Judgment Aggregation and Collective Annotation

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam

Mini Course on the Theory of Aggregation (Lecture 2)
LIP6, Pierre & Marie Curie University, Paris

Opening Example

Suppose three robots are in charge of climate control for this building. They need to make judgments on p (the temperature is below 17°C), q (we should switch on the heating), and $p \rightarrow q$.

	p	$p \rightarrow q$	q
Robot 1:	Yes	Yes	Yes
Robot 2:	No	Yes	No
Robot 3:	Yes	No	No

- What should be the collective decision?

Plan for Today

Recall: Last time we discussed the axiomatics of preference aggregation and its generalisation in the form of graph aggregation.

Today we'll start with an introduction to *judgment aggregation* (JA) and then discuss the *collective annotation* of crowdsourced data.

These slides are available online:

`https://staff.science.uva.nl/u.endriss/teaching/paris-2016/`

Most of the material is covered in the two papers cited below.

U. Endriss. Judgment Aggregation. In *Handbook of Computational Social Choice*, Cambridge University Press, 2016.

C. Qing. U. Endriss, R. Fernández, and J. Kruger. Empirical Analysis of Aggregation Methods for Collective Annotation. Proc. 25th International Conference on Computational Linguistics (COLING), 2014.

The Doctrinal Paradox

Suppose a court with three judges is considering a case in contract law. Legal doctrine stipulates that the defendant is *liable* (r) *iff* the contract was *valid* (p) and it has been *breached* (q): $r \leftrightarrow p \wedge q$.

	p	q	r
Judge 1:	Yes	Yes	Yes
Judge 2:	No	Yes	No
Judge 3:	Yes	No	No
Majority:	Yes	Yes	No

Paradox: Taking majority decisions on the *premises* (p and q) and then inferring the conclusion (r) yields a different result from taking a majority decision on the *conclusion* (r) directly.

L.A. Kornhauser and L.G. Sager. The One and the Many: Adjudication in Collegial Courts. *California Law Review*, 81(1):1–59, 1993.

Variants

Our judges were expressing judgments on *atoms* (p, q, r) and consistency of a judgment set was evaluated w.r.t. an *integrity constraint* ($r \leftrightarrow p \wedge q$).

Alternatively, we could allow judgments on *compound formulas*, like so:

	p	q	$p \wedge q$
Judge 1:	Yes	Yes	Yes
Judge 2:	No	Yes	No
Judge 3:	Yes	No	No
Majority:	Yes	Yes	No

	p	q	$r \leftrightarrow p \wedge q$	r
Judge 1:	Yes	Yes	Yes	Yes
Judge 2:	No	Yes	Yes	No
Judge 3:	Yes	No	Yes	No
Majority:	Yes	Yes	Yes	No

Thus, we can also work within a framework without integrity constraints ("legal doctrines"), where all inter-relations between propositions stem from the logical structure of those propositions themselves.

And we do not need to distinguish premises from conclusions either.

Formal Framework

Notation: Let $\sim\varphi := \varphi'$ if $\varphi = \neg\varphi'$ and let $\sim\varphi := \neg\varphi$ otherwise.

An *agenda* Φ is a finite nonempty set of propositional formulas (w/o double negation) closed under complementation: $\varphi \in \Phi \;\Rightarrow\; \sim\varphi \in \Phi$.

A *judgment set* J on an agenda Φ is a subset of Φ. We call J:

- *complete* if $\varphi \in J$ or $\sim\varphi \in J$ for all $\varphi \in \Phi$
- *complement-free* if $\varphi \notin J$ or $\sim\varphi \notin J$ for all $\varphi \in \Phi$
- *consistent* if there exists an assignment satisfying all $\varphi \in J$

Let $\mathcal{J}(\Phi)$ be the set of all complete and consistent subsets of Φ.

A finite set of *agents* $\mathcal{N} = \{1, \ldots, n\}$, with $n \geqslant 2$, express judgments on the formulas in Φ, producing a *profile* $\boldsymbol{J} = (J_1, \ldots, J_n)$.

An *aggregation rule* for an agenda Φ and a set of n agents is a function mapping a profile of complete and consistent individual judgment sets to a single collective judgment set: $F : \mathcal{J}(\Phi)^n \to 2^\Phi$.

Example: Majority Rule

The (strict) *majority rule* accepts those proposition that have been accepted by more than half of the agents.

Suppose three agents ($\mathcal{N} = \{1, 2, 3\}$) express judgments on the propositions in the agenda $\Phi = \{p, \neg p, q, \neg q, p \vee q, \neg(p \vee q)\}$.

For simplicity, we only show the positive formulas in our tables:

	p	q	$p \vee q$	formal notation
Agent 1:	Yes	No	Yes	$J_1 = \{p, \neg q, p \vee q\}$
Agent 2:	Yes	Yes	Yes	$J_2 = \{p, q, p \vee q\}$
Agent 3:	No	No	No	$J_3 = \{\neg p, \neg q, \neg(p \vee q)\}$

In our example: $F_{\text{maj}}(\boldsymbol{J}) = \{p, \neg q, p \vee q\}$ [complete and consistent!]

More Aggregation Rules

Various rules have been proposed in the literature. Examples:

- A (uniform) *quota rule* accepts an issue if at least k individuals do (e.g., weak *majority rule* for $k = \lceil \frac{n}{2} \rceil$).
- The *Kemeny rule* returns the rational ballot(s) minimising the sum of the Hamming distances to the individual ballots.
- A *representative-voter rule* returns the "most representative" input ballot (e.g., *average-voter rule* or *plurality-voter rule*).

F. Dietrich and C. List. Judgment Aggregation by Quota Rules: Majority Voting Generalized. *Journal of Theoretical Politics*, 19(4):391–424, 2007.

M.K. Miller and D. Osherson. Methods for Distance-based Judgment Aggregation. *Social Choice and Welfare*, 32(4):575–601, 2009.

U. Endriss and U. Grandi. Binary Aggregation by Selection of the Most Representative Voter. Proc. AAAI-2014.

Basic Axioms

What makes for a "good" aggregation rule F? The following *axioms* all express intuitively appealing (yet, always debatable!) properties:

- *Anonymity:* Treat all agents symmetrically!

  Formally: for any profile $\boldsymbol{J}$ and any permutation $\pi : \mathcal{N} \to \mathcal{N}$ we have $F(J_1, \ldots, J_n) = F(J_{\pi(1)}, \ldots, J_{\pi(n)})$.

- *Neutrality:* Treat all propositions symmetrically!

  Formally: for any φ, ψ in the agenda Φ and any profile $\boldsymbol{J}$, if for all $i \in \mathcal{N}$ we have $\varphi \in J_i \Leftrightarrow \psi \in J_i$, then $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \psi \in F(\boldsymbol{J})$.

- *Independence:* Only the "pattern of acceptance" should matter!

  Formally: for any φ in the agenda Φ and any profiles $\boldsymbol{J}$ and $\boldsymbol{J'}$, if $\varphi \in J_i \Leftrightarrow \varphi \in J'_i$ for all $i \in \mathcal{N}$, then $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \varphi \in F(\boldsymbol{J'})$.

Observe that the *majority rule* satisfies all of these axioms.

(But so do some other rules! Can you think of some examples?)

Impossibility Theorem

We have seen that the majority rule is *not consistent*. Is there some other "reasonable" aggregation rule that does not have this problem? *Surprisingly, no!* (at least not for certain agendas)

Theorem 1 (List and Pettit, 2002) *No judgment aggregation rule for two or more agents and an agenda Φ with $\{p, q, p \wedge q\} \subseteq \Phi$ that satisfies anonymity, neutrality, and independence will always return a complete and consistent judgment set.*

This is the main result in the original paper introducing the formal framework of JA and proposing to apply the axiomatic method.

Remark: Similar impossibilities arise for other agendas.

C. List and P. Pettit. Aggregating Sets of Judgments: An Impossibility Result. *Economics and Philosophy*, 18(1):89–110, 2002.

Proof: Part 1

Notation: $N_\varphi^{\boldsymbol{J}}$ is the set of agents who accept formula φ in profile $\boldsymbol{J}$.

Let F be any aggregator that is independent, anonymous, and neutral.

We observe:

- Due to *independence*, whether $\varphi \in F(\boldsymbol{J})$ only depends on $N_\varphi^{\boldsymbol{J}}$.
- Then, due to *anonymity*, whether $\varphi \in F(\boldsymbol{J})$ only depends on $|N_\varphi^{\boldsymbol{J}}|$.
- Finally, due to *neutrality*, the manner in which the status of $\varphi \in F(\boldsymbol{J})$ depends on $|N_\varphi^{\boldsymbol{J}}|$ must itself *not* depend on φ.

Thus: if φ and ψ are accepted by the same number of agents, then we must either accept both of them or reject both of them.

Proof: Part 2

Recall: For all $\varphi, \psi \in \Phi$, if $|N_\varphi^{\boldsymbol{J}}| = |N_\psi^{\boldsymbol{J}}|$, then $\varphi \in F(\boldsymbol{J}) \Leftrightarrow \psi \in F(\boldsymbol{J})$.

First, suppose the number n of agents is *odd* (and $n > 1$):

Consider a profile $\boldsymbol{J}$ where $\frac{n-1}{2}$ agents accept p and q; one accepts p but not q; one accepts q but not p; and $\frac{n-3}{2}$ accept neither p nor q. That is: $|N_p^{\boldsymbol{J}}| = |N_q^{\boldsymbol{J}}| = |N_{\neg(p \wedge q)}^{\boldsymbol{J}}|$. Then:

- Accepting all three formulas contradicts consistency. ✓
- But if we accept none, completeness forces us to accept their complements, which also contradicts consistency. ✓

If n is *even*, we can get our impossibility even without having to make (almost) any assumptions regarding the structure of the agenda:

Consider a profile $\boldsymbol{J}$ with $|N_p^{\boldsymbol{J}}| = |N_{\neg p}^{\boldsymbol{J}}|$. Then:

- Accepting both contradicts consistency. ✓
- Accepting neither contradicts completeness. ✓

Annotation and Crowdsourcing

Disciplines such as computer vision and computational linguistics require large corpora of annotated data.

Examples from linguistics: grammaticality, word senses, speech acts

People need corpora with *gold standard* annotations:

- set of *items* (e.g., text fragment with one utterance highlighted)
- assignment of a *category* to each item (e.g., it's a *question*)

Classical approach: ask a handful of experts (who hopefully agree).

Modern approach is to use *crowdsourcing* (e.g., Mechanical Turk) to collect annotations: fast, cheap, more judgments from more speakers.

<u>But:</u> how to *aggregate* individual annotations into a gold standard?

U. Endriss and R. Fernández. Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model. Proc. ACL-2013.

Formal Framework

Idea: think of this as a problem of judgment aggregation.

An annotation task has three components:

- infinite set of *agents* N
- finite set of *items* J
- finite set of *categories* K

A finite subset of agents annotate some of the items with categories (one each), resulting is a *group annotation* $A \subseteq N \times J \times K$.

$(i, j, k) \in A$ means that agent i annotates item j with category k.

An *aggregation rule* F maps group annotations to annotations:

$$F : 2^{N \times J \times K}_{<\omega} \to 2^{J \times K}$$

Remark: For $|K| = 2$, collective annotation is like standard judgment aggregation (with atomic propositions only), except that ballots can be incomplete and aggregation rules can be irresolute.

Axioms

Examples for desirable properties of an aggregation rule F (expressed using notation that's handy for highly incomplete inputs):

- *Nontriviality:* $|A \upharpoonright j| > 0$ should imply $|F(A) \upharpoonright j| > 0$
- *Groundedness:* $\mathrm{cat}(F(A) \upharpoonright j)$ should be a subset of $\mathrm{cat}(A \upharpoonright j)$
- *Item-Independence:* $F(A) \upharpoonright j$ should be equal to $F(A \upharpoonright j)$
- *Agent-Symmetry:* $F(\sigma(A)) = F(A)$ for all $\sigma : N \to N$
- *Category-Symmetry:* $F(\sigma(A)) = \sigma(F(A))$ for all $\sigma : K \to K$
- *Positive Responsiveness:* $k \in \mathrm{cat}(F(A) \upharpoonright j)$ and $(i, j, k) \notin A$ should imply $\mathrm{cat}(F(A \cup (i, j, k)) \upharpoonright j) = \{k\}$

Reminder: annotation A, agents $i \in N$, items $j \in J$, categories $k \in K$

Characterisation Result

An elegant characterisation of the most basic aggregation rule (a slight generalisation of May's Theorem):

Theorem 2 (Simple Plurality) *An aggregator F is nontrivial, item-independent, agent-symmetric, category-symmetric, and positively responsive iff F is the simple plurality rule:*

$$F : A \mapsto \{(j, k^\star) \in J \times K \mid k^\star \in \operatorname*{argmax}_{k \in \mathrm{cat}(A \restriction j)} |A \restriction j, k|\}$$

<u>Proof:</u> Omitted.

J. Kruger, U. Endriss, R. Fernández, and C. Qing. Axiomatic Analysis of Aggregation Methods for Collective Annotation. Proc. AAMAS-2014.

Concrete Aggregation Rules

We have three proposals for concrete aggregation rules that are more sophisticated than the simple plurality rule and that try to account for the *reliability of individual annotators* in different ways:

- Bias-Correcting Rules
- Greedy Consensus Rules
- Agreement-Based Rule

Proposal 1: Bias-Correcting Rules

If an annotator appears to be *biased* towards a particular category, then we could try to correct for this bias during aggregation.

- $\text{Freq}_i(k)$: relative frequency of annotator i choosing category k
- $\text{Freq}(k)$: relative frequency of k across the full profile

$\text{Freq}_i(k) > \text{Freq}(k)$ suggests that i is biased towards category k.

A *bias-correcting rule* tries to account for this by varying the weight given to k-annotations provided by annotator i:

- **Diff** (difference-based): $1 + \text{Freq}(k) - \text{Freq}_i(k)$
- **Rat** (ratio-based): $\text{Freq}(k) \,/\, \text{Freq}_i(k)$
- **Com** (complement-based): $1 + 1 \,/\, |K| - \text{Freq}_i(k)$
- **Inv** (inverse-based): $1 \,/\, \text{Freq}_i(k)$

For comparison: the *simple majority rule* SPR always assigns weight 1.

Proposal 2: Greedy Consensus Rules

If there is *(near-)consensus* on an item, we should adopt that choice. And: we might want to classify annotators who disagree as *unreliable*.

The *greedy consensus rule* **GreedyCR**^{t} (with *tolerance threshold* t) repeats two steps until all items are decided:

(1) *Lock in* the majority decision for the item with the strongest majority not yet locked in.

(2) *Eliminate* any annotator who disagrees with more than t decisions.

Variations are possible: any nondecreasing function from disagreements with locked-in decisions to annotator weight might be of interest.

Greedy consensus rules appar to be good at recognising *item difficulty*.

Proposal 3: Agreement-Based Rule

Suppose each item has a *true* category (its *gold standard*). If we knew it, we could compute each annotator i's *accuracy* acc_i.

If we knew acc_i, we could compute annotator i's *optimal weight* w_i (using maximum likelihood estimation, under certain assumptions):

$$w_i \quad = \quad \log \frac{(|K| - 1) \cdot \mathrm{acc}_i}{1 - \mathrm{acc}_i}$$

But we don't know acc_i. However, we can try to *estimate* it as annotator i's *agreement* agr_i with the plurality outcome:

$$\mathrm{agr}_i \quad = \quad \frac{|\{j \in J \mid i \text{ agrees with SPR on } j\}| + 0.5}{|\{j \in J \mid i \text{ annotates } j\}| + 1}$$

The agreement rule **Agr** thus uses weights $w'_i = \log \frac{(|K|-1) \cdot \mathrm{agr}_i}{1 - \mathrm{agr}_i}$.

Empirical Analysis

We have implemented our three types of aggregation rules and compared the results they produce to *existing gold standard* annotations for three tasks in computational linguistics:

- RTE: *recognising textual entailment* (2 categories)
- PSD: *proposition sense disambiguation* (3 categories)
- QDA: *question dialogue acts* (4 categories)

For RTE we used readily available crowdsourced annotations.
For PSD and QDA we collected new crowdsourced datasets.

GreedyCR so far has only been implemented for the binary case.

The crowdsourced data is available here:

`http://www.illc.uva.nl/Resources/CollectiveAnnotation/`

C. Qing, U. Endriss, R. Fernández, and J. Kruger. Empirical Analysis of Aggregation Methods for Collective Annotation. Proc. COLING-2014.

Case Study 1: Recognising Textual Entailment

In RTE tasks you try to develop algorithms to decide whether a given piece of text entails a given hypothesis. Examples:

Text	Hypothesis	GS
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	Yahoo bought Overture.	1
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology.	Israel was established in May 1971.	0

We used a dataset collected by Snow et al. (2008):

- Gold standard: 800 items (T-H pairs) with an 'expert' annotation
- Crowdsourced data: 10 AMT annotations per item (164 people)

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proc. EMNLP-2008.

Case Study 2: Preposition Sense Disambiguation

The PSD task is about choosing the sense of the preposition "*among*" in a given sentence, out of three possible senses from the ODE:

(1) situated more or less centrally in relation to several other things,
e.g., *"There are flowers hidden among the roots of the trees."*

(2) being a member or members of a larger set,
e.g., *"Snakes are among the animals most feared by man."*

(3) occurring in or shared by some members of a group or community,
e.g., *"Members of the government bickered among themselves."*

We crowdsourced data for a corpus with an existing GS annotation:

- Gold standard: 150 items (sentences) from *SemEval 2007*
- Crowdsourced data: 10 AMT annotations per item (45 people)

K.C. Litkowski and O. Hargraves. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. Proc. SemEval-2007.

Case Study 3: Question Dialogue Acts

The QDA task consists in selecting a *question dialogue act*, for a highlighted utterance in a dialogue fragment, out of four possibilities:

(1) **Yes-No:** Questions with a standard form that could be answered with *yes* or *no*, e.g., *"Is that the only pet that you have?"*

(2) **Wh:** Questions with a standard form that ask for specific information using wh-words, e.g., *"What kind of pet do you have?"*

(3) **Declarative:** Questions with a statement-like form that nevertheless ask for an answer, e.g., *"You have how many pets."*

(4) **Rhetorical:** Questions that do not need to be answered, but are asked only to make a point, e.g., *"If I had a pet, how could I work?"*

We crowdsourced data for a corpus with an existing GS annotation:

- Gold standard: 300 questions from the *Switchboard Corpus*
- Crowdsourced data: 10 AMT annotations per item (63 people)

D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL: Shallow-Discourse-Function-Annotation Coders Manual. Univ. of Colorado Boulder, 1997.

Case Studies: Results

How well did we do? Observed *agreement* with the gold standard annotation (any ties are counted as instances of disagreement):

- Recognising Textual Entailment (two categories):
  - SPR: 85.6%
  - Best BCR's: Com 91.6%, Diff 91.5%
  - Agr: 93.3%
  - GreedyCR^0: 86.6%, GreedyCR^{15}: 92.5%
- Preposition Sense Disambiguation (three categories):
  - SPR: 81.3% [caveat: gold standard appears to have errors]
  - Best BCR: Rat 84%, Diff 83.3%
  - Agr: 82.7%
- Question Dialogue Acts (four categories):
  - SPR: 85.7%
  - Best BCR: Inv 87.7% [shared bias $\leadsto$ agent-indep. rules better]
  - Agr: 86.7%

Last Slide

This has been an introduction to *judgment aggregation*, followed by a discussion of applications to *collective annotation*. Topics covered:

- formal framework, aggregation rules, axioms
- doctrinal paradox: majority rule may be inconsistent
- impossibility theorem: no collectively rational rule is A+N+I
- collective annotation: non-binary, highly incomplete, unconstrained
- rules: bias-correcting, greedy, agreement-based
- empirical study: new data available, encouraging results

Note that judgment aggregation is more general than last time's *preference aggregation* (or *graph aggregation*), as we may ask agents to judge propositions of the form "$x \succ y$".

Again, the slides are available online:

`https://staff.science.uva.nl/u.endriss/teaching/paris-2016/`