Fast Dawid-Skene: A Fast Vote Aggregation Scheme for Sentiment Classification

Vaibhav B Sinha, Sukrut Rao, Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad

KDD WISDOM '18

https://sites.google.com/view/fast-dawid-skene/

https://github.com/GoodDeeds/Fast-Dawid-Skene/

Introduction

Applications of supervised machine learning

- Image classification
- Sentiment/Opinion classification from text/media
- Object detection
- etc.

Key: lots of labeled data

Introduction

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Getting labeled data for classification tasks
  - Expensive
  - Time-consuming
  - May require specialized domain knowledge (eg. Medicine)

Introduction

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Getting labeled data for classification tasks
  - Expensive
  - Time-consuming
  - May require specialized domain knowledge (eg. Medicine)

- Possible solution: crowdsource labels
  - Obtain labels for each data point from a group of non-experts
  - Apply aggregation algorithm to estimate true label

Introduction

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Getting labeled data for classification tasks
  - Expensive
  - Time-consuming
  - May require specialized domain knowledge (eg. Medicine)
- Possible solution: crowdsource labels
  - Obtain labels for each data point from a group of non-experts
  - Apply aggregation algorithm to estimate true label
- Simple aggregation algorithm: Majority Voting
  - Estimate label chosen by majority of aggregators

Introduction

- Crowdsourced aggregation
  - Not all annotators are equally reliable
  - Some data points are difficult to label

Majority Voting does not take these characteristics into account

Can we do better?

Existing Techniques: Dawid-Skene

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Dawid-Skene algorithm [Dawid and Skene, 1979]
  - EM algorithm
  - Efficient and widely used till date
- Dawid-Skene takes time to converge – increases with increasing dataset sizes
- Fast, real-time sentiment analysis required
- Proposals
  - Iterated Weighted Majority Voting (IWMV) [Li and Yu, 2014]
  - Fast Dawid-Skene (FDS) (ours)

Problem Setting

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Each data-point (question) has exactly one true label (option), from a fixed set of choices.
- Participants (annotators) provide labels for questions.
- Each participant chooses one option per question.
- A participant may answer only a subset of questions.
- Each question is presented to multiple participants.
- **Task:** Aggregate the label chosen by the participants for each question to estimate the true label.

Fast Dawid-Skene

	Q1 (b)	**Q2 (a)**	**Q3 (b)**	**Q4 (a)**	**Q5 (a)**
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b

Fast Dawid-Skene: Majority Voting

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	a				

Fast Dawid-Skene: Majority Voting

	Q1	**Q2**	**Q3**	**Q4**	**Q5**
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

First E step: Majority Voting

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

First M step

The fraction:

$$\frac{\text{Number of questions answered by P1 whose correct answer was a and (s)he chose a}}{\text{Number of questions answered by P1 whose answer was a}}$$

$= 3 / 3 = 1$

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

First M step

The fraction:

$$\frac{\text{Number of questions answered by P1 whose correct answer was a and (s)he chose a}}{\text{Number of questions answered by P1 whose answer was a}}$$

$= 3 / 3 = 1$

	a	**b**
a	1	
b		

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

First M step

Similarly complete the table for P1

	a	**b**
a	1	0
b	0.5	0.5

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

First M step

Similarly complete the table for P2 and P3

P1	**a**	**b**
a	1	0
b	0.5	0.5

P2	**a**	**b**
a	0.67	0.33
b	0	1

P3	**a**	**b**
a	0.33	0.67
b	0.5	0.5

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

P1	**a**	**b**
a	1	0
b	0.5	0.5

P2	**a**	**b**
a	0.67	0.33
b	0	1

P3	**a**	**b**
a	0.33	0.67
b	0.5	0.5

First M step

Also calculate the probabilities of each option being correct (priors)

a	0.6
b	0.4

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

P1	**a**	**b**
a	1	0
b	0.5	0.5

P2	**a**	**b**
a	0.67	0.33
b	0	1

P3	**a**	**b**
a	0.33	0.67
b	0.5	0.5

a	0.6
b	0.4

Second E step

Now we reestimate the answers for each questions.

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

P1	**a**	**b**
a	1	0
b	0.5	0.5

P2	**a**	**b**
a	0.67	0.33
b	0	1

P3	**a**	**b**
a	0.33	0.67
b	0.5	0.5

a	0.6
b	0.4

Second E step

Now we reestimate the answers for each questions.

Probability that answer to first question is a:

(Prior) 0.6 x 1 x 0.33 x 0.33

= 0.067

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

P1	a	b
a	1	0
b	0.5	0.5

P2	a	b
a	0.67	0.33
b	0	1

P3	a	b
a	0.33	0.67
b	0.5	0.5

a	0.6
b	0.4

Second E step

Now we reestimate the answers for each questions.

Probability that answer to first question is a:

(Prior) 0.6 x 1 x 0.33 x 0.33
= 0.067

Similarly probability that answer to first question is b:

(Prior) 0.4 x 0.5 x 1 x 0.5 = 0.1

Fast Dawid-Skene

	Q1	Q2	Q3	Q4	Q5
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV	**a**	**b**	**b**	**a**	**a**

P1	**a**	**b**
a	1	0
b	0.5	0.5

P2	**a**	**b**
a	0.67	0.33
b	0	1

P3	**a**	**b**
a	0.33	0.67
b	0.5	0.5

a	0.6
b	0.4

Second E step

Now we reestimate the answers for each questions.

Probability that answer to first question is a:

(Prior) 0.6 x 1 x 0.33 x 0.33
= 0.067

Similarly probability that answer to first question is b:

(Prior) 0.4 x 0.5 x 1 x 0.5 = 0.1

Thus (b) becomes the answer

Fast Dawid Skene

	Q1 (b)	**Q2 (a)**	**Q3 (b)**	**Q4 (a)**	**Q5 (a)**
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV (First E step)	a	b	b	a	a
After 2 E steps	**b**	**a**	**b**	**a**	**a**

Fast Dawid Skene

	Q1 (b)	**Q2 (a)**	**Q3 (b)**	**Q4 (a)**	**Q5 (a)**
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV (First E step)	a	b	b	a	a
After 2 E steps	b	a	b	a	a
After 3 steps	**b**	**a**	**b**	**a**	**a**

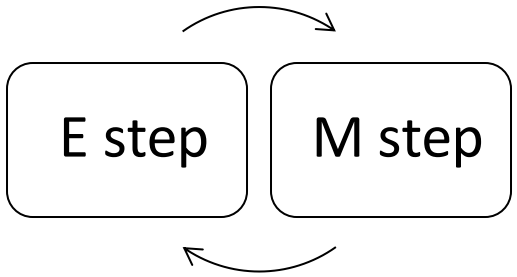
Fast Dawid Skene

	Q1 (b)	**Q2 (a)**	**Q3 (b)**	**Q4 (a)**	**Q5 (a)**
P1	a	a	b	a	a
P2	b	b	b	a	a
P3	a	b	a	b	b
MV (First E step)	a	b	b	a	a
After 2 E steps	b	a	b	a	a
After 3 steps	**b**	**a**	**b**	**a**	**a**

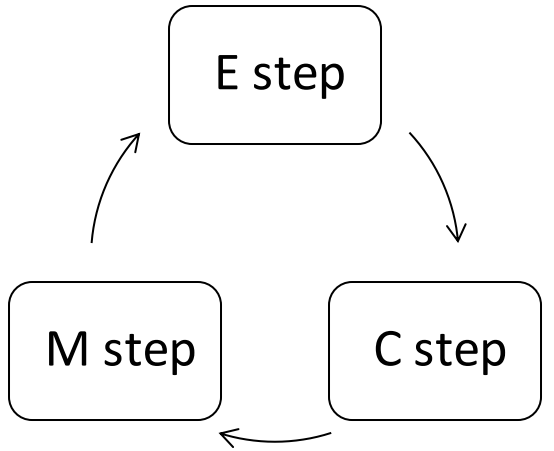
The algorithm converges.

Fast Dawid Skene Algorithm

- E step: Estimate the answers to the questions
- C step: Give the 'hard' estimates.
- M step: Compute the parameters.



Dawid-Skene Algorithm



Fast Dawid-Skene Algorithm

Guarantees for Convergence

Theorem 1:

Fast Dawid-Skene converges to a stationary point.

Guarantees for Convergence

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Theorem 1:

Fast Dawid-Skene converges to a stationary point.

Theorem 2:

If the algorithm is started from an area close to a local maximum of the likelihood, Fast Dawid-Skene is guaranteed to converge to the maximum at a linear rate.

More details in our paper

Improvement: Hybrid Algorithm

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- FDS: Empirical Observations
  - Likelihood is not maximized to the same extent as DS
  - DS converges to a better maxima
- Proposal: Hybrid algorithm
  - Start with DS
  - Switch to FDS after difference in priors is below a certain threshold
  - Best of both worlds – procedure of DS, speed of FDS

Extensions to FDS

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

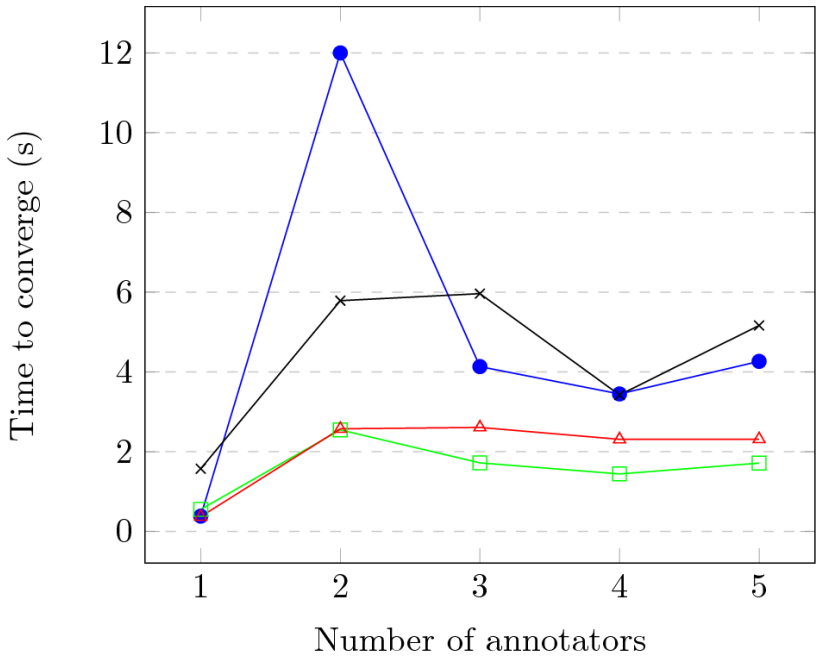
- Online FDS
  - Online setting: Initial set of questions and annotations available, new questions with annotations become available with time
  - Perform aggregation as questions arrive, using information from past data
- Multiple Answers Correct
  - Assumption: truth value of each option is independent
  - Treat each question-option pair as a separate binary question
  - Run FDS/Hybrid algorithm on each question-option pair

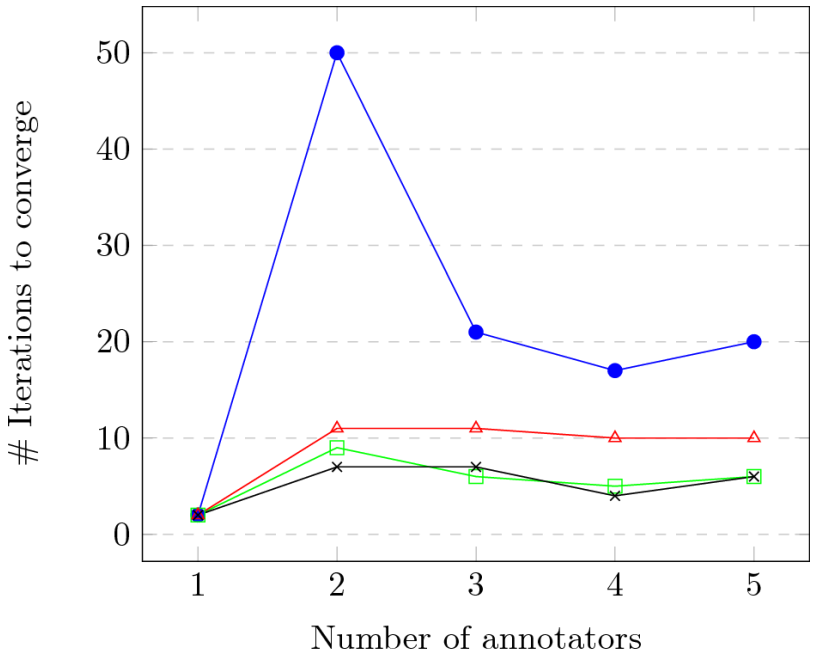
Experiments and Results

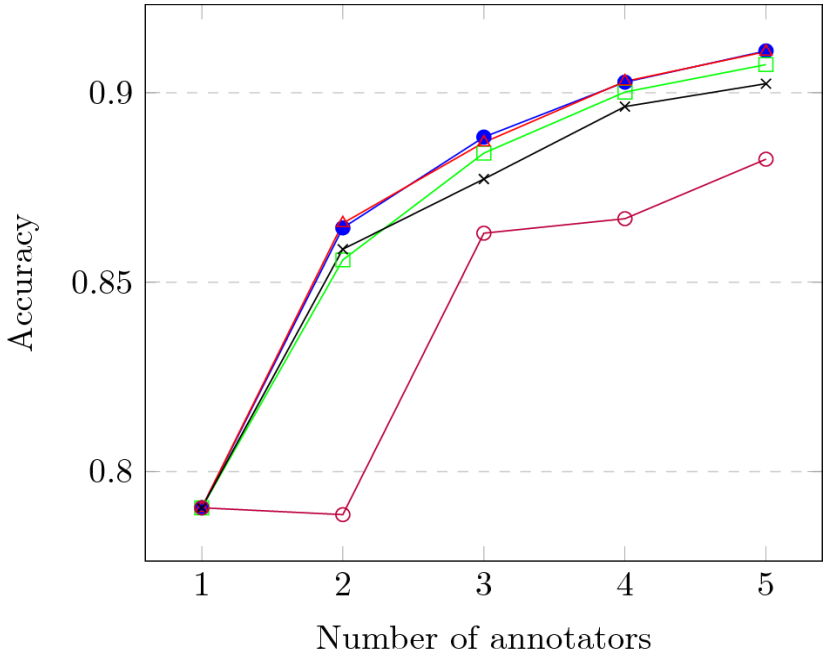
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Experiments: Comparison of DS, FDS, Hybrid, MV, IWMV, and GLAD [Whitehill et al., 2009] across seven real-world datasets
- Results
  - 3.00x - 7.84x speed of FDS compared to DS
  - 1.49x - 5.15x speed of Hybrid compared to DS
  - 0.54x - 6.09x speed of FDS compared IWMV

Results: Sentiment Polarity Dataset







DS IWMV MV FDS Hybrid

Questions: 4968, Options per question: 2, Maximum number of annotators per question: 5

Thank you

Questions?