Annotation Quality Framework - Accuracy, Credibility, and Consistency

Liliya Lavitas
Twitter, llavitas@twitter.com
Olivia Redfield
Twitter, oredfield@twitter.com
Allen Lee
Twitter, allenl@twitter.com
Daniel Fletcher
Twitter, dfletcher@twitter.com
Matthias Eck
Twitter, meck@twitter.com
Sunil Janardhanan
Twitter, sjanardhanan@twitter.com

November 20, 2021

Abstract

Success of many machine learning and offline measurement efforts is highly dependent on the quality of labeled data that they use. Development of supervised machine learning models and quantitative research rely on the assumption of annotation obtained through human reviewers being "ground truth". Annotation quality issues result in violation of this assumption and corrupt quality of all the downstream work and analysis. Through a series of analyses we have identified a highly pressing need for development of a quality framework that will allow creation of a robust system of label quality monitoring and improvement. In this paper we will present an overview of the Accuracy, Credibility, and Consistency (ACC) framework, which consists of three elements: (1) understanding of what annotation quality is and what metrics are required to be tracked (2) implementation of the concepts and measurements and (3) intervention protocols for identified annotation quality issues.

1 Introduction

Human labeling is an essential component of machine learning model development and offline measurement efforts. It is important that the data is high quality: low quality or unreliable data can yield poor model performance or unreliable

measurement results [10]. Especially in the case of continuous measurement systems, annotations must remain consistent and stable over time in order to produce reliable metrics.

The definition of data quality may differ depending on the constraints of a given annotation task, but for the purpose of this framework, we define quality as how well annotators can agree with themselves and one another in a manner that is aligned with task guidelines.

In this paper we introduce the ACC framework, which stands for Accuracy, Credibility, and Consistency. We will discuss: (1) definition of the facets of annotation quality with an illustrative example, (2) implementation of quality metrics, (3) intervention protocols for quality issues.

We want to highlight three important contributions of the proposed framework: (a) it has been designed as problem-agnostic, i.e. it is applicable for multiple human-annotated programs, (b) it covers temporal aspects of annotation quality, thus is applicable for not only one-time data collections, but for continuous programs as well, (c) it allows iterative or partial development because the facets of annotation quality can be implemented and measured independently of one another.

2 Key concepts of the framework

In this section we discuss aspects of annotation quality measured in the ACC framework.

2.1 Data Objects

To build a robust measurement system we need to define which unit of measurements we will consider as atomic. Atomic units can be used as building blocks for all the quality measurements for all data structures of interest. Natural candidates are either: (1) an object reviewers are being asked to annotate or (2) an individual reviewer's decision.

If each object is being annotated more than once (for example N reviewers are annotating each object and then a verdict is being made), then using individual reviewers' decisions as building blocks allow measurement of individual reviewer's quality. On the other hand, the choice of reviewed objects as atomic units has a benefit of simplicity. In the case of each object being annotated by only one reviewer these two approaches will yield same statistics and estimates.

2.2 Measurements

Whenever human-annotated data collection is not a one-time task, but a continuous program (for example for machine model retraining) it is not only initial quality of the data annotation that is of interest, but also changes in annotation quality over time. Changes in annotation quality undermine long-term usability of metrics derived from human-annotated data since changes in a metric might

not be due to real changes in underlying data, and cause biases in machine learning models unless detected quickly.

Thus, in the ACC framework we are focusing on four key metrics of measuring quality:

- Accuracy - annotations' alignment with gold standard, and thus task guidelines
- Credibility - likelihood of object being annotated correctly
- Longitudinal Consistency - stability of annotations over time
- Instant Consistency - reviewers agreement for objects receiving more than one review at the same time

Accuracy, credibility, and instant consistency can be computed either one time or with some predefined periodicity to assess their dynamic over time. Longitudinal consistency is a temporal metric and requires assessment over time. Each metric is described in detail below. In section 3 we will present an illustrative example of such metrics.

2.2.1 Accuracy

Accuracy is a commonly used method for understanding quality of annotations [1]. The Accuracy metric evaluates how closely the annotations are aligned with a gold standard. To evaluate such alignment expert reviewing is required to establish "ground truth" annotations to be compared against the annotations [8]. In order to allow the metric to be problem-agnostic, our design requires implementers to publish their results in a consistent format that counts the correct and incorrect occurrences of each decision in their annotations. This allows each implementation team to use their expertise in defining specifics of accuracy metrics as they see fit for their problem area. Thus the ACC framework is applicable for different kinds of labels, including: binary labels, multi-class classifications, etc. Generally speaking there are multiple accuracy metric approaches, such as *precision*, *recall*, $F1$ score, AUC, loss functions, etc. Different metrics are appropriate for different tasks. For an example of binary labels we use *precision* and *recall* to evaluate the accuracy of annotations.

2.2.2 Credibility

The Credibility metric evaluates the likelihood of an object being annotated correctly. To understand the difference between credibility and accuracy, consider the following example: suppose each object is being annotated by N reviewers using a *majority rule* to assign a verdict. In the case of a correct final verdict coming from a split vote with one tie-breaker one can consider such annotation to be accurate, but with low credibility. Alternatively all reviewers can agree on a decision counter to the guidelines. Another way to interpret credibility is an expected accuracy of an annotation conditioned on reviewers' individual

decisions. Credibility of individual annotations can be computed as a share of reviewers who have agreed with the final object annotation. In our example of N reviewers using a *majority rule* to assign a verdict, credibility is taking values from: $(\lfloor N/2 \rfloor + 1)/N$ to 1. Usability and informativeness of this metric is proportional to the number of individual judgments each object is receiving. For the edge case of each object receiving only one annotation, the credibility of an annotation will always be equal to 1.

For evaluation of credibility of the dataset we propose using bootstrapping [7]: (1) for each object we resample individual reviewers decisions with replacement to create a set of hypothetical possible decisions this object could have received; (2) for the sample of objects we resample the objects with replacement to create a set of hypothetical possible objects we could have obtained for annotation. By doing such resampling sufficient number of times we simulate distribution of hypothetical possible annotations which can be measured against the original annotations.

2.2.3 Consistency

When human-annotated data is being used for monitoring changes in some metrics, consistency of annotations can be considered the most important quality aspect. Gaps in both accuracy and credibility can be accounted for using confidence intervals. Whereas changes in annotation consistency can result in misleading metrics and wrong interpretations of metrics dynamic. Thus for programs that rely on continuous human annotation, consistency plays a crucial role.

A proposed way to measure both Instant and Longitudinal Consistency is through a measure of agreement among reviewers. For the Instant Consistency such measure is estimated using individual reviewers' annotation collected at the time of initial review, and for the Longitudinal Consistency original annotations are compared with the re-annotations. In an ideal scenario, the original results are reproducible across time assuming underlying task remains and guidelines remain the same [2,4].

There are multiple statistics available for measure of agreement, such as kappa coefficients (Cohen kappa [9] and Fleiss kappa [6]), correlation coefficient, etc. In the ACC framework we are using the Fleiss kappa coefficient estimated on the sample of objects using individual reviewers decisions, as it allows more flexibility and does not rely on parametric assumption of the underlying distributions. [].

3 An Illustrative example of annotation quality implementation

We will be using the following illustrative example. Assume that in our sample there are 1000 objects. Each object is reviewed by 5 annotators, who are asked to make a decision about an object belonging to some class T. Possible decisions

Table 1: Summary of initial annotations

Yes verdicts		No verdicts	
Reviewers votes	Count	Reviewers votes	Count
5 - Yes, 0 - No	5	0 - Yes, 5 - No	880
4 - Yes, 1 - No	5	1 - Yes, 4 - No	60
3 - Yes, 2 - No	20	2 - Yes, 3 - No	30

Table 2: Summary of relabeling verdicts

Original overall verdict	Expert verdict	Type	Count
Yes	Yes	True positive	20
Yes	No	False positive	10
No	No	True negative	950
No	Yes	False negative	20

are: Yes, No. The verdict for each object is based on majority rule. Each object in the sample is also re-annotated by expert reviewers.

In this example, as in many real use cases, only a small fraction of objects belong to the class of interest. Summary of initial annotations is in 1. Summary of expert annotations in 2. Summary of ACC results are in table 3.

4 Implementation details

We are currently building an internal platform to perform regularly scheduled annotation quality checks on continuous annotation programs. Since these checks are part of a somewhat complex dependency graph, we use an orchestration platform to manage the relationships between our initial annotation tasks and our annotation quality checks. Once initial annotation tasks are complete, each batch of annotations is automatically ingested by our platform, parsed into a consistent structure, and loaded into a data warehouse. The downstream annotation quality checks run automatically on each batch of annotations once they are present in the data warehouse.

We have a separate pipeline for each type of quality check. One pipeline runs annotations through a function that calculates Fleiss Kappa; another executes

Table 3: ACC measurements results

Yes-prevalence	Accuracy		Consistency	Credibility
	Precision	*Recall*	*Fleiss Kappa*	95% *Credibility C.I. of Yes-prevalence*
3.0%	0.7	0.5	0.6	$[2.9\%, 4.4\%]$

resampling in order to calculate credibility of the results; a third sends samples of labels for relabeling by experts for accuracy review; and a fourth sends samples of objects for relabeling by the original reviewers pool after some time interval has passed (currently 60 days) for consistency checks. The results of all of these quality checks are published in a data warehouse and power dashboards showing quality metrics over time. Future plans include adding alerts if quality dips below certain thresholds.

5 Approaches for addressing quality issues

We suggest performing quality evaluations on a recurring basis. Regular monitoring will help to detect emerging issues early so that corrective measures can be taken. Metrics like Credibility and Instant Consistency do not require additional annotations, thus come at little cost. Accuracy and Longitudinal Consistency measurements are the most expensive because they require additional human labeling. It is possible that due to the budget constraints, these assessments can only be performed on a sub-sample of the data.

Annotation quality metrics can be impacted by a number of factors. While poor agreement can be indicative of bad actors or spam, recent research has explored other levers that can impact agreement, such as (1) differences in annotators' background knowledge or approaches to annotation and (2) ambiguity in the input data, task design, or guidelines[5,3].

In addition, changes over time could be a function of turnover of the annotation workforce, or individual annotator drift, wherein annotators apply labels more liberally or conservatively over time as they gain exposure to a task.

Depending on these reasons different strategies can be used to address quality issues. Among those are: (1) increasing number of reviewers, (2) improving task guidelines and refining task design, (3) improving reviewer training and qualification standards.

6 Conclusion

In this paper we have presented the ACC framework, which allows continuous quality evaluation for any labeling program and provides information about accuracy of annotations and their consistency over time. It also allows credibility of labeling to manifest itself through the confidence intervals of estimates based on human labeling. We believe that implementation of this framework dramatically increases usability and maturity of machine learning models, quantitative research, and measurement programs that are relying on human annotations.

References

[1] Daniel, Florian & Kucherbaev, Pavel & Cappiello, Cinzia & Benatallah, Boualem & Allahbakhsh, Mohammad. (2018) Quality control in crowdsourcing: A survey of

quality attributes, assessment techniques, and assurance actions, *ACM Computing Surveys (CSUR), 51-1*, pp. 1–40. ACM New York, NY, USA.

[2] Welty, Chris & Paritosh, Praveen & Aroyo, Lora (2019) Metrology for AI: From benchmarks to instruments. *arXiv preprint arXiv:1911.01875*

[3] Kairam, Sanjay & Heer, Jeffrey (2016) Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 1637–1648*

[4] Qarout, Rehab Kamal & Checco, Alessandro & Bontcheva, Kalina (2018) Investigating stability and reliability of crowdsourcing output. *CEUR Workshop Proceedings, vol. 2276, 83–87*

[5] Dumitrache, Anca (2015) Crowdsourcing disagreement for collecting semantic annotation. *European Semantic Web Conference, 701–710, Springer*

[6] Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76, 378-382*

[7] Efron,B. (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist. 7 (1) 1 - 26*

[8] Paritosh, P. (2012) Human computation must be reproducible. *In Proceedings of CrowdSearch: Crowdsourcing Web search 2012*

[9] Jacob Cohen (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, vol. 20, 37 –46*

[10] Eric Breck & Marty Zinkevich & Neoklis Polyzotis & Steven Whang & Sudip Roy (2019) Data Validation for Machine Learning. *Proceedings of SysML*