Axiomatic Analysis of Aggregation Methods for Collective Annotation

Justin Kruger
ILLC, University of Amsterdam
justin.g.kruger@gmail.com

Ulle Endriss
ILLC, University of Amsterdam
ulle.endriss@uva.nl

Raquel Fernández
ILLC, University of Amsterdam
raquel.fernandez@uva.nl

Ciyang Qing
ILLC, University of Amsterdam
qciyang@gmail.com

ABSTRACT

Crowdsourcing is an important tool, e.g., in computational linguistics and computer vision, to efficiently label large amounts of data using nonexpert annotators. The individual annotations collected need to be aggregated into a single collective annotation. The hope is that the quality of this collective annotation will be comparable to that of a traditionally sourced expert annotation. In practice, most scientists working with crowdsourcing methods use simple majority voting to aggregate their data, although some have also used probabilistic models and treated aggregation as a problem of maximum likelihood estimation. The observation that the aggregation step in a collective annotation exercise may be considered a problem of social choice has only been made very recently. Following up on this observation, we show that the axiomatic method, as practiced in social choice theory, can make a contribution to this important domain and we develop an axiomatic framework for collective annotation, focusing amongst other things on the notion of an annotator's bias. We complement our theoretical study with a discussion of a crowdsourcing experiment using data from dialogue modelling in computational linguistics.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence; J.5 [**Arts and Humanities**]: Linguistics

Keywords

Crowdsourcing; Annotation; Computational Social Choice

1. INTRODUCTION

Many fields of science and engineering rely on the availability of annotated data, e.g., images labelled with object names for computer vision or part-of-speech annotations of words in text corpora for computational linguistics. Traditionally, such annotations have been provided by small numbers of experts, each labelling large amounts of data. Today, the availability of crowdsourcing technologies makes it possible to instead collect annotations from large numbers of individuals, who may not be experts and who may each only annotate a small subset of a given dataset. This new technology offers spectacular opportunities but also raises methodological questions. To produce a definitive annotation of the data we need to *aggregate* the individual annotations provided by the participants in a crowdsourcing exercise. The most common approach is to label a given item with the category chosen most often by the individuals. This method is usually referred to as *majority voting* and is often combined with some form of quota requirement [4, 15].

A more sophisticated approach is to use unsupervised methods from machine learning. By using an individual annotator's agreement with the choices made by the full population of annotators, we can estimate certain parameters relating to her competence and then weight her vote accordingly [14, 16]. Such maximum likelihood estimation methods can give good results, but they also do not easily lend themselves to a principled reflection on the rule of aggregation implemented. A useful perspective that does permit such reflection is to consider aggregating crowdsourced data as a problem of *social choice* [6].

Social choice theory is the systematic study of methods of aggregating information provided by individuals into a collective view of that information, with the study of voting rules (aggregating voter preferences to obtain election winners) being the best-known example. In social choice theory, aggregators have been analysed both in terms of the *axioms* (formal renderings of normative desiderata) they satisfy and as maximum likelihood estimators given certain assumptions on the nature of the noisy signals regarding an assumed ground truth received by the individuals [18]. In the crowdsourcing literature, on the other hand, only the maximum likelihood estimation perspective has been considered. Our aim in this paper is to address this imbalance and to initiate an axiomatic analysis of aggregation methods that can be applied in the context of crowdsourcing.

The remainder of the paper is organised as follows. Section 2 presents our formal model for collective annotation. In Section 3 we adapt several axioms from other areas of social choice to collective annotation and in Section 4 we use these axioms to prove two characterisation results. Section 5 is devoted to the notion of *bias:* we formulate axioms for aggregation methods that can handle annotator bias and we define four particularly natural representatives of the class of methods we characterise. In Section 6 we test these methods using data on dialogue modelling in computational lin-

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

guistics collected through crowdsourcing. Finally, Section 7 reviews related work and Section 8 concludes.

2. THE MODEL

In this section we introduce our formal model for collective annotation. We also define a number of domain restrictions regarding the space of possible annotations collected.

2.1 Group Annotations

Let N be an infinite set of *agents* (which we shall also refer to as *annotators* or *individuals*). This represents the (for all practical purposes unlimited) set of individuals that may, in principle, be recruited to take part in an annotation exercise. Furthermore, let J be a finite set of *items* and let K be a finite set of *categories* (possibly including a *'don't know'* category). A finite number of agents annotate some of the items by assigning categories to them. This results in a finite *group annotation* $A \subseteq N{\times}J{\times}K$. Read $(i,j,k) \in A$ as agent i annotating item j with category k.[1]

For a given group annotation A and sets $N' \subseteq N$, $J' \subseteq J$, and $K' \subseteq K$, we refer to subsets of A obtained by means of relevant *restrictions* like this:

$$A \restriction N', J', K' \;:=\; \{(x,y,z) \in A \mid x \in N',\, y \in J',\, z \in K'\}$$

Various liberties will be taken with this notation. These include omitting set brackets for single elements, and omitting whole sets when, say, $N' = N$. Thus, for instance, $A \restriction j = \{(x,y,z) \in A \mid y = j\}$ is the set of annotations concerning item j.

We also introduce the following notation to *extract* relevant information from sets of tuples $A \subseteq N \times J \times K$:

$$\begin{aligned} \mathrm{agt}(A) &:= \{i \mid (i,j,k) \in A\} \\ \mathrm{itm}(A) &:= \{j \mid (i,j,k) \in A\} \\ \mathrm{cat}(A) &:= \{k \mid (i,j,k) \in A\} \end{aligned}$$

Combining our notation for restriction and extraction, we can succinctly express complex concepts, e.g., the set of categories used by agent i is $\mathrm{cat}(A \restriction i)$.

2.2 Domain Restrictions

It is sometimes useful (and appropriate) to assume certain *domain restrictions.* We call a group annotation A *complete* if every agent annotating any item at all has actually annotated all items, i.e., if $A \restriction i, j \neq \emptyset$ for all $(i,j) \in \mathrm{agt}(A) \times J$. A is *category-exclusive* if no agent annotates the same item with more than one category, i.e., if $|A \restriction i,j| \leqslant 1$ for all $(i,j) \in N \times J$. Finally, A is *item-covering* if each item is annotated at least once, i.e., if $A \restriction j \neq \emptyset$ for all $j \in J$.

Throughout this paper, we shall assume that *all group annotations involved are category-exclusive.* This is also the most common scenario in practice: most crowdsourcing exercises will require annotators to choose exactly one category (possibly *'don't know'*) for any item they are presented with. At the same time, it is important to note that there are exceptions. For instance, one of the steps in taxonomy creation tasks via crowdsourcing consists in showing annotators an item (e.g., a piece of text or a picture) and n labels and asking them to select *all* labels that apply to the item [4].

Item-coverage is a very weak restriction. It can always be assumed *by definition*, by simply removing any items from J that have not actually been annotated. However, it will not be necessary to make this assumption here.

Completeness, finally, is a much more demanding restriction, which we usually do *not* want to impose (although we will study complete annotations as an extreme case). Indeed, in practice it would be very costly to collect an annotation from every agent on every item.

2.3 Aggregators

An *aggregator* is a function $F : 2^{N \times J \times K}_{<\omega} \to 2^{J \times K}$, mapping any given group annotation (i.e., any given finite subset of $N \times J \times K$) to a single annotation (i.e., to a subset of $J \times K$), associating every item with a (possibly empty, and ideally singleton) set of categories. We refer to $F(A)$ as the *collective annotation* obtained as the result of aggregating the information in group annotation A. An example for an aggregator is the *simple plurality rule:*[2]

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

That is, when given a group annotation A, the SPR returns an annotation in which each item j is annotated with the category chosen most often for j (note that $|A \restriction j,k|$ is the number of agents choosing category k for item j in A). In case of a (multi-way) tie, j is annotated with all categories receiving maximal support. In case j has not been annotated at all in A, it will not get annotated in $\mathrm{SPR}(A)$ either.[3]

We extend our notation for restriction of and extraction from group annotations to collective annotations. For instance, $\mathrm{itm}(F(A) \restriction k)$ is the set of items annotated with category k in the collective annotation returned by F.

3. FUNDAMENTAL AXIOMS

Next, we introduce several *axioms* that encode simple structural (usually desirable) properties of an aggregator F.

Unless explicitly stated to the contrary, all our axioms apply to all group annotations A, all agents i and i', all items j and j', and all categories k and k' (to improve readability, this kind of quantification is often left implicit).

3.1 Outcome Restrictions

Our first group of axioms impose basic constraints on the outcome of an aggregation. Let us call an aggregator F *decisive* if it assigns a single category to every item: $|F(A \restriction j)| = 1$ for all items $j \in J$. In case A is not item-covering, this is an unreasonably demanding requirement (as it requires us to choose a category also for items that have not been annotated by anyone). This consideration leads to our next defi-

[1] Compare this to the standard notation used in social choice theory: Usually, to model an aggregation problem such as this, we would first define what kind of information an agent can supply (e.g., a preference order, or in our case an annotation of some of the items), and then define the input to the aggregation problem, called the *profile*, as a vector of such individual pieces of information, one for each agent. We could have followed this approach also here [6]. However, as we are specifically interested in *incomplete* annotations, where distinct agents will typically annotate very different (small) parts of the data, our notation is much more flexible than the standard, agent-centric, approach.

[2] This rule is often referred to as *simple majority rule*, although strictly speaking that term should be restricted to scenarios with only *two* possible categories. Note that if we were to drop our assumption of category-exclusivity, then the SPR would become a form of (item-wise) *approval voting* [3].

[3] In this case, $\mathrm{cat}(A \restriction j) = \emptyset$. We take the argmax-operator over the empty set as returning the empty set.

nition: F is *weakly decisive* if $|F(A \restriction j)| = \min\{1, |A \restriction j|\}$ for all items $j \in J$, i.e., if it always assigns a single category—unless the item in question has not been annotated at all (in which case it remains unannotated also in the outcome).

Furthermore, let us call F *nontrivial* if it does not leave items unannotated for which there is at least one individual annotation: $|F(A) \restriction j| > 0$ whenever $|A \restriction j| > 0$. That is, weak decisiveness implies nontriviality.

Observe that the SPR, for instance, is nontrivial, but neither decisive nor weakly decisive: there may be ties.

3.2 Unanimity and Related Properties

Unanimity is a fundamental principle requiring that any decision supported by all members of a community should be implemented. There are several variants of this principle that are of interest here, including at least these two:

- F is *unanimous* if it is the case that, whenever there exists a set $K' \subseteq K$ such that $\mathrm{cat}(A \restriction i, j) = K'$ for all agents $i \in N$, then also $\mathrm{cat}(F(A) \restriction j) = K'$.[4]
- F is *grounded* if $\mathrm{cat}(F(A) \restriction j) \subseteq \mathrm{cat}(A \restriction j)$.

That is, groundedness expresses a *'no new categories'* requirement: no item should be annotated with a category that has not been used by at least one individual for that item (and items that have not been annotated at all should not get assigned any category during aggregation).

PROPOSITION 1. *Any nontrivial aggregator that is grounded must also be unanimous.*

PROOF. Suppose all agents agree on category k for item j. By nontriviality, our aggregator F must annotate j with one or more categories. By groundedness, they can only include categories chosen by at least one agent. But k is the only such category.[5] Hence, F must choose (only) k for j. □

All of the aggregators to be discussed in this paper are grounded (and thus also unanimous).[6]

3.3 Independence and Symmetry Properties

Our intuitively most demanding axiom is independence: F is *item-independent* if $F(A) \restriction j = F(A \restriction j)$ for all $j \in J$. That is, F is item-independent if we can determine the category (or categories) to assign to a given item j by only considering the individual annotations for j.

Two standard axioms in social choice theory are anonymity and neutrality. Anonymity imposes a symmetry requirement on individuals (they should all be treated the same). Neutrality has several possible interpretations in our context: neutrality w.r.t. items means that two items annotated in exactly the same way in the input should also receive the same annotation in the output; neutrality w.r.t. categories means that categories should be treated symmetrically (e.g., if a certain group of agents choosing category k is sufficient for k to win for a given item, then the same should be true for category k'). We formulate all these properties as symmetry requirements:[7]

- F is *agent-symmetric* if $F(\sigma(A)) = F(A)$ for all permutations $\sigma : N \to N$ and all group annotations A.
- F is *item-symmetric* if $F(\sigma(A)) = \sigma(F(A))$ for all permutations $\sigma : J \to J$ and all group annotations A.
- F is *category-symmetric* if $F(\sigma(A)) = \sigma(F(A))$ for all permutations $\sigma : K \to K$ and all group annotations A.

For the definitions above, permutations σ on one component of a tuple are extended to full tuples in the natural manner, e.g., $\sigma(A) = \{(\sigma(i), j, k) \mid (i, j, k) \in A\}$ for $\sigma : N \to N$. Note that in our definition of agent-symmetry (i.e., of anonymity) we do not need to apply σ on the righthand side, because the set $F(A)$ does not refer to agents at all. It is not difficult to see that the SPR satisfies all four axioms above.

PROPOSITION 2. *The four axioms of item-independence, agent-symmetry, item-symmetry, and category-symmetry are mutually independent.*

PROOF (SKETCH). Given that the SPR satisfies all four axioms, to show independence of one axiom w.r.t. the other three, it suffices to identify an aggregator that violates the former but satisfies the latter three. We demonstrate this for the case of item-symmetry. Suppose there are two items (say, 1 and 2). Consider the aggregator F that will assign all those categories to item 1 that were chosen, for that item, by an odd number of agents, and all those categories to item 2 that were chosen, for that item, by an even number of agents. F is item-independent (you can decide the outcome for a given item by only looking at the input for that item), agent-symmetric (all agents are treated the same), and category-symmetric (for a given item, all categories are awarded using the same rule). However, F clearly violates item-symmetry, as items 1 and 2 will receive different collective annotations in case their individual annotations are the same. □

3.4 Monotonicity Properties

Intuitively speaking, if we accept a given category k for a given item j, then if that category receives additional support, we should still accept it.

We say that F is *monotonic* if $(j, k) \in F(A)$ implies $(j, k) \in F(A \cup \{(i, j, k)\})$. Later we shall require a slightly stronger form of monotonicity, inspired by the seminal work of May [13], which in addition stipulates that in case k was a tied winner for j, after receiving additional support k will become the sole winner. Formally, F satisfies *positive responsiveness* if $k \in \mathrm{cat}(F(A) \restriction j)$ and $(i, j, k) \notin A$ together imply $\mathrm{cat}(F(A \cup (i, j, k)) \restriction j) = \{k\}$. That is, when category k is selected for item j and then agent i, who has not previously annotated j at all, now also annotates j with k, then k should become the sole collective annotation for j.

4. CHARACTERISATION RESULTS

In this section, we present two characterisation results. The first is a characterisation of the SPR in terms of some of our axioms. The second is an axiomatic characterisation of the class of all aggregators that can be defined in terms of an assignment of weights to individual annotations.

[4] Note that if A is category-exclusive, then K' will be a singleton.

[5] Note that here we use our assumption of category-exclusiveness.

[6] It is not inconceivable that one might want to violate groundedness: Suppose the available categories are *'good'*, *'medium'*, and *'bad'*; and suppose that half of the agents chose *'good'* for a given item and the other half chose *'bad''*. Then we might want our aggregator to return *'medium'*. But defining such an aggregator in a principled manner requires taking the semantics of the categories into account, which is beyond the scope of this paper.

[7] The need to distinguish between different forms of neutrality has also be recognised in binary (and judgment) aggregation. Specifically, for the special case of two categories, what we call *category-symmetry* has been called *domain-neutrality* before [9].

4.1 The Simple Plurality Rule

Recall the definition of the SPR: it selects those categories for a given item j that are tied for having been selected most often by the agents (with the one exception that items not annotated at all are also not annotated in the outcome).

Theorem 3. *An aggregator is nontrivial, item-independent, agent-symmetric, category-symmetric, and positively responsive if and only if it is the simple plurality rule.*

Proof. Clearly, the SPR meets all five axioms. For the opposite direction, first observe that item-independence means that it is sufficient to prove the claim for the case of a single annotated item j. Then agent-symmetry implies that $F(A) \upharpoonright j$, the collective annotation of j, must be computable from the $|A \upharpoonright j, k|$'s alone, i.e., considering only the *cardinalities* of the sets of agents choosing a given category for j. Furthermore, due to category-symmetry, if $|A \upharpoonright j, k| = |A \upharpoonright j, k'|$ for two categories $k, k' \in K$, then we must have $(j,k) \in F(A) \Leftrightarrow (j,k') \in F(A)$.

Now let F be an arbitrary aggregator that satisfies our five axioms. We need to exclude two possible scenarios where F would differ from the SPR (the scenario of a plurality-winner losing, and that of a plurality-loser winning):

- Suppose $k^+ \in \mathrm{argmax}_{k\in K} |A \upharpoonright j, k|$ and $\left|A \upharpoonright j, k^+\right| \neq 0$. Then $(j, k^+) \in \mathrm{SPR}(A)$. For the sake of contradiction, assume $(j,k^+) \not\in F(A)$. By our characterisation above, no other category k' with $|A \upharpoonright j, k'| = \left|A \upharpoonright j, k^+\right|$ may then be in $\mathrm{cat}(F(A) \upharpoonright j)$ either. Still, by nontriviality, $F(A) \upharpoonright j \neq \emptyset$. So pick any $k \in \mathrm{cat}(F(A) \upharpoonright j)$. Then keep adding new agents i providing the single annotation (i,j,k) to the group annotation, until k^+ and k are used equally often to label j. By positive responsiveness, after the first such addition, k becomes the only category labelling j in the outcome, and this remains the case until the end. At that point we have reached a situation in which k and k^+ are chosen equally often, but have a different status in the outcome. This contradicts our characterisation above, i.e., we have derived our contradiction.
- Now suppose $\left|A \upharpoonright j, k^-\right| < \left|A \upharpoonright j, k^+\right|$, i.e., $(j,k^-) \not\in \mathrm{SPR}(A)$. For the sake of contradiction, assume $(j,k^-) \in F(A)$. By positive responsiveness, if we add one more agent choosing k^-, then k^- will become the only winning category for j. This remains the case if we keep on adding such agents until k^- and k^+ have equal support. But then we have a contradiction with our earlier characterisation, which would require that either both or none of k^- and k^+ are winning.

Hence, F must coincide with the SPR in all cases. □

For the special case of exactly *two categories* (i.e., for binary aggregation) and under the domain restriction of *complete annotations*, Theorem 3 is, essentially, a known result [9]. If we furthermore restrict ourselves to annotations of a single item, then we obtain a variant of May's Theorem [13].

Interestingly, while the SPR satisfies item-symmetry, this axiom is not required for our characterisation. Rather, it is entailed by the other axioms (see also Proposition 2).

We stress that Theorem 3 relies on our assumption of category-exclusivity. A natural direction for future work would be to search for a similar characterisation result without this assumption. We expect that it will be possible to do so by looking for a suitable generalisation of known axiomatisations of approval voting with a variable electorate [7, 17].

4.2 Weighted Plurality Rules

We now want to generalise the SPR and consider *weighted* plurality rules. For instance, we may have high confidence in the competence of agent 1 and want to give her a higher weight than the other agents. Or we may want to give her a higher weight as far as the annotation of items 1–10 are concerned, on which she is an acknowledged expert. Or, in case we observe that agent 2 annotates 90% of all items with category 20, we may want to lower her weight for those instances. In general, we may assign a distinct weight to any triple (i,j,k), and this weight may depend on any feature of the group annotation A we are given. That is, in the most general case, a weighted plurality rule will be defined in terms of a weight function *wt*:

$$wt : 2^{N\times J\times K}_{<\omega} \to (N\times J\times K \to \mathbb{R}^+_0)$$

We write wt_A for $wt(A)$, the function from individual annotations (i,j,k) to weights, as pinpointed by group annotation A. The *weighted plurality rule* F_{wt} with weight function *wt* is defined as follows:

$$F_{wt} : A \mapsto \{(j,k^\star) \in J\times K \mid k^\star \in \operatorname*{argmax}_{k\in \mathrm{cat}(A\upharpoonright j)} \sum_{i\in \mathrm{agt}(A\upharpoonright j,k)} wt_A(i,j,k)\}$$

That is, category $k^\star$ will be chosen for item j if $k^\star$ maximises the weighted sum we get when we add up the weights for every individual annotation of j with a given category, using the weights prescribed by *wt*. As for the SPR, in case no agent annotates j at all, $F_{wt}(A) \upharpoonright j$ will be empty. Observe how F_{wt} reduces to the SPR in case $wt \equiv 1$.

We say that F is a weighted plurality rule if there *exists* a weight function *wt* such that $F \equiv F_{wt}$. Note that any aggregator that is a weighted plurality rule can be defined in terms of many different weight functions; in particular, we have $F_{wt} \equiv F_{c\cdot wt}$ for any constant $c \in \mathbb{R}^+$.

Theorem 4. *An aggregator is nontrivial and grounded if and only if it is a weighted plurality rule.*

Proof. First, F_{wt} is certainly nontrivial and grounded for any choice of *wt*. This follows immediately from the properties of the argmax-operator. For the other direction, take an arbitrary nontrivial and grounded aggregator F. We need to devise a function *wt* such that $F \equiv F_{wt}$. By nontriviality, if $F(A) \upharpoonright j = \emptyset$, then there are no annotations containing j, so any weight function gives the correct (empty) set of outcomes. We only need to ensure that, if $F(A) \upharpoonright j \neq \emptyset$, then the weight function returns the correct outcomes.

Fix an order $\gg$ on N. Define $wt_A(i,j,k) := 1$ if $(j,k) \in F(A)$ and i is the $\gg$-first agent in N with $(i,j,k) \in A$. Note that there must be at least one such agent, because of groundedness (if not, we could not have $(j,k) \in F(A)$). Define all other weights to be 0. Then the sum of weights for (j,k) will be 1 exactly when we want (j,k) to be part of the outcome, and 0 otherwise. □

Of course, the class of weighted plurality rules is huge and includes many unattractive aggregators. In the sequel, we will focus on a specific class of weighted rules with intuitively appealing features. Thanks to Theorem 4, when discussing such rules, we may switch freely between, on the one hand, descriptions of aggregators themselves and, on the other, descriptions of the weight functions defining them.

5. BIAS CORRECTION

Annotator bias is a common problem, not only in crowdsourcing, but also in traditionally sourced expert annotations. For instance, an annotator may have misunderstood the instructions given, she may use a faulty heuristic to make annotations quickly, or she may be an outright spammer simply annotating all or most items with the same category. In this section, we want to give a basic axiomatic account of the phenomenon of bias, complementing the growing amount of work on probabilistic models of bias [2, 16].

We want to distinguish bias from mere lack of reliability: an agent who is biased towards a certain category will systematically overuse that category, an agent who is biased against a category will systematically underuse it, while an agent that is merely unreliable will not display such a clear pattern in their mistakes. Still, bias is more than the agent in question just often (or rarely) using a given category; if a category is highly prevalent in the data to begin with (which to a certain extent will be reflected by the frequency with which it is chosen by other annotators), then a heavy use of that category does not necessarily constitute a mistake.

As we saw in Section 3.4, the axiom of monotonicity constrains the outcomes of an aggregator w.r.t. two neighbouring group annotations that only differ in terms of a single individual annotation. There are also other situations where the addition (or the exchange) of a single individual annotation can provide additional support for the currently winning category for a given item. We will now identify several of them that relate to the notion of bias.

5.1 Category Prevalence and Scarcity

Consider the following scenario:

> Suppose our aggregator has assigned category k to item j. Now we observe an agent i' (not involved in the annotation of j) annotating item j' (different from j) with k. This provides additional evidence that k is a common category: the *prevalence* of category k has increased. Thus, we should have increased confidence in any agent i who chose k for j (more specifically, we should have increased confidence in her *recall* for items of category k). Hence, we should certainly keep our collective annotation of k for j.

Similarly, if agent i' were to *delete* an annotation using category k' (different from k), then k' would become more *scarce*, and again our confidence in any agent choosing k should increase. These considerations lead to two new axioms:

- F is *prevalence-sensitive* if $(j,k) \in F(A)$ implies $(j,k) \in F(A \cup \{(i',j',k)\})$ for $i' \not\in \mathrm{agt}(A \upharpoonright j)$, $j \neq j'$.
- F is *scarcity-sensitive* if $(j,k) \in F(A \cup \{(i',j',k')\})$ implies $(j,k) \in F(A)$ for $i' \not\in \mathrm{agt}(A \upharpoonright j)$, $j \neq j'$, $k \neq k'$.

5.2 Category Overuse and Underuse

Now consider this scenario:

> Suppose on current evidence our aggregator has assigned category k to item j, and agent i was one of the annotators who labelled j with k. Now we observe agent i annotating a different item j' with a different category k'. This extra evidence suggests that i is less biased towards k (more specifically, we should have increased confidence in her *precision* as far as k is concerned). Thus, after our observation the aggregator should still assign k to item j.

That is, adding a k'-annotation is evidence for k'-*overuse*, which should increase our confidence in k-choices. Similarly, if i *deletes* one of her k-annotations elsewhere, then this suggests an *underuse* of k, and our confidence in her k-choices should again increase. We formulate two axioms:

- F is *overuse-sensitive* if $(j,k) \in F(A)$ implies $(j,k) \in F(A \cup \{(i,j',k')\})$ for $(i,j,k) \in A$, $i \not\in \mathrm{agt}(A \upharpoonright j')$, $j \neq j'$, $k \neq k'$.
- F is *underuse-sensitive* if $(j,k) \in F(A \cup \{(i,j',k)\})$ implies $(j,k) \in F(A)$ for $(i,j,k) \in A$, $i \not\in \mathrm{agt}(A \upharpoonright j')$, $j \neq j'$.

It could be argued that these last two axioms are too strong. Consider the second one: as i deletes her k-annotation of j', not only does her use of k decrease (which we argued should increase our confidence into her remaining k-annotations), but at the same time k becomes more scarce in general, which should decrease our confidence in non-k choices made by other agents. We may not wish to commit to one of these two effects necessarily outweighing the other. The following variants of our axioms account for this point. They are designed to apply to pairs of group annotations in which the overall number of k-annotations (i.e., the prevalence/scarcity of k) remains constant:

- F is *weakly overuse-sensitive* if $(j,k) \in F(A \cup \{(i',j',k')\})$ implies $(j,k) \in F(A \cup \{(i,j',k')\})$ for $(i,j,k) \in A$, $i \neq i'$, $i' \not\in \mathrm{agt}(A \upharpoonright \{j,j'\})$, $j \neq j'$, $k \neq k'$.
- F is *weakly underuse-sensitive* if $(j,k) \in F(A \cup \{(i,j',k)\})$ implies $(j,k) \in F(A \cup \{(i',j',k)\})$ for $(i,j,k) \in A$, $i \neq i'$, $i' \not\in \mathrm{agt}(A \upharpoonright \{j,j'\})$, $j \neq j'$.

5.3 The Space of Bias-Correcting Rules

We are now ready to offer a definition of *bias*, by means of fixing a class of aggregators that are able to correct for bias. First, we want any such aggregator to satisfy the axioms regarding prevalence/scarcity and (the weak variants of) over/underuse defined above. Second, we want any such aggregator to satisfy all of the fundamental axioms defined in Section 3—with the sole exception of item-independence, which is in direct conflict with the basic idea of using information gathered for one item to improve the aggregation outcome for another item.[8] That is,

> F is a *bias-correcting rule* (BCR) if it is nontrivial, grounded, agent-symmetric, item-symmetric, category-symmetric, monotonic, prevalence-sensitive, scarcity-sensitive, weakly overuse-sensitive, and weakly underuse-sensitive.

The space of BCR's is not small: we may vary the significance we ascribe to the different components of bias (prevalence, scarcity, overuse, underuse) and we may vary the degree to which we are willing to violate the independence axiom. By Theorem 4, any BCR must be a weighted plurality rule, so we can define any such rule in terms of a suitable weight function. In previous work [6], we have (for the case of *binary* categories) identified what arguably are three of the most natural representatives of the class of BCR's; here we generalise these definitions and add a

[8] One additional, very minor, exception is that we use monotonicity rather than the slightly more demanding positive responsiveness (because we do not want to insist on even the weakest form of additional support to necessarily have the power to break ties, e.g., when that support comes from a highly biased annotator).

Diff	*difference-based BCR*	$1 + \mathrm{Freq}(k) - \mathrm{Freq}_i(k)$
Rat	*ratio-based BCR*	$\mathrm{Freq}(k)/\mathrm{Freq}_i(k)$
Com	*complement-based BCR*	$1 + 1/\lvert K\rvert - \mathrm{Freq}_i(k)$
Inv	*inverse-based BCR*	$1/\mathrm{Freq}_i(k)$

Table 1: Weights used for canonical BCR's.

fourth rule. To define these rules, we use the *global frequency* $\mathrm{Freq}(k) := \frac{|A\restriction k|}{|A|}$ of category k to measure prevalence and scarcity, and we use the *individual frequency* $\mathrm{Freq}_i(k) := \frac{|A\restriction i,k|}{|A\restriction i|}$ of agent i in using category k to measure over- and underuse.[9] Our four rules are defined in Table 1 in terms of the corresponding weight function by fixing the weight assigned to annotation (i, j, k).[10] That is, the difference-based BCR, for instance, is the aggregator F_{wt} with $wt_A : (i, j, k) \mapsto 1 + \mathrm{Freq}(k) - \mathrm{Freq}_i(k)$.

It is not difficult to verify that all four rules meet all the axioms defining the class of BCR's.

The weight functions of all four rules are monotonically decreasing in the individual frequencies; Diff and Rat are also monotonically increasing in the global frequencies. The functions of Table 1 are particularly simple functions with these monotonicity properties; hence our claim that our four rules are natural representatives of the class of BCR's.

For Diff and Rat, if an agent's individual frequency for category k is equal to the global frequency of k, then her weight for k is 1, i.e., in this case the rules coincide with the SPR. Com and Inv coincide with Diff and Rat, respectively, in case the global frequencies of all categories are the same, i.e., in case $\mathrm{Freq}(k) = \frac{1}{|K|}$ for all $k \in K$. Indeed, the simpler rules (ignoring $\mathrm{Freq}(k)$) may be preferred, if we do not want to assume that it is possible to estimate the gold standard frequency of a category from its observed global frequency in the group annotation.

Next, we show how two simple axioms (on weight functions) allow us to neatly separate our four rules:

- Call F *agent-independent* if there exists a weight function wt with $F \equiv F_{wt}$ that satisfies the following property for all A and all (i, j, k):

  $$wt_A(i, j, k) \quad = \quad wt_{A\restriction i}(i, j, k)$$

  Com and Inv satisfy agent-independence, while Diff and Rat do not. That is, for the former two we can calculate the weight given to (i, j, k) by only considering the annotations of agent i, while for the latter two we also need to take the annotations of the other agents into account (to compute $\mathrm{Freq}(k)$).[11]

- Call F *weight-bounded* if there exist a weight function wt with $F \equiv F_{wt}$ and a constant $c \in \mathbb{R}^+$ such that the following property holds for all A and all (i, j, k):

  $$wt_A(i, j, k) \quad \leqslant \quad c$$

  Both Diff and Com are weight-bounded with $c = 2$. Rat and Inv, on the other hand, are not weight-bounded. For Inv, if agent i annotates item j with k and all other items with k', then $\mathrm{Freq}_i(k) = \frac{1}{|J|}$, i.e., the weight for (i, j, k) will be $|J|$, which is not bounded from above. For Rat, if i annotates j with k and all other items with k', while all other agents annotate all items with k, then $\mathrm{Freq}(k) = \frac{(|I|-1)\cdot|J|+1}{|I|\cdot|J|}$ and $\mathrm{Freq}_i(k) = \frac{1}{|J|}$, i.e., (i, j, k)'s weight is in $\Omega(|J|)$.

A rule that is not weight-bounded allows an agent to have an arbitrarily strong influence on the collective annotation of one item (albeit at the expense of losing influence on many other items), while weight-bounded rules put clear limits on such trade-off effects.

5.4 Complete Annotations

We conclude our discussion of the axiomatics of BCR's with an intriguing result showing that for the special case of *complete annotations* and *binary categories* (i.e., for $|K| = 2$), the difference-based BCR reduces to the SPR.[12]

PROPOSITION 5. *For binary categories, if all agents annotate all items, then Diff and SPR return the same result.*

PROOF. Let $K = \{0, 1\}$. Category 1 will win for a given item j under Diff if and only if the following holds:

$$\sum_{i\in \mathrm{agt}(A\restriction j,1)} 1 + \frac{|A\restriction 1|}{|A|} - \frac{|A\restriction i,1|}{|A\restriction i|} \geqslant \sum_{i\in \mathrm{agt}(A\restriction j,0)} 1 + \frac{|A\restriction 0|}{|A|} - \frac{|A\restriction i,0|}{|A\restriction i|}$$

Using $\frac{|A\restriction 0|}{|A|} = 1 - \frac{|A\restriction 1|}{|A|}$ and $\frac{|A\restriction i,0|}{|A\restriction i|} = 1 - \frac{|A\restriction i,1|}{|A\restriction i|}$ (i.e., using the fact that there are exactly two categories, with no possibility of abstaining), we rewrite:

$$\sum_{i\in \mathrm{agt}(A\restriction j,1)} 1 + \frac{|A\restriction 1|}{|A|} - \frac{|A\restriction i,1|}{|A\restriction i|} \geqslant \sum_{i\in \mathrm{agt}(A\restriction j,0)} 1 - \frac{|A\restriction 1|}{|A|} + \frac{|A\restriction i,1|}{|A\restriction i|}$$

Pushing all terms involving $\frac{|A\restriction 1|}{|A|}$ to the left and all those involving $\frac{|A\restriction i,1|}{|A\restriction i|}$ to the right, we get:

$$|A\restriction j,1| + |A\restriction j| \cdot \frac{|A\restriction 1|}{|A|} \quad \geqslant \quad |A\restriction j,0| + \sum_{i\in \mathrm{agt}(A\restriction j)} \frac{|A\restriction i,1|}{|A\restriction i|}$$

If we can simplify this further to $|A\restriction j,1| \geqslant |A\restriction j,0|$, i.e., to the winning condition for 1 under the SPR, then we are done. That is, we are done if we can show:

$$|A\restriction j| \cdot \frac{|A\restriction 1|}{|A|} \quad = \sum_{i\in \mathrm{agt}(A\restriction j)} \frac{|A\restriction i,1|}{|A\restriction i|}$$

As every agent annotates every item exactly once, we have $\mathrm{agt}(A\restriction j) = N$, $|A\restriction j| = |N|$, $|A\restriction i| = |I|$, and $|A| = |N|\cdot|I|$. Hence, we can rewrite as follows:

$$|A\restriction 1| \quad = \quad \sum_{i\in N} |A\restriction i,1|$$

But this is immediately seen to be true, so we are done. □

This result illustrates that aggregation problems with highly incomplete profiles are qualitatively different from the more commonly studied case in which all individuals are asked exactly the same questions: rules that are not distinguishable in the latter case may differ greatly in the former.

[9] We shall assume $|A| \neq 0$. Although $|A\restriction i| = 0$ for some $i \in N$, we will never need to compute the individual frequency for those i.

[10] The rule Com had originally been defined using weights $1 - \mathrm{Freq}_i(k)$ [6]. Arguably, adding $1/|K|$ is a more natural choice, e.g., it ensures that Com and Diff coincide when the global frequencies of all categories are the same.

[11] An alternative way of separating Com/Inv from Diff/Rat would be the non-weak variants of overuse- and underuse-sensitivity.

[12] The same is not true for Rat, Com, or Inv.

	SPR	Com	Inv	Diff	Rat
Overall	.857	.870	.877	.867	.870
Yes-No	.86/.98	.87/.98	.91/.91	.84/.98	.84/.99
Wh	.87/1.0	.87/1.0	.94/.98	.87/1.0	.87/1.0
Declarative	.92/.75	.88/.77	.84/.77	.89/.78	.92/.77
Rhetorical	.90/.42	.88/.49	.72/.73	.91/.44	.91/.47

Table 2: Observed agreement with the gold standard and precision/recall per category for different rules.

6. CASE STUDY

In this section, we report on the results of an experimental case study in which we have tested the BCR's of Table 1.

6.1 Data Collected

To carry out this study, we created a new dataset of crowdsourced annotations using the Switchboard Corpus [8]—a corpus of telephone conversations that includes a gold standard annotation assigning a *dialogue act* type (such as *assert*, *answer*, *reject*) to each utterance [10]. We restricted ourselves to four types of question dialogue acts: *yes-no questions* (e.g., "Do you sell your projects?"), *wh-questions* (e.g., "What area do you live in?"), *declarative questions* (e.g., "I was wondering if all vans did that."), and *rhetorical questions* (e.g., "How high are the taxes going to be when my children are my age? That's the scary thing.").

We extracted 300 questions from the corpus, 35% of which were tagged as yes-no in the gold standard annotation, 30% as wh, 20% as declarative, and 15% as rhetorical. We then used Amazon's *Mechanical Turk* (AMT) to collect 10 non-expert annotations for each of the 300 items. Each item consisted of a short dialogue fragment such as the one below and the AMT workers were asked to classify the highlighted question using one of the four question categories.

A: You know, because he's had all this room to run in.
B: **Well, how did he get out?**
A: He dug a hole under the fence.
B: Oh, boy.

A total of 63 AMT workers took part in the annotation task, each of them annotating between 10 items (24 annotators) and 200 items (only one annotator). Amongst the nonexpert annotations, the relative frequencies per category were approximately 37% for yes-no, 34% for wh, 18% for declarative, and 11% for rhetorical questions.

6.2 Results

We then applied our four bias-correcting rules to this data and compared their performance to the SPR. The results are shown in Table 2. The first row shows the overall observed agreement with the gold standard. We can see that all BCR's outperform the SPR. (There were 7 ties for the SPR, which we count as disagreements.) The remaining rows show precision and recall for each category.[13]

We can see that the AMT workers tend to overuse the most prevalent categories (yes-no and wh), resulting in high recall but lower precision. In contrast, the less frequent categories (declarative and rhetorical) tend to be underused, resulting in high precision but low recall. Note that Inv, the only rule that is both agent-independent and weight-unbounded, is distinct from the other rules in that it has a substantially higher recall for rhetorical questions (73%). This can be explained by two features of the data collected: First, the AMT workers displayed a *common* bias against labelling items as rhetorical questions, rather than just some individual annotators displaying an *individual* bias. Therefore, agent-dependent rules such as Diff and Rat that attempt to temper the effects of the individual frequencies observed by relating them to the corresponding global frequencies result in outcomes that tend towards the SPR outcome. Agent-independent rules, on the other hand, correctly discount votes for high-frequency categories also in this case. Second, while some rhetorical questions were easy to recognise and received clear majorities for all rules, others were particularly hard to spot and only had a chance of winning if a few strong annotators were able to turn around majorities. This is only possible for weight-unbounded rules.

The price to pay for the good performance of Inv in recalling rhetorical questions is, naturally, the drop in precision compared to the other rules. The dual effect is that the precision for yes-no and wh is higher with Inv than with other rules, while the recall is lower.

If we compare the performance of our rules in terms of their *F-score* (the harmonic mean of precision and recall), the most striking finding is that for rhetorical questions (the category that AMT workers had most difficulty recognising), all four bias-correcting rules (F-scores between 0.59 for Diff and 0.73 for Inv) outperform the SPR (F-score 0.57).

7. RELATED WORK

The potential of using principles of social choice theory to aggregate information obtained via crowdsourcing has been noted before [6, 12]. Mao et al. [12], for instance, have made proposals for identifying realistic models of distortion for crowdsourcing experiments to be able to apply the maximum likelihood approach of social choice theory more effectively. They focus on aggregation problems where agents provide rankings of possible answers and a single best answer needs to be selected (as in classical voting theory), so their results are not immediately relevant here. Instead, our work may be considered a refinement of the model of ('plain') collective annotation put forward in our own previous work [6].

Our model is similar to *binary aggregation* as studied in (computational) social choice [5, 9]. The main difference is the latter's restriction to binary categories and to complete profiles. As we have argued in Section 5.4, not assuming completeness results in a richer landscape of aggregators.

There are also connections to *voting in combinatorial domains* [11]. The main difference is that for the latter, ballots (corresponding to our individual annotations) have a more complex structure, e.g., they might be CP-nets describing an agent's preferences over the range of possible choices, while in our model each agent only contributes their top choice.

We took 'bias' as bias *for* (or against) a given category, but one could also adopt a broader perspective. This has been done by Artstein and Poesio [2], who define bias of a group annotation as the average variance of the frequencies for the categories. As for Diff and Rat, this means that bias is taken to reduce when global and individual category frequencies become more alike. Note that the *use* of the concept of bias by Artstein and Poesio is rather different from ours: they use it to quantify the quality of a group annotation (using so-called measures for inter-annotator agreement), while we use it to define aggregation methods.

[13] The *precision* of rule F for category k is the proportion of items classified as k by F on which the gold standard agrees. The *recall* of F for k is the proportion of items classified as k by the gold standard for which F returns k as well.

8. CONCLUSION

We have shown that the axiomatic method of social choice theory can make a contribution to the study of aggregation methods used in crowdsourcing and, more generally, in collective annotation. We have done so by (1) axiomatically characterising two base-line methods (the simple plurality rule and the class of all rules that can be described in terms of a weight function); by (2) proposing a class of methods, described in terms of axioms, that are concerned with addressing annotator bias; and by (3) discussing how certain axioms can explain the difference in performance of different aggregators on real data.

A restriction of our model is that we assume that the categories available for use are the same for each and every item. This may not be appropriate for certain types of annotation tasks (although it is for many). For instance, in word-sense labelling tasks in computational linguistics, annotators need to choose the right sense amongst the possible senses for each word [14, 15]. Of course, we may think of K as the union of all sets of categories needed for all items, but this will hardly be the most natural way of modelling the problem. Extending our model to handle different sets of categories for different items is possible in principle [6], but requires great care, for instance, when defining category-symmetry.

A second restriction, which we had already discussed, is our assumption of category-exclusivity. If we wanted to drop this assumption, our plurality rule would become a form of (item-wise) approval voting. Future work aimed at extending our model in this manner should carefully consider the semantics of annotations of the same item by the same agent with multiple categories. For example, if by choosing several categories together an agent is expressing uncertainty between those categories, a rule such a *even-and-equal cumulative voting*, where the voter evenly distributes 1 point over all approved alternatives, might seem appropriate. On the other hand, if an annotation may belong to more than one category and an agent is expressing this fact through their multi-category annotation, classical approval voting (giving 1 full point to each of the chosen alternatives) would seem more appropriate. In general, the full class of *size approval voting rules* may be of interest [1].

Finally, we should stress that our bias-correcting rules are not intended to be a 'one-fits-all' approach to collective annotation. For instance, they rely on the inherent assumption that every annotator is sincerely trying to provide an accurate annotation. This clearly is not the case in most crowdsourcing exercises. If we remove the most obvious spammers from a group annotation (the agents that disagree the most with the plurality choice, for instance), then we can easily improve the performance of our rules. The reason is that, while agents that, say, always choose category 1, are easily picked out by our rules, agents that simply annotate at random are not. If there are too many agents of the latter type, this significantly reduces the quality of the observed global frequency as an estimator for gold standard frequency, and thus negatively affects the performance of aggregation methods such as Diff and Rat.

Acknowledgements. We are grateful for the feedback received from four anonymous AAMAS reviewers.

9. REFERENCES

[1] J. Alcalde-Unzu and M. Vorsatz. Size approval voting. *Journal of Economic Theory*, 144(3):1187–1210, 2009.

[2] R. Artstein and M. Poesio. Bias decreases in proportion to the number of annotators. *Proc. FG-MoL: Formal Grammar and Mathematics of Language*, 2005.

[3] S. J. Brams and P. C. Fishburn. Voting procedures. In K. J. Arrow, A. K. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, pages 173–236. North-Holland, 2002.

[4] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proc. Conference on Human Factors in Computing Systems (CHI)*, 2013.

[5] E. Dokow and R. Holzman. Aggregation of binary evaluations. *Journal of Economic Theory*, 145(2):495–511, 2010.

[6] U. Endriss and R. Fernández. Collective annotation of linguistic resources: Basic principles and a formal model. In *Proc. 51st Annual Meeting of the Association for Computat. Linguistics (ACL)*, 2013.

[7] P. C. Fishburn. Axioms for approval voting: Direct proof. *Journal of Economic Theory*, 19(1):180–185, 1978.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[9] U. Grandi and U. Endriss. Lifting integrity constraints in binary aggregation. *Artificial Intelligence*, 199–200:45–66, 2013.

[10] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report, University of Colorado at Boulder, 1997.

[11] J. Lang. Logical preference representation and combinatorial vote. *Annals of Mathematics and Artificial Intelligence*, 42(1–3):37–71, 2004.

[12] A. Mao, A. D. Procaccia, and Y. Chen. Better human computation through principled voting. In *Proc. 27th AAAI Conference on Artificial Intelligence*, 2013.

[13] K. O. May. A set of independent necessary and sufficient conditions for simple majority decisions. *Econometrica*, 20(4):680–684, 1952.

[14] R. J. Passonneau, V. Bhardwaj, A. Salleb-Aouissi, and N. Ide. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252, 2012.

[15] N. Venhuizen, V. Basile, K. Evang, and J. Bos. Gamification for word sense labeling. In *Proc. 10th International Conference on Computational Semantics (IWCS)*, 2013.

[16] F. L. Wauthier and M. I. Jordan. Bayesian bias mitigation for crowdsourcing. In *Proc. 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

[17] Y. Xu. Axiomatizations of approval voting. In M. R. Sanver and J.-F. Laslier, editors, *Handbook on Approval Voting*, pages 91–102. Springer-Verlag, 2010.

[18] H. P. Young. Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244, 1988.