Gossip-based Truth Discovery

Zhiying Xu

May 7, 2017

1 Objective Function

The objective function is based on [1]. However, we adopt a discretized truth vector here to avoid in complicated EM algorithm.

Suppose there are n sensors in an arbitrary connected sensor network and each sensor is asked to answer m questions. The questions for each sensor are same and each answer can only be 0 or 1. $\boldsymbol{X}^{n\times m} \in \{0,1\}^{n\times m}$ is the observed matrix of the answer from the whole network. In $\boldsymbol{X}$, x_{ij} denote the answer of question j from sensor i, $\boldsymbol{x}_i$ is the ith row of the $\boldsymbol{X}$, and $\boldsymbol{x}_j$ is the jth column of the $\boldsymbol{X}$. $\boldsymbol{t}^m \in \{0,1\}^m$ is the truth vector, the true answers of the m questions. $\boldsymbol{r}^n \in [0,1]^n$ is the reliability vector, the possibility of telling truth of each sensor.

We try to discover the truth vector through observing the $\boldsymbol{X}$, that is,

$$\max p\left(\boldsymbol{X}|\boldsymbol{t},\boldsymbol{r}\right).$$

Here $\boldsymbol{t}$ and $\boldsymbol{r}$ are related to each other, we further simplify the object function.

For question j, we denote sensors that observe 0 the set S_{j0} and those that observe 1 the set S_{j1}.

$$p\left(\boldsymbol{x}_j|\boldsymbol{r},t_j=0\right) = \prod_{i\in S_{j0}} r_i \prod_{j\in S_{j1}} \left(1-r_i\right)$$
$$p\left(\boldsymbol{x}_j|\boldsymbol{r},t_j=1\right) = \prod_{i\in S_{j0}} \left(1-r_i\right) \prod_{j\in S_{j1}} r_i$$

We decide the value of t_i by comparing $p\left(\boldsymbol{x}_j|t_i=0\right)$ and $p\left(\boldsymbol{x}_j|t_i=1\right)$. Taking all the m into consideration, we have

$$p\left(\boldsymbol{X}|\boldsymbol{t},\boldsymbol{r}\right) = p\left(\boldsymbol{X}|\boldsymbol{r}\right) = \prod_{j=1}^{m} \max\left\{\prod_{i\in S_{j0}} r_j \prod_{i\in S_{j1}}\left(1-r_i\right), \prod_{i\in S_{j0}}\left(1-r_i\right)\prod_{i\in S_{j1}} r_i\right\}$$

$$\begin{aligned}
\ln p\left(\boldsymbol{X}|\boldsymbol{r}\right) &= \sum_{j=1}^{m}\max\left\{\sum_{i\in S_{j0}}\ln r_i + \sum_{i\in S_{j1}}\ln\left(1-r_i\right), \sum_{i\in S_{j0}}\ln\left(1-r_i\right) + \sum_{i\in S_{j1}}\ln r_i\right\} \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{1}{2}\left(\sum_{i=1}^{n}\ln r_i + \sum_{i=1}^{n}\ln\left(1-r_i\right) + |\sum_{i\in S_{j0}}\ln r_i + \sum_{i\in S_{j1}}\ln\left(1-r_i\right) - \sum_{i\in S_{j0}}\ln\left(1-r_i\right) - \sum_{i\in S_{j1}}\ln r_i|\right) \\
&= \frac{1}{2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\ln r_i\left(1-r_i\right) + |\sum_{i=1}^{n} x_{ij}\ln\frac{r_i}{1-r_i}|\right).
\end{aligned}$$

We get reliability vector $\boldsymbol{r}$ from $\arg\max_{\boldsymbol{r}} \ln p\left(\boldsymbol{X}|\boldsymbol{r}\right)$. Traditionally, to maximize $p\left(\boldsymbol{X}|\boldsymbol{t},\boldsymbol{r}\right)$, an EM algorithm is required. However, there is no need for iteration here. The reason is that future $\boldsymbol{r}$ won't change after we updating $\boldsymbol{t}$ based on current $\boldsymbol{r}$. Specifically, the result of r_i base on $\boldsymbol{t}$ is $\frac{||\boldsymbol{x}_i-\boldsymbol{t}||_1}{m}$.

In this way, the object function can also be written as

$$\arg\max_{\boldsymbol{t}}\sum_{i=1}^{n}\left(d_i\ln d_i + \left(1-d_i\right)\ln\left(1-d_i\right)\right), \text{ where } d_i = \frac{1}{m}||\boldsymbol{x}_i-\boldsymbol{t}||_1.$$

2 Opitimal Solution

Theorem 1. *The optimal solution depends on the rank of the observed matrix.*

Proof. content... □

3 NP-hardness

Theorem 2. *Finding the truth vector is NP-hard.*

Proof. We prove the NP-hardness through reduction from exact 3-cover problem. The exact 3-cover asks, given set U and $S \subseteq \binom{U}{3}$, to decide if there exists $S' \subseteq S$, where S' is a partition of U.

Our proof can be divided into 3 parts. (1) We construct a graph based on U and S and define function on the nodes. (2) We prove minimizing the sum of function of all nodes is NP-hard. (2)We derive a matrix from the graph, and the value of object function of the matrix equals to the sum of function in all nodes.

(1) We construct a graph $G = (V, E)$ same as graph described in [3]. We creates a vertex s_i per set S_i and a copy of gadgets per elements u_j. We links $u_{j,k}$ into $S_{j,k}$, where j_1, j_2, j_3 are the indices of the three sets containing u_j.

We transform the undirected graph into a directed graph. The total degree of node v is denoted as $d(v)$ The in-degree and out-degree of the node v is denoted $d_{in}(v)$ and $d_{in}(v)$. We define a function for v.

$$f(v) = -()$$

□

References

[1] Wang, Dong, et al. "On truth discovery in social sensing: A maximum likelihood estimation approach." Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on. IEEE, 2012.

[2] Boyd, Stephen, et al. "Randomized gossip algorithms." IEEE/ACM Transactions on Networking (TON) 14.SI (2006): 2508-2530.

[3] Cardinal, Jean, Samuel Fiorini, and Gwenal Joret. "Minimum entropy orientations." Operations research letters 36.6 (2008): 680-683.