From Truth Discovery to Trustworthy Opinion Discovery: An Uncertainty-Aware Quantitative Modeling Approach

Mengting Wan[1], Xiangyu Chen[2], Lance Kaplan[3], Jiawei Han[2], Jing Gao[4], Bo Zhao[5]

[1] University of California, San Diego, USA [2] University of Illinois, Urbana-Champaign, USA [3] U.S. Army Research Laboratory, USA [4] SUNY Buffalo, USA [5] LinkedIn, USA

Introduction

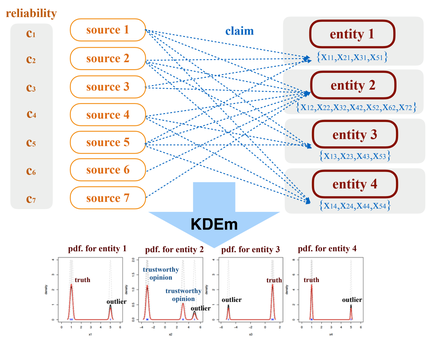

Figure 1: General Workflow: from Truth Discovery to Trustworthy Opinion Discovery.

Numerous claims about the same entity can be collected from multiple sources and they are usually not consistent. How to integrate and summarize conflicting claims and find reliable information?

- **Truth Discovery**: given conflicting information, resolve it and find the most trustworthy fact (i.e. the concept of truth) by introducing **source reliability**.
- **Trustworthy Opinion Discovery**: replace the concept of truth by trustworthy opinion, regard it as a **random variable**, estimate its probability distribution and summarize representative values (i.e. modes)
  - categorical data: easy to tackle since claim confidence scores can be obtained;
  - numerical data: nontrivial to model in an uncertainty-aware way! (we will solve it in this study)

	Truth Discovery	Trustworthy Opinion Discovery
input	entities; claims; sources.	
target	truth (fixed value)	trustworthy opinion (random variable)
output	*value* for truth	*probability distribution* for opinion - if truth exists: value for truth - otherwise: single or multiple representative values
source reliability?	Yes	
multi-modality detection?	No	Yes
Anomaly detection?	No	Yes
Robust to outliers? (numeric data)	No	Yes

Table 1: Truth Discovery v.s. Trustworthy Opinion Discovery.

Method

- Intuition: from a **fixed value** to a **random variable**, from a real coordinate space to a function space.
- Proposed method: Kernel Density Estimation from Multiple Sources (**KDEm**)
- Achieve it using kernel techniques, define a mapping using Gaussian kernel:

$$\Phi_i : \mathbb{R}^d \to \mathcal{H}_i$$
$$\boldsymbol{x} \mapsto K_{h_i}(\cdot, \boldsymbol{x}) := \Phi_i(\boldsymbol{x})$$

Output format with and without modeling source reliability: (i: entity; j: source)

$$\overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}\boldsymbol{x}_{ij}}^{\text{sample mean}} \mapsto \overbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}\Phi_i(\boldsymbol{x}_{ij})}^{\text{sample mean function, i.e. KDE}}$$
$$\underbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}w_{ij}\boldsymbol{x}_{ij}}_{\text{weighted sample mean}} \mapsto \underbrace{\frac{1}{m_i}\sum_{j\in\mathcal{S}_i}w_{ij}\Phi_i(\boldsymbol{x}_{ij})}_{\text{weighted sample mean function}}$$

We need to find $f_i \in \mathcal{H}_i$, $i = 1, ..., n$ and $c_j \in \mathbb{R}^+$, $j = 1, ..., m$, which can minimize

$$J(f_1, ..., f_n; c_1, ..., c_m) = \sum_{i=1}^{n}\frac{1}{m_i}\sum_{j\in\mathcal{S}_i} c_j \|\Phi_i(\boldsymbol{x}_{ij}) - f_i\|^2_{\mathcal{H}_i}$$

where $c_1, ..., c_m$ satisfy

$$\sum_{j=1}^{m} n_j \exp(-c_j) = 1.$$

The algorithm to solve proposed optimization problem:

(a) Initialize $c_1^{(0)} = ... = c_j^{(0)} = ... = c_m^{(0)}$;

(b) Update *opinion density function* $\hat{f}_i$ by $\hat{f}_i^{(k+1)} = \sum_{j\in\mathcal{S}_i} w_{ij}^{(k)}\Phi_i(\boldsymbol{x}_{ij})$, where $w_{ij}^{(k)} = \frac{c_j^{(k)}}{\sum_{j'\in\mathcal{S}_i} c_{j'}^{(k)}}$, $i = 1, ..., n$;

(c) Update *source reliability score* c_j by

$$c_j^{(k+1)} = -\log\left(\frac{\frac{1}{n_j}\sum_{i\in\mathcal{N}_j}\frac{1}{m_i}\|\Phi_i(\boldsymbol{x}_{ij}) - \hat{f}_i^{(k+1)}\|^2_{\mathcal{H}_i}}{\sum_{j'=1}^{m}\sum_{i\in\mathcal{N}_{j'}}\frac{1}{m_i}\|\Phi_i(\boldsymbol{x}_{ij'}) - \hat{f}_i^{(k+1)}\|^2_{\mathcal{H}_i}}\right);$$

$j = 1, ..., m$

(d) Repeat (b) and (c) until the total loss $J(f_1, ..., f_n; c_1, ..., c_m)$ does not change.

The output for f_i is defined as the density estimation for the trustworthy opinion of the i-th entity. Then we can summarize representative values based on the density functions (eg. DENCLUE[1]).

	Entity 1	Entity 2	Entity 3	Entity 4
Source 1	1.00	3.00	1.00	0.95
Source 2	1.10	3.10	0.90	1.00
Source 3	0.90	-3.00	-	-
Source 4	-	-3.10	1.10	1.05
Source 5	5.00	5.00	-5.00	5.00
Source 6	-	-2.90	-	-
Source 7	-	-3.05	-	-

Table 2: Ex.1: A toy example for trustworthy opinion discovery.

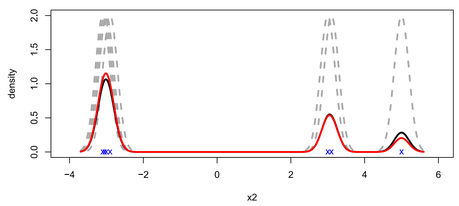

Figure 2: Probability density estimation for Entity 2 in Ex.1.

Experiment

Task 1: Traditional truth discovery from contaminated data (single truth existence can be ensured).

- Performance measure: Rooted Mean Squared Error (RMSE) and Mean Absolute Error (MAE).
- Datasets: *Synthetic(unimodal)* and *Population(outlier)* dataset [2].
- Baselines: **KDE** [3], **RKDE** [4], **Mean**, **Median**, **Voting**, **TruthFinder**[5], **AccuSim**[6], **GTM**[7], **CRH**[8] and **CATD**[9].
- Result: our method **KDEm** has the best performance.

Method	MAE	RMSE
KDEm	**1547**	**8884**
KDE	1630	8900
RKDE	1687	9093
Mean	200917	1136605
Median	11075	129850
Voting	18813	259066
TruthFinder	1551	8892
AccuSim	20819	259948
GTM	317444	1989964
CRH	219596	1289422
CATD	53750	304781

Table 3: Results on the *Population(outlier)* dataset.

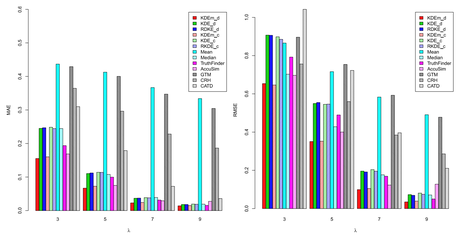

Figure 3: Results on the synthetic uni-modal datasets *Synthetic(uni)*.

Experiment (Continued)

Task 2: Multi-modality detection and anomaly detection (truth existence cannot be ensured).

- Performance measure: Area Under Curve (AUC).
- Dataset: *Synthetic(mix)* and *Tripadvisor* [10] (review rating scores for 8 aspects: *value*, *rooms*, *location*, *cleanliness*, *check in/front desk*, *service*, *business service* and *overall*).
- Baselines: **KDE** [3] and **RKDE** [4].
- Result: our method **KDEm** has the best performance.

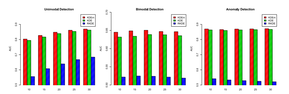

Figure 4: Results on synthetic mixed multi-modal datasets *Synthetic(mix)*.

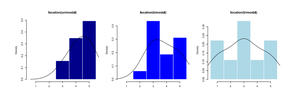

Figure 5: Example histograms in *Tripadvisor(location)*.

Figure 6: Pairwise correlation of source reliability scores and predicted numbers of modals for the *Tripadvisor* datasets.

- Darker ellipse indicates stronger correlation.
- For source reliability scores, the correlation is calculated based on sources which provide claims for both aspects of interest.

References

[1] A. Hinneburg and H.-H. Gabriel, "Denclue 2.0: Fast clustering based on kernel density estimation," in *Advances in Intelligent Data Analysis VII*. Springer, 2007, pp. 70–80.

[2] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *COLING*, 2010.

[3] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[4] J. Kim and C. D. Scott, "Robust kernel density estimation," *JMLR*, vol. 13, no. 1, pp. 2529–2565, 2012.

[5] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 6, pp. 796–808, 2008.

[6] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: the role of source dependence," *PVLDB*, vol. 2, no. 1, pp. 550–561, 2009.

[7] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," *QDB Workshop*, 2012.

[8] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *SIGMOD*, 2014.

[9] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *PVLDB*, vol. 8, no. 4, pp. 425–436, 2014.

[10] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: a rating regression approach," in *SIGKDD*, 2010.

Contact Information

- Data and code: `https://github.com/MengtingWan/KDEm`
- Email: m5wan@ucsd.edu