Seeking the Truth in a Decentralized Manner

Zhiying Xu

January 4, 2018

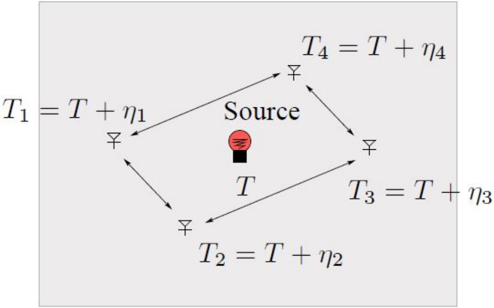
Outline

Outline

Outline

Gossip Algorithm

Gossip algorithms: schemes which distribute the computation burden and in which a node communicates with a randomly chosen neighbor.



Seeking the Truth

- Seeking the truth from noisy crowd sensing data
- Evaluate the reliability of sensors

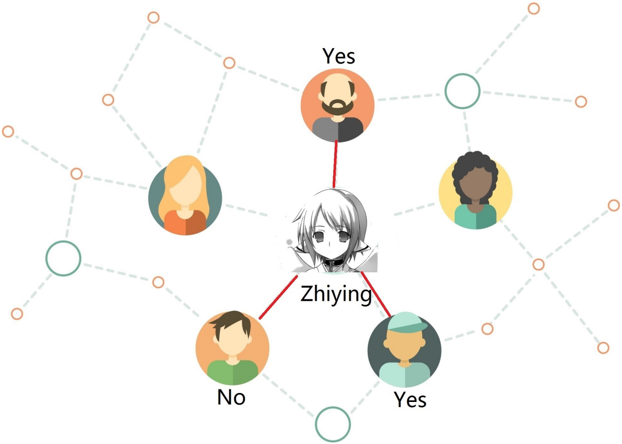
Outline

A Real-life Case

Seeking the Truth in a Decentralized Manner

Zhiying Xu

Whether the class is canceled?



Outline

Model

Definition (Problem Formulation)

Given a crowd sensing model where n sensors make individual observations $\boldsymbol{SC}^{n \times m} \in \{0,1\}^{n \times m}$ about a set of m variables, determine the true value of these variables $\boldsymbol{t} \in \{0,1\}^m$ which satisfies

$$\langle \boldsymbol{t}, \boldsymbol{r} \rangle = \arg\max_{\langle \boldsymbol{t}, \boldsymbol{r} \rangle} p\left(\boldsymbol{SC} | \boldsymbol{t}, \boldsymbol{r}\right),$$

where $\boldsymbol{r} \in [0,1]^n$ is the reliability of sensors.

Model

Definition (Problem Formulation)

Given a crowd sensing model where n sensors make individual observations $\boldsymbol{SC}^{n\times m} \in \{0,1\}^{n\times m}$ about a set of m variables, determine the true value of these variables $\boldsymbol{t} \in \{0,1\}^m$ which satisfies

$$\langle \boldsymbol{t}, \boldsymbol{r} \rangle = \arg\max_{\langle \boldsymbol{t}, \boldsymbol{r} \rangle} p\left(\boldsymbol{SC} | \boldsymbol{t}, \boldsymbol{r}\right),$$

where $\boldsymbol{r} \in [0,1]^n$ is the reliability of sensors.

- $\boldsymbol{t}, \boldsymbol{r}$ are related to each other.
- There is no need for iteration by EM algorithm.
- We denote the ith row of the $\boldsymbol{SC}$ as $\boldsymbol{s}_i$ and the jth column of the $\boldsymbol{SC}$ as $\boldsymbol{c}_j$

Objective Function

Using the relationship between $\boldsymbol{t}$ and $\boldsymbol{r}$, We can simplify the previous problem into the optimization of the objective function

$$\boldsymbol{t} = \arg\min_{\boldsymbol{t}} \sum_{i=1}^{n} f\left(\boldsymbol{s}_i, \boldsymbol{t}\right),$$

where $f\left(\boldsymbol{s}_i, \boldsymbol{t}\right)$ denotes
$-\frac{||\boldsymbol{s}_i - \boldsymbol{t}||}{m} \ln \frac{||\boldsymbol{s}_i - \boldsymbol{t}||}{m} - \left(1 - \frac{||\boldsymbol{s}_i - \boldsymbol{t}||}{m}\right) \ln \left(1 - \frac{||\boldsymbol{s}_i - \boldsymbol{t}||}{m}\right)$.

Truth Discovery Strategy

- All sensors are in a connected networks.
- Sensors have limited memory and can't store much data.
- Each sensor communicates with a randomly chosen neighbor asynchronously.

Outline

NP Hardness of the Decision Version

Given a matrix $\boldsymbol{SC}^{n \times m}$, is there a vector $\boldsymbol{t}$ which satisfies $\sum_{i=1}^{n} f\left(\boldsymbol{s}_i, \boldsymbol{t}\right) < C$?

Proof of NP Hardness

Theorem

Discovering the truth $\boldsymbol{t}$ *from the observed matrix* $\boldsymbol{SC}$ *is NP-hard.*

- By reduction from 3 exact cover problem.
- Inspired by Cardinal, J., Fiorini, S. and Joret, G., 2012. Minimum entropy combinatorial optimization problems. Theory of Computing Systems, 51(1), pp.4-21.

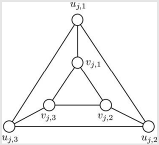
Proof of NP Hardness

Proof: Step 1

3 exact cover: Given a set system (U, S), decide whether U can be covered using pairwise disjoint sets from S

$$\Downarrow$$

minimum orientation defined on a concave function:

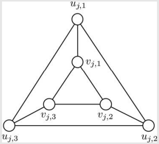


Construct a graph by creating a gadget for each element in U.
Given the graph, decide whether there is an orientation that satisfies $\sum_{i=1}^{|V|} g\left(\frac{d_i^+}{|E|}\right) < C_1$.
Here $g(\cdot)$ is a concave function and d_i^+ is the in-degree of the node i.

Proof of NP Hardness

Proof: Step 2

minimum orientation defined on a concave function:



Given the graph, decide whether there is an orientation that satisfies $\sum_{i=1}^{|V|} g\left(\frac{d_i^+}{|E|}\right) < C_1$. Here $g(\cdot)$ is a concave function and d_i^+ is the in-degree of the node i.

$$\Downarrow$$

Truth Discovery:

Convert the graph into a matrix $\boldsymbol{SC}$, decide whether there is a vector $\boldsymbol{t}$ which satisfies $\sum_{i=1}^{n} f(\boldsymbol{s}_i, \boldsymbol{t}) < C_2$.

$\boldsymbol{SC}$ = ancillary part + graph part.

Outline

Outline

The Complexity of Exact Algorithm

Theorem

The complexity of the exact algorithm is $\Theta\left(mn2^{m'}\right)$.

The Complexity of Exact Algorithm

Theorem

The complexity of the exact algorithm is $\Theta\left(mn2^{m'}\right)$.

- m denotes the number of vectors in the largest set $\{\boldsymbol{c}_i'\} \subset \{\boldsymbol{c}_i\}$ with $\boldsymbol{c}_i' \neq \boldsymbol{c}_j'$, $\boldsymbol{c}_i' \neq \tilde{\boldsymbol{c}}_j'$, which can be obtained in polynomial time.
- The truth vector $\boldsymbol{t}$ is a combination of vectors in $\{\boldsymbol{c}_i'\}$.

Outline

Trade-off in Dencentralized Algorithm

- It's hard to calculate the base vectors in a decentralized ways, so we decide to try every vector in $\{0,1\}^m$
- There is a trade-off between run time and accuracy.

$$\text{run time: } \Theta\left(\frac{2^{m'}\left(m'+\log^{-1}\epsilon\right)\log n}{n\log\lambda_2\left(\boldsymbol{W}\right)^{-1}}\right)$$

$$\text{accuracy: } 1+\epsilon$$

Algorithm

Algorithm 1 Gossip-based Decentralized Exact Algorithm (GDEA)

Input: Observed matrix $\boldsymbol{SC}$

Output: Truth vector $\boldsymbol{t}^*$

Initialize: $\boldsymbol{t}^* = \boldsymbol{t} = 0, \boldsymbol{s}^* = \boldsymbol{0}, \min = 1, \max = 0$

for $j = 1$ to m **do**

  if the j^{th} element in $\boldsymbol{s}^*$ equals 0 **then**

    for $i = 1$ to n **do**

      if $S_iC_j = 0$ **then**

        Invert $\boldsymbol{s}_i$

    Implement $\mathcal{A}(P, \text{ave})$ to obtain $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i$ and mark the elements in $\boldsymbol{s}^*$ as j whose corresponding elements in $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i$ are 0s or 1s

for $j = 1$ to $2^{m'}$ **do**

  We obtain $\boldsymbol{s}_i'$ and $\boldsymbol{s}''$ from $\boldsymbol{s}^*$ and set $\boldsymbol{t} = \bigoplus_{i=1}^{m'} \boldsymbol{s}_i'''$, where $\boldsymbol{s}_i''' \in \{\boldsymbol{s}_i', \boldsymbol{s}_i''\}$. Implement $\mathcal{A}(P, \text{ave})$ to obtain $\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t})$ and $\frac{1}{n}\sum_{i=1}^{n} r_i$

  if $\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t}) < \min$ or $\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t}) = \min$ and $\frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\|\boldsymbol{t}^* - \boldsymbol{s}_i\|_1}{m} > \max$ **then**

    $\boldsymbol{t}^* = \boldsymbol{t}, \min = \frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t}), \max = \frac{1}{n}\sum_{i=1}^{n} r_i$

Acceleration

- Accelerated by optimizing the probability of gossiping with neighbors.
- Preferential Attachment Graphs: $\Theta\left(\frac{m'2^{m'}\log n}{n}\right)$
- Ramdom Geometric Graphs: $\Theta\left(\frac{m'2^{m'}\log n}{n\ell^2}\right)$

Outline

The Complexity of Exact Algorithm

Theorem

The approximation algorithm achieves an approximation ratio of 1.64.

The Complexity of Exact Algorithm

Theorem

The approximation algorithm achieves an approximation ratio of 1.64.

- The approximation ratio is an upper bound, for the worst case is difficult to find.
- Use triangle inequality $f(\boldsymbol{x}, \boldsymbol{y}) + f(\boldsymbol{y}, \boldsymbol{z}) > f(\boldsymbol{x}, \boldsymbol{z})$.

Outline

Trade-off in Dencentralized Algorithm

- Use triangle inequality in metric space and Chebyshev Inequility.
- There is a trade-off between run time and accuracy.

$$\text{run time: } \Theta\left(\frac{\log n\left(\log n+\log \epsilon_1^{-1}\right)}{\epsilon_2^2\left(1-\lambda_{\max}\left(\boldsymbol{W}_1\right)\right)}+\frac{\log n}{n\Phi\left(\boldsymbol{P}_2\right)}+\frac{\log n+\log m}{n\log\lambda_2\left(\boldsymbol{W}_3\right)}\right)$$

$$\text{accuracy: } 1+\epsilon_1+\epsilon_2$$

Algorithm

Algorithm 2 Gossip-based Decentralized Approximation Algorithm (GDAA)

Input: Observation matrix $\boldsymbol{SC}$
Output: Truth vector $\boldsymbol{t}^*$

Initialize: $t_1 = t_2 = 0$
for $i = 1$ to k_2 **do**
  Implement $\mathcal{A}(\boldsymbol{P}_1, \text{mix})$ and sample every k_1 ticks to estimate $\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{s}_k)$ in participant S_k
Implement $\mathcal{A}(\boldsymbol{P}_2, \text{min})$ to find 1-median point $\boldsymbol{t}_1$
if $||\boldsymbol{s}_i - \boldsymbol{t}_1||_1/m > \frac{2}{3}$ **then**
  Invert $\boldsymbol{s}_i$
else if $\frac{1}{3} \leq ||\boldsymbol{s}_i - \boldsymbol{t}_1||_1/m \leq \frac{2}{3}$ **then**
  $\boldsymbol{s}_i = \frac{\boldsymbol{1}}{\boldsymbol{2}}$
Implement $\mathcal{A}(\boldsymbol{P}_3, \text{ave})$ to find $\boldsymbol{t}_2 = round\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i\right)$
Implement $\mathcal{A}(\boldsymbol{P}_3, \text{ave})$ to calculate $\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t}_1)$, $\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{s}_i, \boldsymbol{t}_2)$, $\frac{1}{n}\sum_{i=1}^{n} \frac{||\boldsymbol{s}_i - \boldsymbol{t}_1||_1}{m}$ and $\frac{1}{n}\sum_{i=1}^{n} \frac{||\boldsymbol{s}_i - \boldsymbol{t}_2||_2}{m}$, and choose the one with the larger average value of distances between $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ as $\boldsymbol{t}^*$
if $\frac{1}{n}\sum_{i=1}^{n} r_i < 0.5$ **then**
  Inverse $\boldsymbol{t}^*$

Acceleration

- Accelerated by optimizing the probability of gossiping with neighbors.
- Preferential Attachment Graphs: $\Theta\left(\frac{\log^2 n}{\ell^2}\right)$
- Ramdom Geometric Graphs: $\Theta\left(\log^2 n\right)$

Boyd, S., Ghosh, A., Prabhakar, B. and Shah, D., 2005, March. Gossip algorithms: Design, analysis and applications. In INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE (Vol. 3, pp. 1653-1664). IEEE.

Wang, D., Kaplan, L., Le, H. and Abdelzaher, T., 2012, April. On truth discovery in social sensing: A maximum likelihood estimation approach. In Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on (pp. 233-244). IEEE.

Indyk, P., 1999, May. Sublinear time algorithms for metric space problems. In Proceedings of the thirty-first annual ACM symposium on Theory of computing (pp. 428-434). ACM.

Cardinal, J., Fiorini, S. and Joret, G., 2012. Minimum entropy combinatorial optimization problems. Theory of Computing Systems, 51(1), pp.4-21.

Thank You!