Tabular Data

Definition: We target the crowdsourcing of a two-dimensional table $C = \{c_{ij}\}$, where $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$. $C$ has an entity attribute which is the key attribute of the table. Each column is a categorical or a continuous attribute. Each cell $c_{ij}$ represents the value of the $i$-th entity in the $j$-th attribute, whose true value (i.e., truth, or ground truth) is denoted as $T_{ij}$.

Example:

<table>
<thead>
<tr>
<th>Name</th>
<th>Nationality</th>
<th>Age</th>
<th>Notability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leonardo DiCaprio</td>
<td>United States</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>Jet Li</td>
<td>China</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>James Purefoy</td>
<td>Great Britain</td>
<td>53</td>
<td>3</td>
</tr>
</tbody>
</table>

Entity attr | Continuous attr | Categorical attr

Problem: A task is related to a cell $c_{ij}$ and the workers are asked to answer the task, by providing values for the cell. Given the set of answers $\{a_{ij}^u\}$, by workers $u$ to cells $c_{ij}$, our target is to obtain the truth table including all the accurate estimates $\hat{T}_{ij}$ for each cell $c_{ij}$’s true value $T_{ij}$.

Existing Methods[3]

For categorical data only:
- Majority Voting (MV) determines the correct labels based on the majority of answers from workers.
- EM iteratively estimates each worker’s confusion matrix, which is used to infer the correct labels.
- GLAD is a probabilistic approach for categorical data.

For continuous data only:
- Median uses the median of workers’ answers as the estimated true value.
- GTM[2] is a truth-finding method specially designed for continuous data.

For both categorical and continuous data:
- CRH[1] detects truth from heterogeneous data types by minimizing a loss function.

Weaknesses: Existing work often treats related attributes independently, leading to suboptimal performance. T-Crowd integrates each worker’s answers on different attributes to effectively learn his/her trustworthiness and true values.

Quality of a Worker

For categorical types, $q_u \in [0,1]$ indicates the probability that the worker $u$ would correctly answer a task traditionally [1], i.e.,

$$P(a_{ij}^u = z) = (q_u)^{1_{\{\text{true}\}}} \cdot (1 - q_u)^{1_{\{\text{false}\}}}$$

where $1_{\{\text{true}\}}$ is an indicator function and $z$ is possible answer in $L$.

For continuous types, we model the distribution of $a_{ij}^u$ given by worker $u$ as a normal distribution $a_{ij}^u \sim \mathcal{N}(\hat{T}_{ij}, \phi_u)$ [2]:

$$P(a_{ij}^u = z) = \frac{1}{\sqrt{2\pi}\phi_u} \exp\left(-\frac{(x - \hat{T}_{ij})^2}{2\phi_u}\right)$$

Where $\hat{T}_{ij}$ is the expected value of $c_{ij}$ and $\phi_u$ is the variance of $u$. Inspired by the idea of discretization of continuous data, we model categorical quality $q_u$ as the area under the normal distribution curve $\mathcal{N}(\hat{T}_{ij}, \phi_u)$ in a small range $\varepsilon$ around the truth $\hat{T}_{ij}$:

$$q_u = P(\hat{T}_{ij} - \varepsilon < a_{ij}^u < \hat{T}_{ij} + \varepsilon) = \text{erf}(\varepsilon / \sqrt{2\phi_u})$$

where $\varepsilon$ is a general parameter that controls the shape of the area and erf is the Gauss error function.

Difficulty of a Cell

The quality of answer $a_{ij}^u$ depends on the quality of worker $u$, the difficulty $\beta_j$ of attribute (i.e., column) $j$, and the difficulty $\alpha_i$ of entity (i.e., row) $i$. To incorporate the difficulty, the variance of continuous answer to $c_{ij}$ as $\phi_u = \alpha_i \beta_j \phi_u$. Based on above equation, the categorical quality is $q_u = \text{erf}(\varepsilon / \alpha_i \beta_j \phi_u)$.

Inference Process

Note that $\hat{T}_{ij}$, $\alpha_i$, $\beta_j$ and $\phi_u$ are unknown which need to compute. The objective function of the problem is to maximize the likelihood of workers’ answers, i.e.,

$$\arg \max_{\alpha_i, \beta_j, \phi_u} P(A|\alpha, \beta, \phi) = \arg \max_{\alpha_i, \beta_j, \phi_u} \sum_{T} P(A, T|\alpha, \beta, \phi)$$

where $A$ is the current set of answers and $T_{ij}$ denotes the estimated distribution of truth in cell $c_{ij}$.

To optimize this non-convex function, we use the EM algorithm, which takes an iterative approach. In each iteration of EM, the E-step computes the hidden variables $T_{ij}$, and the M-step computes the parameters $\alpha_i$, $\beta_j$ and $\phi_u(q_u)$.

Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Celebrity Error Rate</th>
<th>Celebrity MNAD</th>
<th>Restaurant Error Rate</th>
<th>Restaurant MNAD</th>
<th>Emotion Error Rate</th>
<th>Emotion MNAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Crowd</td>
<td>0.0441</td>
<td>0.6339</td>
<td>0.1855</td>
<td>0.5607</td>
<td>0.5961</td>
<td></td>
</tr>
<tr>
<td>CRH</td>
<td>0.0460</td>
<td>0.6737</td>
<td>0.1921</td>
<td>0.5835</td>
<td>0.7224</td>
<td></td>
</tr>
<tr>
<td>Maj. Voting</td>
<td>0.0573</td>
<td>/</td>
<td>0.2003</td>
<td>/</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td>0.0620</td>
<td>/</td>
<td>0.2463</td>
<td>/</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>GLAD</td>
<td>0.0498</td>
<td>/</td>
<td>0.1905</td>
<td>/</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>/</td>
<td>0.6998</td>
<td>/</td>
<td>0.6784</td>
<td>0.7026</td>
<td></td>
</tr>
<tr>
<td>GTM</td>
<td>/</td>
<td>0.6516</td>
<td>/</td>
<td>0.5871</td>
<td>0.6792</td>
<td></td>
</tr>
</tbody>
</table>

References