CROWDSOURCING

Stephen Mayhew, Daniel Khashabi

May 5, 2015

University of Illinois at Urbana Champaign

BASELINES

1. Majority Voting
2. Hubs and Authorities
3. Singular vector approach
4. EM
5. EM with priors
6. Iterative Weighted Majority Voting
7. Simplified BP
8. Discretized BP
9. Tree-reweighted message passing[1]
10. Spectral Meets EM[1]

[1] never finished...

BASELINES

1. Majority Voting
2. Hubs and Authorities
3. Singular vector approach
4. EM
5. **EM with priors**
6. **Iterative Weighted Majority Voting**
7. Simplified BP
8. **Discretized BP**
9. Tree-reweighted message passing[1]
10. Spectral Meets EM[1]

[1] never finished...

Define

$$g_{ij}(t_i, p_j) = p_j \mathbf{I}\left\{A_{ij} = t_i\right\} + (1 - p_j)\mathbf{I}\left\{A_{ij} \neq t_i\right\}$$

EM WITH PRIORS

Define

$$g_{ij}(t_i, p_j) = p_j \mathbf{I}\left\{A_{ij} = t_i\right\} + (1 - p_j)\mathbf{I}\left\{A_{ij} \neq t_i\right\}$$

Full likelihood (with Beta priors):

$$\mathcal{L}(p; t) = \prod_{i,j} g_{ij}(t_i, p_j) \prod_{j} \left(c + \frac{1 - c}{B(\alpha, \beta)} p_j^{\alpha - 1} (1 - p_j)^{\beta - 1} \right)$$

EM WITH PRIORS

Define

$$g_{ij}(t_i, p_j) = p_j \mathbf{I}\left\{A_{ij} = t_i\right\} + (1 - p_j)\mathbf{I}\left\{A_{ij} \neq t_i\right\}$$

Full likelihood (with Beta priors):

$$\mathcal{L}(p; t) = \prod_{i,j} g_{ij}(t_i, p_j) \prod_j \left(c + \frac{1 - c}{B(\alpha, \beta)} p_j^{\alpha - 1} (1 - p_j)^{\beta - 1} \right)$$

M-Step becomes:

$$p_j = \frac{\alpha - 1 + \sum_j q(A_{ij})}{\alpha + \beta - 2 + \sum_j q(A_{ij}) + \sum_j q(-A_{ij})}$$

(Note: this same result is a lower bound on the situation where priors come from $p_j = 0.1 + 0.9Z$)

ITERATIVE WEIGHTED MAJORITY VOTING (IWMV)

From *Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing* by Hongwei Li and Bin Yu[2]

Algorithm 1 The iterative weighted majority voting algorithm (IWMV)

Input: Number of workers= M; Number of items= N; data matrix: $Z \in \overline{[L]}^{M \times N}$;
Output: the predicted labels $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_N\}$
Initialization: $\nu_i = 1, \quad \forall i \in [M]; \quad T_{ij} = \mathrm{I}\left(Z_{ij} \neq 0\right), \forall i \in [M], \forall j \in [N]$.
repeat

$$\hat{y}_j \leftarrow \underset{k \in [L]}{\operatorname{argmax}} \sum_{i=1}^{M} \nu_i \mathrm{I}\left(Z_{ij} = k\right), \qquad \forall j \in [N].$$

$$\hat{w}_i \leftarrow \frac{\sum_{j=1}^{N} \mathrm{I}\left(Z_{ij} = \hat{y}_j\right)}{\sum_{j=1}^{N} T_{ij}}, \qquad \forall i \in [M].$$

$$\nu_i \leftarrow L\hat{w}_i - 1, \qquad \forall i \in [M].$$

until converges or reaches S iterations.
Output the predictions $\{\hat{y}_j\}_{j \in [N]}$ by $\hat{y}_j = \operatorname{argmax}_{k \in [L]} \sum_{i=1}^{M} \nu_i \mathrm{I}\left(Z_{ij} = k\right)$.

[2] http://arxiv.org/pdf/1411.4086v1.pdf

DISCRETIZED BP

Original BP udpates are:

$$y^{(k)}_{a\to i}(p_a) \propto \mathcal{F}(p_a) \prod_{j\in\partial a\setminus i} \Big\{(p_a + \bar{p}_a + (p_a - \bar{p}_a)A_{ja})x^{(k)}_{j\to a}(+1) + (p_a + \bar{p}_a - (p_a$$

$$x^{(k+1)}_{i\to a}(\hat{t}_i) \propto \prod_{b\in\partial i\setminus a} \int \Big(y^{(k)}_{b\to i}(p_b)(p_b\mathbb{I}_{(A_{ib}=\hat{t}_i)} + \bar{p}_b\mathbb{I}_{(A_{ib}\neq\hat{t}_i)})\Big)dp_b$$

Original BP udpates are:

$$y_{a\to i}^{(k)}(p_a) \propto \mathcal{F}(p_a) \prod_{j\in\partial a\setminus i} \Big\{(p_a + \bar{p}_a + (p_a - \bar{p}_a)A_{ja})x_{j\to a}^{(k)}(+1) + (p_a + \bar{p}_a - (p_a$$

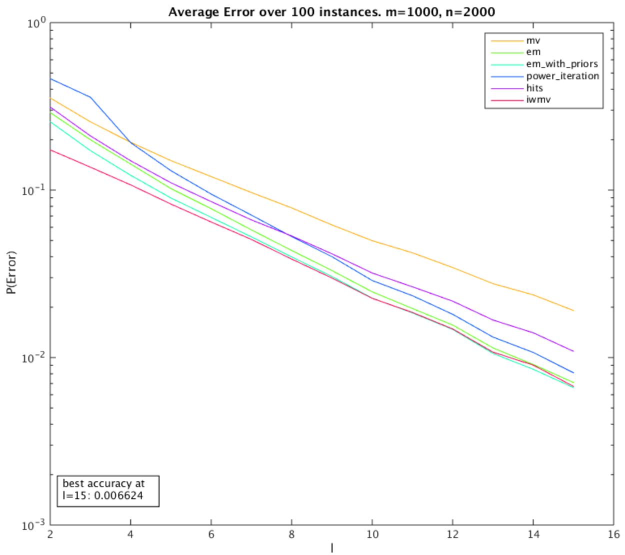
$$x_{i\to a}^{(k+1)}(\hat{t}_i) \propto \prod_{b\in\partial i\setminus a} \int \Big(y_{b\to i}^{(k)}(p_b)(p_b \mathbb{I}_{(A_{ib}=\hat{t}_i)} + \bar{p}_b \mathbb{I}_{(A_{ib}\neq\hat{t}_i)})\Big) dp_b$$

Discretize p_j:

$$x_{i\to a}^{(k+1)}(\hat{t}_i) \propto \prod_{b\in\partial i\setminus a} \sum_{p_b\in P} \Big(y_{b\to i}^{(k)}(p_b)(p_b \mathbb{I}_{(A_{ib}=\hat{t}_i)} + \bar{p}_b \mathbb{I}_{(A_{ib}\neq\hat{t}_i)})\Big)$$

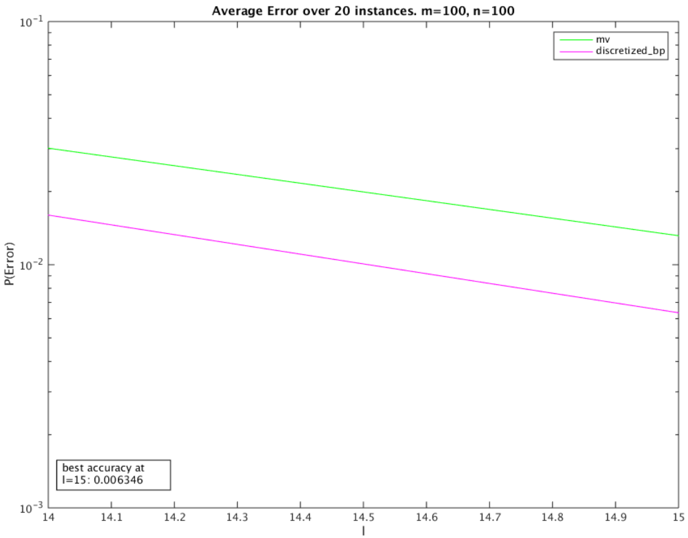
For $P = \{p_1, p_2, p_3, ..., p_k\}$

GRAPH



Average Error over 100 instances. m=1000, n=2000
P(Error)
l
mv
em
em_with_priors
power_iteration
hits
iwmv
best accuracy at
l=15: 0.006624

GRAPH



Average Error over 20 instances. m=100, n=100
mv
discretized_bp
best accuracy at
l=15: 0.006346
P(Error)
l

TABLE

Algorithm	m, n	$\ell = 15$
Discretized BP	100,100	0.006346
EM with Priors	100,100	0.008478
EM with Priors	1000,2000	0.006624
EM with Priors	250,2000	0.005512
Simplified BP	250,2000	0.005466
IWMV	250,2000	0.0057