfellas}) < \varphi(\text{God Father I}, \text{Star Wars})$$

Rank aggregation III

To find

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{m} M_\varphi(\pi, \sigma_i)$$

we approximate M_φ by D where

$$D(\pi, \sigma) = \sum_{i=1}^{n} \text{cost}\left(p^*(\pi(i), \sigma(i))\right).$$

We can find

$$\pi' = \arg\min_{\pi} \sum_{i=1}^{m} D(\pi, \sigma_i)$$

and can show that

$$\sum_{i=1}^{m} M_\varphi(\pi', \sigma_i) \leq 4 \sum_{i=1}^{m} M_\varphi(\pi^*, \sigma_i).$$

Rank aggregation IV

Social sciences: Ranking candidates based on job interview performance; rankings of schools.

Rankings of schools: recent study in Huffington post shows that Stony Brooks is the school with "worst" teaching performance.

UIUC ranked number **three** and UCSD ranked number **five**.

Gender Studies: conducted experiments on students assessment of quality of institution based on different criteria such as diversity, tuition cost, dorm cost, proximity of friends and family...

Rank aggregation V

Undergraduate students were asked to rank the following items in order of importance in their academic life:

- A Campus friendliness and inclusiveness
- B Availability of recreational and cultural facilities
- C Quality of classrooms and dorms
- D Extracurricular student groups and activities
- E Geographical proximity to your family/boyfriend/girlfriend
- F Commitment of campus to build a diverse community
- G Being able to express one's personal identity freely
- H Being able to make friends on campus
- I Safety and security
- J Availability of financial support/ scholarship
- K Availability of personal counseling/ academic tutoring
- L Friendliness/ academic prowess of faculty members/ instructors

---

Data collected with help of Dr. Chaplin, History department, UIUC.

Rank aggregation V

... and this is how they ranked the items
(73 students, 32 female, 31 male)

Rank	**Item**
1	Friendliness/ academic prowess of faculty members/ instructors
2	Safety and security
3	Being able to make friends on campus
4	Campus friendliness and inclusiveness
5	Quality of classrooms and dorms
6	Availability of financial support/ scholarship
7	Extracurricular student groups and activities
8	Availability of recreational and cultural facilities
9	Being able to express one's personal identity freely
10	Availability of personal counseling/ academic tutoring
11	Geographical proximity to your family/boyfriend/girlfriend
12	Commitment of campus to build a diverse community

Rank aggregation V

... and this is how they ranked the items
(73 students, 32 female, 31 male)

Rank	Female students	Male students
1	faculty members/ instructors	faculty members/ instructors
2	safety and security	able to make friends on campus
3	able to make friends on campus	safety and security
4	friendliness and inclusiveness	classrooms and dorms
5	classrooms and dorms	friendliness and inclusiveness
6	financial support	financial support
7	groups and activities	counseling/ tutoring
8	rec. and cultural facilities	rec. and cultural facilities
9	able to express identity	groups and activities
10	counseling/ tutoring	able to express identity
11	proximity to your family/gf/bf	proximity to family/gf/bf
12	to build a diverse community	to build a diverse community

Topics in rank aggregation

Use cost-constrained distance for rank aggregation and rank prediction.

- Application of cost-constrained distance to rank predictions (rank collaborative filters)
- Extension of costs-constrained distances when cost depends both on location and object: $\varphi(i, j, \pi^{-1}(i), \sigma^{-1}(j))$.
- Extension of costs-constrained distance to partial rankings.

Thank you!