Approximating the Wisdom of the Crowd

Seyda Ertekin MIT

Haym Hirsh Rutgers University

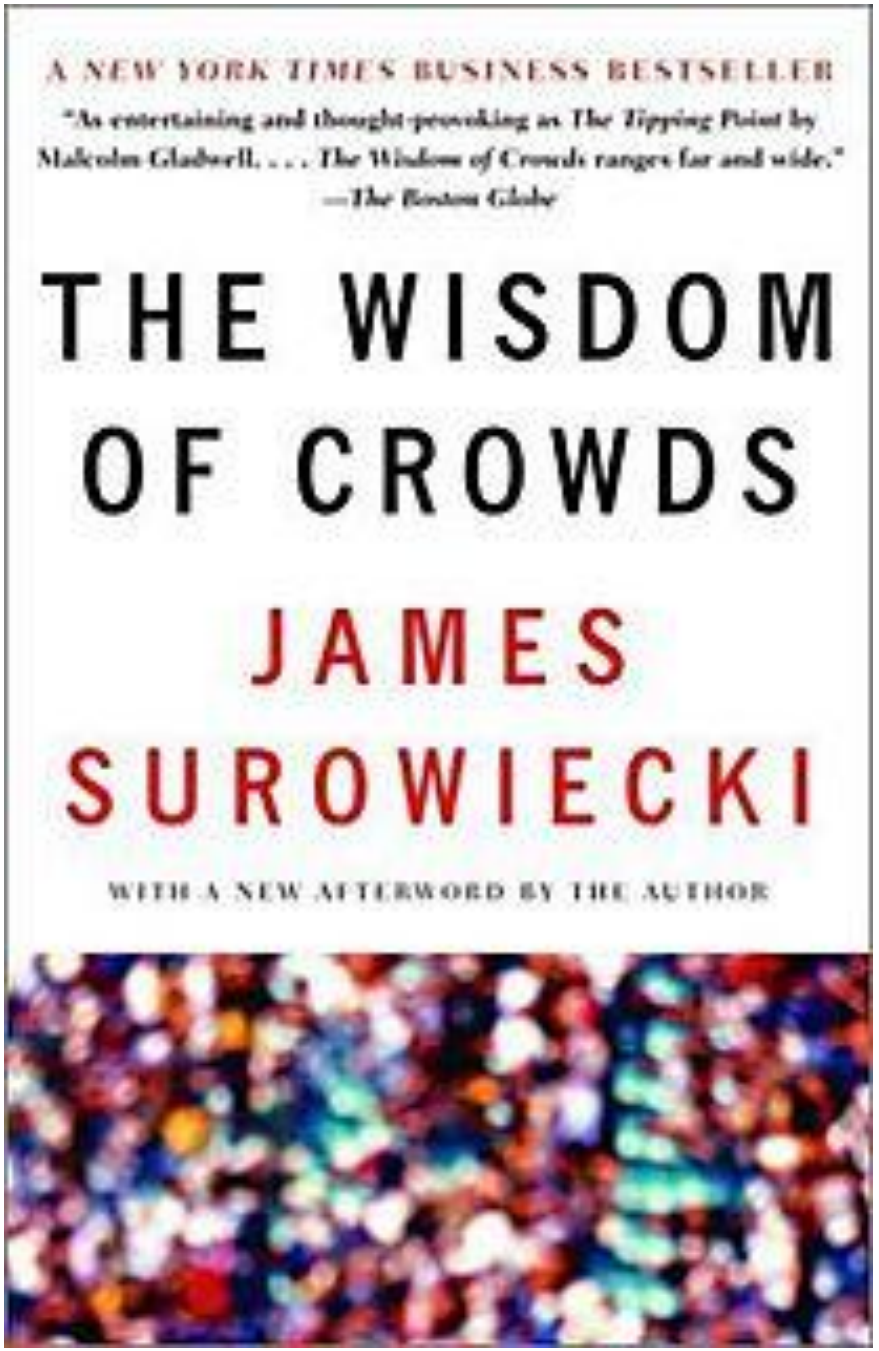
Cynthia Rudin MIT

Approximating the Wisdom of the Crowd

Seyda Ertekin MIT

Haym Hirsh Rutgers University

Cynthia Rudin MIT



© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

17°·0 at Moyeni, Basutoland, on August 23. The mean yearly value of the absolute maxima was 86°·9, and of the corresponding minima 41°·6. The mean temperature for the year was 0°·9 below the average. The stormiest month was October, and the calmest was April.

We have also received the official meteorological year-books for South Australia (1904) and Mysore (1905). Both of these works contain valuable means for previous years.

Forty Years of Southern New Mexico Climate.—Bulletin No. 59 of the New Mexico College of Agriculture contains the meteorological data recorded at the experimental station from 1892 to 1905 inclusive, together with results of temperature and rainfall observations at other stations in the Mesilla Valley for most of the years between 1851 and 1890, published some years ago by General Greely in a "Report on the Climate of New Mexico." The station is situated in lat. 32° 15′ N., long. 106° 45′ W., and is 3868 feet above sea-level. The data have a general application to those portions of southern New Mexico with an altitude less than 4000 feet. The mean annual temperature for the whole period was 61°·6, mean maximum (fourteen years) 76°·8, mean minimum 41°·4, absolute maximum 106° (which occurred several times), absolute minimum 1° (December, 1895). The mean annual rainfall was 8·8 inches; the smallest yearly amount was 3·5 inches, in 1873, the largest 17·1 inches, in 1905. Most of the rain falls during July, August, and September. The relative humidity is low, the mean annual amount being about 51 per cent. The bulletin was prepared by J. D. Tinsley, vice-director of the station.

Meteorological Observations in Germany.—The results of the observations made under the system of the Deutsche Seewarte, Hamburg, for 1905, at ten stations of the second order, and at fifty-six storm-warning stations, have been received. This is the twenty-eighth yearly volume published by the Seewarte, and forms part of the series of German meteorological year-books. We have frequently referred to this excellent series, and the volume in question is similar in all respects to its predecessors; it contains most valuable data relating to the North Sea and Baltic coasts. We note that the sunshine at Hamburg was only 29 per cent. of the possible annual amount, and that there were 103 sunless days; the rainfall was 25·9 inches, the rainy days being 172 in number.

VOX POPULI.

IN these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. The material about to be discussed refers to a small matter, but is much to the point.

A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6d. each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered and "dressed." Those who guessed most successfully received prizes. About 800 tickets were issued, which were kindly lent me for examination after they had fulfilled their immediate purpose. These afforded excellent material. The judgments were unbiassed by passion and uninfluenced by oratory and the like. The sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best. The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies. The average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case.

After weeding thirteen cards out of the collection, as being defective or illegible, there remained 787 for discussion. I arrayed them in order of the magnitudes of the estimates, and converted the *cwt.*, *quarters*, and *lbs.* in which they were made, into lbs., under which form they will be treated.

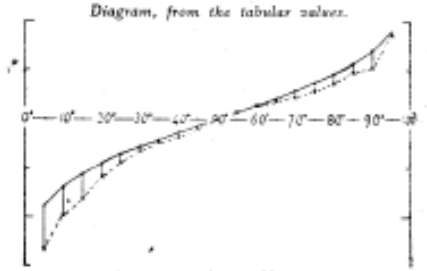
Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array e^{ν/100}	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e. =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
M 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	− 10
85	1254	+ 47	+57	− 10
90	1267	+ 52	+70	− 18
95	1293	+ 86	+90	− 4

q1, q3, the first and third quartiles, stand at 25′ and 75′ respectively.
M, the median or middlemost value, stands at 50′.
The dressed weight proved to be 1198 lbs.

According to the democratic principle of "one vote one value," the middlemost estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters (for fuller explanation see "One Vote, One Value," NATURE, February 28, p. 414). Now the middlemost estimate is 1207 lb., and the weight of the dressed ox proved to be 1198 lb.; so the *vox populi* was in this case 9 lb., or 0·8 per cent. of the whole weight too high. The distribution of the estimates about their middlemost value was of the usual type, so far that they clustered closely in its neighbourhood and became rapidly more sparse as the distance from it increased.

Diagram, from the tabular values.



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

But they were not scattered symmetrically. One quarter of them deviated more than 45 lb. above the middlemost (3·7 per cent.), and another quarter deviated more than 29 lb. below it (2·4 per cent.), therefore the range of the two middle quarters, that is, of the middlemost half, lay within those limits. It would be an equal chance that the estimate written on any card picked at random out of the collection lay within or without those limits. In other words, the "probable error" of a single observation may be reckoned as ½(45+29), or 37 lb. (3·1 per cent.). Taking this for the p.e. of the normal curve that is best adapted for comparison with the observed values, the results are obtained which appear in above table, and graphically in the diagram.

CSS Workshop 17 Dec 2011
© Seyda Ertekin, Haym Hirsh, and Cynthia Rudin

The Wisdom of Crowds: "the aggregation of information in groups, resulting in decisions that […] are often better than could have been made by any single member of the group."

Wikipedia

15 Dec 2011

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

How efficiently (and accurately) can you approximate the crowd?

How efficiently (and accurately) can you approximate the crowd?

Problem Setting

- You have a roomful of people
- Each can give answers to yes/no questions that you pose
- Each time you ask anyone for an answer, it costs you
- The “correct” answer is the majority vote of the room

© Şeyda Ertekin, Haym Hirsh,
and Cynthia Rudin

Problem Addressed

- (How) Can you guess the majority vote of the crowd without asking everyone for their answers?

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Problem Addressed

- (How) Can you guess the majority vote of the crowd without asking everyone for their answers?
- (How) Can you do this “on line,” learning to approximate the crowd during the act of approximating the crowd?

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

What This Is Not

- Polling
  - No demographic information to generalize from

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

What This Is Not

- Polling
  - No demographic information to generalize from
- Estimating “ground truth”
  - “Truth” is crowd-specific

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Key Ideas

1. Associate a weight with each labeler based on performance on past items
   - Weight = labeler accuracy
   - Do the right Bayesian smoothing on these weights

$$Q_{it} = \frac{a_{it} + K}{c_{it} + 2K}$$

Q_{it}: The weight of labeler i after seeing item t

c_{it}: How many times we asked i about items

a_{it}: How many times i was right

K, $2K$: Beta-binomial distribution with $\alpha = K$ and $\beta = 2K$

Key Ideas

2. Mix exploration and exploitation
   - Exploitation: Select the labelers for each item based on the weights
   - Exploration: Select a random labeler for each item

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Key Ideas

3. Build up the set of labelers dynamically for each item
   - Start with 3 labelers
     - Exploitation: Pick 2 based on weights
     - Exploration: Pick 1 uniformly at random
   - Get their answers
   - Keep adding labelers and getting their answers until you're confident with the prediction

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Key Ideas

3. Build up the set of labelers dynamically for each item
   - Start with 3 labelers
     - Exploitation: Pick 2 based on weights
     - Exploration: Pick 1 uniformly at random
   - Get their answers
   - Keep adding labelers and getting their answers until you're confident with the prediction

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Key Ideas

4. "until you're confident with the prediction":

   If the next best labeler has enough weight to change the vote (or come close), add it in

$$\frac{|\text{Score}(S_t)| - Q_{l_{\text{candidate}},t}}{|S_t| + 1} < \varepsilon$$

   Low ε: Exploitation

   High ε: Exploration

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

The CrowdSense Algorithm

1. **Input:** Examples $\{x_1, x_2, \ldots, x_N\}$, Labelers $\{l_1, l_2, \ldots, l_M\}$, confidence threshold ε, smoothing parameter K.
2. **Define:** $L_Q = \{l^{(1)}, \ldots, l^{(M)}\}$, labeler id's in descending order of their quality estimates.
3. **Initialize:** $a_{i1} \leftarrow 0$, $c_{i1} \leftarrow 0$ for $i = 1, \ldots, M$.
4. **Loop for** $t = 1, ..., N$
   (a) Compute quality estimates $Q_{it} = \frac{a_{it}+K}{c_{it}+2K}$, $i = 1, \ldots, M$. Update L_Q.
   (b) $S_t = \{l^{(1)}, l^{(2)}, k\}$, where k is randomly sampled from the set $\{l^{(3)}, \ldots l^{(M)}\}$.
   (c) **Loop for** $j = 3 \ldots M$, $j \neq k$
      i. $\text{Score}(S_t) = \sum_{i \in S_t} V_{it} Q_{it}$, $l_{\text{candidate}} = l^{(j)}$.
      ii. If $\frac{|\text{Score}(S_t)| - Q_{l_{\text{candidate}},t}}{|S_t|+1} < \varepsilon$, then $S_t \leftarrow S_t \cup l_{\text{candidate}}$. Otherwise exit loop to stop adding new labelers to S_t.
   (d) Get the weighted majority vote of the labelers $V_{S_t t} = \text{sign}\left(\sum_{i \in S_t} V_{it} Q_{it}\right)$
   (e) $\forall i \in S_t$ where $V_{it} = V_{S_t t}$, $a_{it} \leftarrow a_{it} + 1$
   (f) $\forall i \in S_t$, $c_{it} \leftarrow c_{it} + 1$
5. **End**

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Evaluation

Dataset	# of Labelers	Type of Labeler
MovieLens	11	Human
ChemIR	11	Software
Reuters	13	Learned classifiers

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

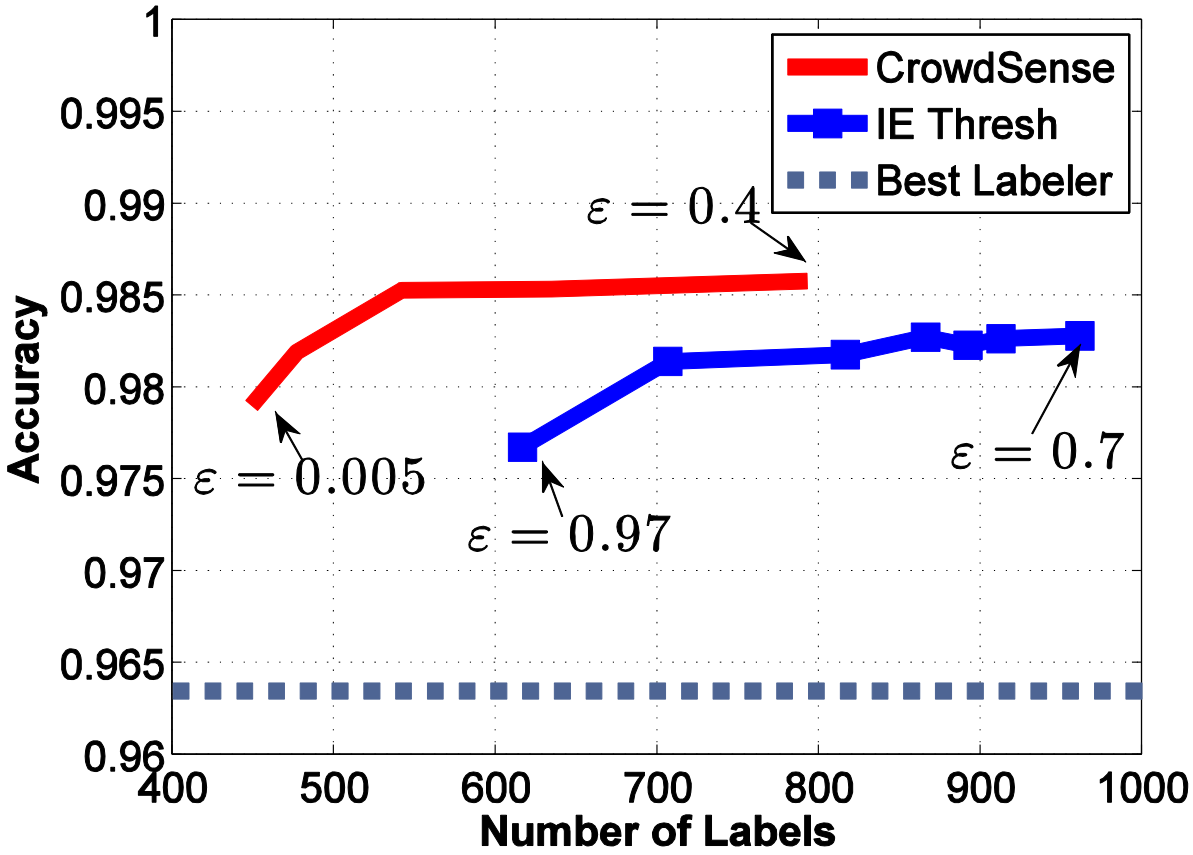
Baseline Methods

1. The accuracy of the overall best labeler (in hindsight)
2. Mean accuracy of the labelers
3. The accuracy of unweighted random labelers
4. IEThresh:

   Order labelers using the upper confidence interval for the probability that a labeler will agree with the majority vote

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

MovieLens

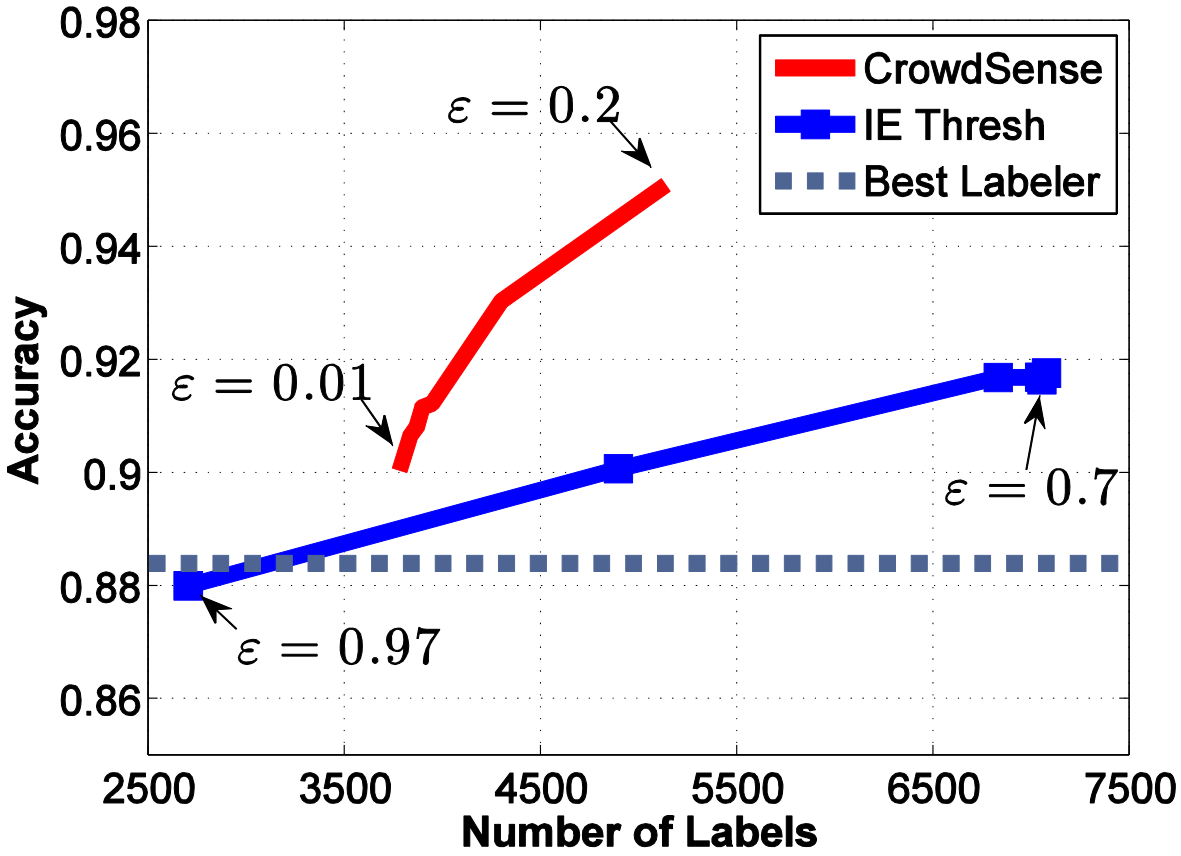


Baseline a: 74%

Baseline c: 34%

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

ChemIR

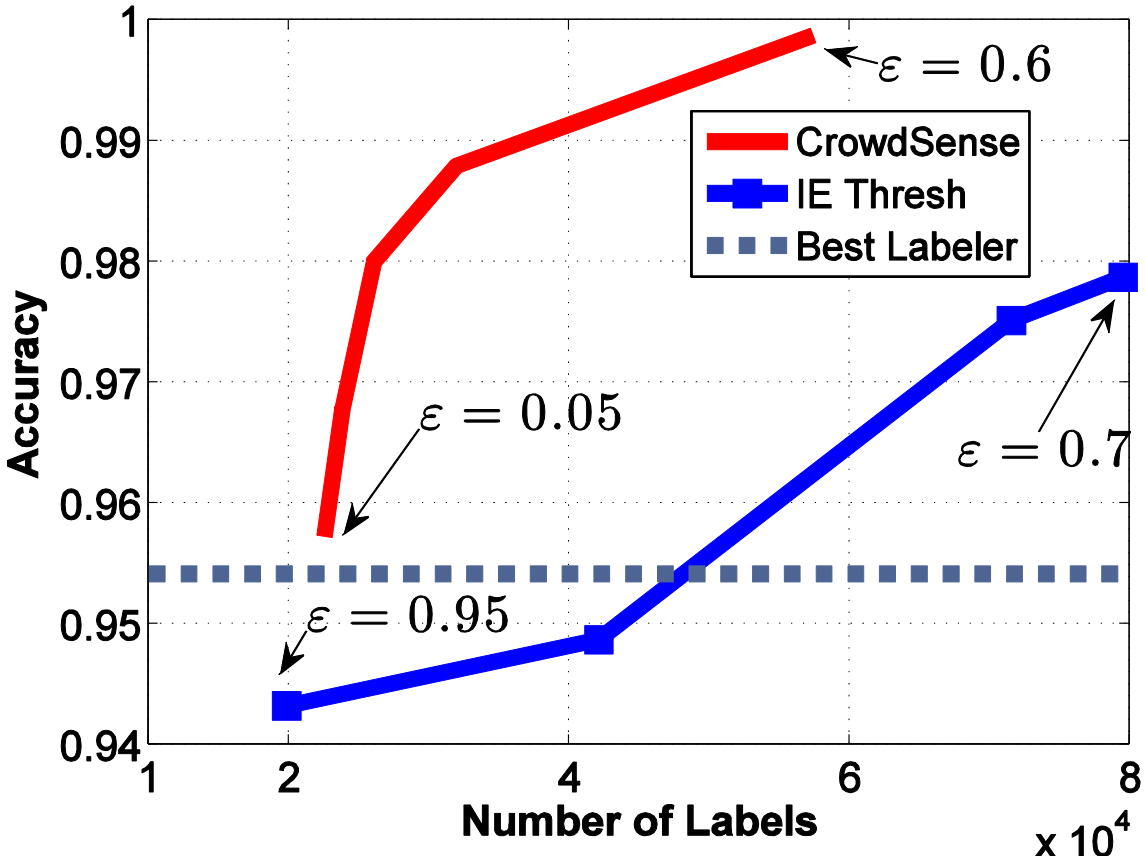


Baseline a: 69%

Baseline c: 73%

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Reuters

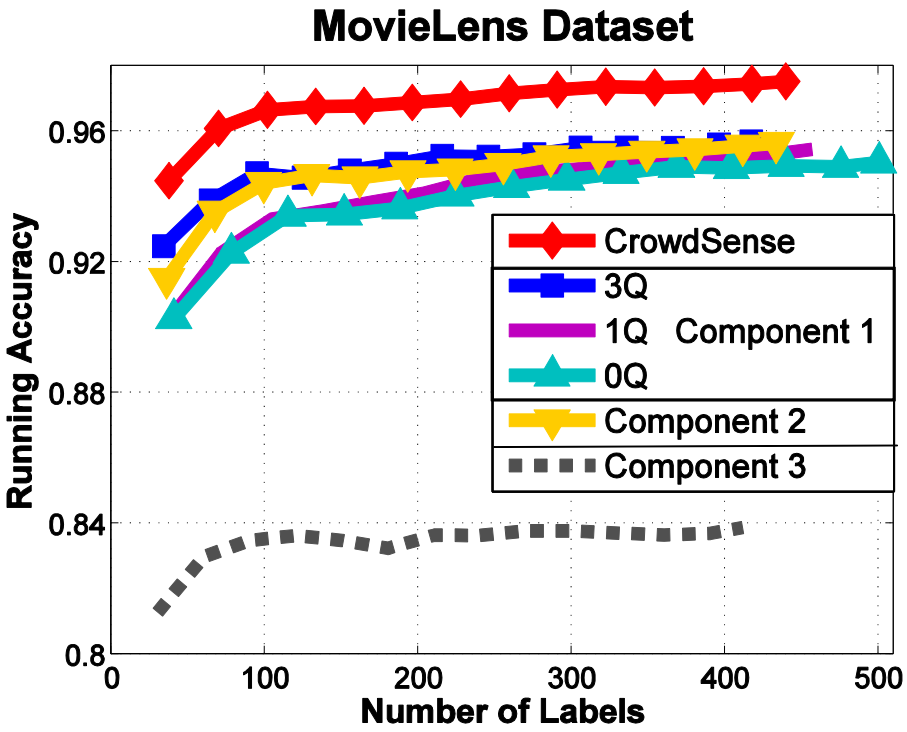


Baseline a: 85%

Baseline c: 95%

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Analysis of CrowdSense Design



Components of CrowdSense:

1. Composition of the initial seed set of labelers
2. How subsequent labelers are added to the set
3. The weighting scheme that affects 1, 2, and combining the votes of the individual labelers.

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Future Work

- Beyond classification
- Greater number of labelers (CrowdSense 2)
- Item features
- Labeler features
- Still early, other algorithms possible
  - Active learning
  - Sleeping experts
  - Budget-sensitive learning

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin

Summary

- Introduced the problem of approximating the wisdom of crowds
- Developed an algorithm for approximating the wisdom of the crowd
  - Balance exploration and exploitation
  - Select labelers based on past accuracy (with appropriate smoothing)
  - Incrementally accrues only enough labelers to reach some confidence in prediction

© Şeyda Ertekin, Haym Hirsh, and Cynthia Rudin