Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

LabelBoost: An Ensemble Model for Ground Truth Inference Using Boosted Trees

Siamak Faridani, Georg Buscher

Microsoft
One Microsoft Way
Redmond, WA

Abstract

We introduce *LabelBoost*, an ensemble model that utilizes various label aggregation algorithms to build a higher precision algorithm. We compare this algorithm with majority vote, GLAD and an Expectation Maximization model on a publicly available dataset. The results suggest that by building an ensemble model, one can achieve higher precision value for aggregating crowd-sourced labels for an item. These higher values are shown to be statistically significant.

Introduction

Inferring the ground truth label for an item from a set of crowd-generated labels is an active research area that has direct impact on the practice of crowdsourcing. In this work, we provide early experimental results of a boosted tree model that combines the results of 3 simpler models to achieve a more accurate ensemble. We compare the ensemble model and the three alternatives on the TREC dataset (Tang and Lease 2011). The evaluation is done on a crowd-sourced binary classification experiment; however, the model is extendable to multi-class classification and ground truth inference when the labels are ordinal.

Problem Statement

We assume *m* items are classified into *2* different binary categories [0,1] by online judges. The number of judgments per item can vary from one item to the next. The true class of the item is not known a priori and the goal of the model is to infer this latent category by observing the labels that the judges have provided on the item. Most of the recent models (Smyth et al. 1995) also utilize the judgments on other items to build a reputation model for each judge. This reputation model can then enhance the ground truth inference on new items.

Dataset

For our evaluation we use the TREC dataset (Tang and Lease 2011). In the set, 762 judges classified 19,033 pairs of topic-document examples into relevant and irrelevant categories. The full set contains 89,624 binary judgments. We took a subset of the dataset with 3,122 topic-documents for which the publishers had provided gold labels (relevance judgments that are performed by experts and can be considered as ground truth). This subset contained 13,750 judgments. Availability of expert labels allowed us to test our model on this set. The task is binary classification and we are in the process of putting together larger test sets from various types of tasks for further evaluation.

Proposed Solution

We propose an ensemble of many simpler ground truth inference models to serve as a higher precision model. For this work we took 4 models that are frequently used by practitioners and researchers:

- Majority Vote
- GLAD (Whitehill et al. 2009)
- Expectation-Maximization (a variation of Smyth et al. 1995)
- Minimax Entropy Model (Zhou et al. 2012)

We later removed the Minimax entropy from the ensemble since tuning two parameters for the task was challenging. We ran each one of the three models separately on the TREC dataset (Tang and Lease 2011) to classify each item.

We also hypothesize that there are various characteristics of the item that can help the ensemble model arrive at the better final label for an item. For example, the mean, the variance and the number of judgments for the items can improve the output of the ensemble. The final feature set for each item was as following:

- Output from Expectation-Maximization Model
- Output of GLAD
- Median

- Mean
- Mode
- Variance of Judgments around the mean
- Number of judgments

Note that for a binary classification, median and mode are equal, but they differ for multi-class classification.

Machine learning using a Boosted Tree

We use a boosted tree to learn the ground truth label from the output of the other classifiers and other item features like the variance of judgments and the number of judgments. For this paper we used the publically available GBM package in R. Our model used 2,000 trees, with 4 interaction layers and we used 0.005 as the shrinkage parameter.

Evaluation

We evaluated the ensemble model on 3,122 documents from TREC that had an expert-provided label. We performed 30 rounds of experiments. In each round, we randomly selected 10% of the 3,122 gold labels as evaluation set and trained the boosted tree on the rest of the set. At the end we tested the prediction of the boosted tree as well as the EM, GLAD and the majority vote on the evaluation set. Precision was used as the evaluation metric. The precision values for each model were then averaged over 30 different runs. Table 1 summarizes these results. Figure one shows the precision of LabelBoost (the black line) compared to other models.

	EM	GLAD	Mode	Label-Boost
Average Precision	64.18%	53.21%	63.78%	***65.64%***
Welch's test against LabelBoost	p< 2.2e-16	p<2.2e-16	p< 2.2e-16	---

Table 1 Average Precision of Various Models over 30 runs

We performed a Welch's t-test on the results of 30 iterations to make sure the increase in precision is statistically significant. Table 1 also summarizes the p-value for these comparisons. The small p-values suggest that the differences in precision between LabelBoost and other models are statistically significant.

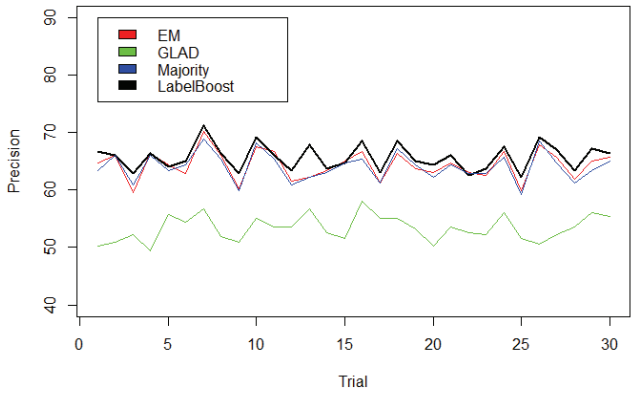

Figure 1 Precision value for various models over 30 runs of experiment

Conclusion and Future Work

We built an ensemble model from a set of simpler label inference models. The ensemble showed a significant improvement in precision over the rest of the models when it was compared on the TREC dataset. We would like to extend this work and test it on a larger set of tasks. The results of this work would be immediately usable by researchers and practitioners who are building crowdsourcing systems.

References

Tang, W., & Lease, M. (2011). *Semi-supervised consensus labeling for crowdsourcing*. In SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR).

Smyth, Padhraic, et al. "*Inferring ground truth from subjective labelling of venus images.*" Advances in neural information processing systems (1995): 1085-1092.

Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). *Whose vote should count more: Optimal integration of labels from labelers of unknown expertise*. In Advances in neural information processing systems (pp. 2035-2043).

Zhou, D., Platt, J., Basu, S., & Mao, Y. (2012). *Learning from the wisdom of crowds by minimax entropy*. In Advances in Neural Information Processing Systems 25 (pp. 2204-2212).