A General Model for Aggregating Annotations AcrossSimple, Complex, and Multi-object Annotation Tasks (Abstract Reprint)

Alexander Braylan[1], **Madalyn Marabella**[1], **Omar Alonso**[2], **Matthew Lease**[3]

[1]Dept. of Computer Science, University of Texas at Austin
[2]Amazon
[3]School of Information, University of Texas at Austin

Abstract Reprint. This is an abstract reprint of a journal article by Braylan, Marabella, Alonso, and Lease (2023).

Abstract

Human annotations are vital to supervised learning, yet annotators often disagree on the correct label, especially as annotation tasks increase in complexity. A common strategy to improve label quality is to ask multiple annotators to label the same item and then aggregate their labels. To date, many aggregation models have been proposed for simple categorical or numerical annotation tasks, but far less work has considered more complex annotation tasks, such as those involving open-ended, multivariate, or structured responses. Similarly, while a variety of bespoke models have been proposed for specific tasks, our work is the first we are aware of to introduce aggregation methods that generalize across many, diverse complex tasks, including sequence labeling, translation, syntactic parsing, ranking, bounding boxes, and keypoints. This generality is achieved by applying readily available task-specific distance functions, then devising a task-agnostic method to model these distances between labels, rather than the labels themselves.

This article presents a unified treatment of our prior work on complex annotation modeling and extends that work with investigation of three new research questions. First, how do complex annotation task and dataset properties impact aggregation accuracy? Second, how should a task owner navigate the many modeling choices in order to maximize aggregation accuracy? Finally, what tests and diagnoses can verify that aggregation models are specified correctly for the given data? To understand how various factors impact accuracy and to inform model selection, we conduct large-scale simulation studies and broad experiments on real, complex datasets. Regarding testing, we introduce the concept of unit tests for aggregation models and present a suite of such tests to ensure that a given model is not mis-specified and exhibits expected behavior.

Beyond investigating these research questions above, we discuss the foundational concept and nature of annotation complexity, present a new aggregation model as a conceptual bridge between traditional models and our own, and contribute a new general semisupervised learning method for complex label aggregation that outperforms prior work.

References

Braylan, A.; Marabella, M.; Alonso, O.; and Lease, M. 2023. A General Model for Aggregating Annotations Across Simple, Complex, and Multi-Object Annotation Tasks. *Journal of Artificial Intelligence Research*, 78: 901–973.