Association for the Advancement of Artificial Intelligence
HCOMP 2013

SQUARE: A Benchmark for Research on Computing Crowd Consensus

Aashish Sheshadri
Department of Computer Science
The University of Texas at Austin
aashishs@cs.utexas.edu

Matthew Lease
School of Information
The University of Texas at Austin
ml@ischool.utexas.edu

Abstract

While many statistical consensus methods now exist, relatively little comparative benchmarking and integration of techniques has made it increasingly difficult to determine the current state-of-the-art, to evaluate the relative benefit of new methods, to understand where specific problems merit greater attention, and to measure field progress over time. To make such comparative evaluation easier for everyone, we present SQUARE, an open source shared task framework including benchmark datasets, defined tasks, standard metrics, and reference implementations with empirical results for several popular methods. In addition to measuring performance on a variety of public, real crowd datasets, the benchmark also varies supervision and noise by manipulating training size and labeling error. We envision SQUARE as dynamic and continually evolving, with new datasets and reference implementations being added according to community needs and interest. We invite community contributions and participation.

1 Introduction

Nascent human computation and crowdsourcing (Quinn and Bederson 2011; Law and von Ahn 2011; Lease 2011) is transforming data collection practices in research and industry. In this paper, we consider the popular statistical aggregation task of *offline consensus*: given multiple noisy labels per example, how do we infer the best consensus label?

While many consensus methods have been proposed, relatively little comparative benchmarking and integration of techniques has occurred. A variety of explanations can be imagined. Some researchers may use consensus methods to improve data quality for another research task with little interest in studying consensus itself. A natural siloing effect of research communities may lead researchers to develop and share new consensus methods only within those communities they participate in. This would lessen awareness of techniques from other communities, especially when research is tightly-coupled with domain-specific tasks. For whatever reason, it has become increasingly difficult to determine the current state-of-the-art in consensus, to evaluate the relative benefit of new methods, and to demonstrate progress.

In addition, relatively few reference implementations or datasets have been shared. While many researchers in other communities simply want to know the best consensus method to use for a given task, lack of a clear answer and reference implementations has led to predominant use of simple majority voting as the most common method in practice. Is this reasonable, or do we expect more sophisticated methods would deliver significantly better performance?

In a recent talk on computational biology, David Tse (2012) suggested a field's progress is often driven not by new algorithms, but by well-defined challenge problems and metrics which drive innovation and enable comparative evaluation. To ease such comparative evaluation of statistical consensus methods, we present SQUARE (**S**tatistical **QU**ality **A**ssurance **R**obustness **E**valuation), a benchmarking framework with defined tasks, shared datasets, common metrics, and reference implementations with empirical results for a number of popular methods. Public shared implementations and/or datasets are used when available, and we provide reference implementations for other methods.

We focus here on evaluating consensus methods which do not require feature representations for examples. This requires consensus to be computed purely on the basis of worker behaviors and latent example properties, excluding hybrid solutions which couple automatic classification with human computation. In addition to measuring performance across datasets of varying scale and properties, SQUARE varies degree of supervision, and we realistically simulate varying noise by preserving empirical traits of each dataset. Beyond empirical analysis, examining multiple techniques in parallel further helps us to organize and compare methods qualitatively, characterizing distinguishing traits, new variants, and potential integration opportunities. We envision SQUARE[1] as a dynamic and evolving community resource, with new datasets and reference implementations added based on community needs and interest.

2 Datasets

We begin by identifying and describing a number of public datasets that are online and provide the foundation for SQUARE 1.0. An early design decision was to include only datasets containing real crowd judgments, thereby increasing validity of experimental findings. While synthetic data can also be useful for sanity checks, carefully controlled experiments, and benchmarking, relatively little synthetic data has been shared. This likely stems from its lesser perceived value and a belief that it can be easily re-generated by others (provided that the generation process is fully and

[1] ir.ischool.utexas.edu/square

Sheshadri, A., & Lease, M. (2013). SQUARE: A benchmark for research on computing crowd consensus. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 1, 156–164.