On the Discovery of Continuous Truth: A Semi-Supervised Approach with Partial Ground Truths

Yi Yang[1], Quan Bai[1], Qing Liu[2]

[1] Auckland University of Technology, New Zealand
Quan.Bai, Yi.Yang@aut.edu.nz
[2] Software & Computational Systems, Data61, CSIRO, Australia
Q.Liu@data61.csiro.au

Abstract In many applications, the information regarding to the same object can be collected from multiple sources. However, these multi-source data are not reported consistently. In the light of this challenge, truth discovery is emerged to identify truth for each object from multi-source data. Most existing truth discovery methods assume that ground truths are completely unknown, and they focus on the exploration of unsupervised approaches to jointly estimate object truths and source reliabilities. However, in many real world applications, a set of ground truths could be partially available. In this paper, we propose a semi-supervised truth discovery framework to estimate continuous object truths. With the help of ground truths, even a small amount, the accuracy of truth discovery can be improved. We formulate the semi-supervised truth discovery problem as an optimization task where object truths and source reliabilities are modeled as variables. The ground truths are modeled as a regularization term and its contribution to the source weight estimation can be controlled by a parameter. The experiments show that the proposed method is more accurate and efficient than the existing truth discovery methods.

Keywords: Truth Discovery, Source relabilities, Semi-supervised learning

1 Introduction

In many applications, the information regarding to the same object can be collected from multiple sources. However, these multi-source data are not reported consistently. For example, different stations may provide different daily high temperature for a city; the stock data provided by different websites may be conflicting. Without resolving conflicts among multi-source data, the data quality cannot be guaranteed, and these data cannot be used in analytic tasks to extract useful information. In the light of this challenge, truth discovery is emerged to resolve conflicts among multi-source data and it has become a hot topic in the community.

Truth discovery resolves conflicts by estimating the truths, i.e. the most trustworthy information, of the objects. Different from the native approaches, such as majority voting (for categorical data) and mean (for continuous data), truth discovery estimates object truths by estimating source reliabilities. The general principle of truth discovery is that sources which frequently provide trustworthy information are reliable, and the data supported by reliable sources is trustworthy and selected as truth of an object. As the source reliabilities are unknown a priori, most existing truth discovery methods explore unsupervised approaches to jointly estimate object truths and source reliabilities, and they assume that the ground truths of objects are entirely unknown. Obtaining the entire set of ground truths from highly reliable sources is usually expensive, but acquiring a small set of ground truths is usually practical. For example, part of the objects' truths may be available from government websites, information released by governments is usually real and we can use it as ground truths. If we can use these partial objects' ground truths and add some supervisions in the truth discovery process, we believe the accuracy of truth discovery can be improved.

A semi-supervised truth discovery method is studied in [19]. However, their method is originally designed for processing categorical object truths. It requires that the ground truths must be among the observations provided by sources. This is impractical for many real world applications, especially when both ground truth and observations are real numbers. As demonstrated in Section 5, the existing methods perform poorly on datasets when object truths are continuous data instead of categorical data.

In this paper, we study the semi-supervised truth discovery problem for continuous object truths. We propose an *Op*timization based *S*emi-supervised *T*ruth *D*iscovery (OpSTD) method for discovering continuous object truths, in which the truth discovery problem is formulated as an optimization task where both object truths and source reliabilities are modeled as variables, and the ground truth is modeled as a regularization term to propagate its trustworthiness to the estimated truths. Furthermore, we present theoretical analysis for the proposed method and conduct a series of experiments on both real datasets and synthetic dataset to demonstrate its effectiveness.

In summary, we make the following contributions:

- We formulate a semi-supervised truth discovery framework, OpSTD, for continuous object truths.
- An iterative based algorithm to estimate source weights and object truths is developed that can converge to an optimal solution.
- We theoretically prove the convergence and analyze the time complexity of OpSTD.
- The experiment results on both real world datasets and synthetic dataset show that the proposed method outperforms the existing methods significantly.

The rest of this paper is organized as follows. In Section 2, we review existing work of truth discovery. In Section 3, we present the optimal framework and the iterative solution. In Section 4, we prove the convergence property of the proposed method and analyze time complexity. In Section 5, it shows the experiments to evaluate the performance of the proposed method. Finally, we conclude in Section 6.

2 Related Work

Truth discovery has received a lot of attentions in the community of data quality, data mining and trust management [2, 8, 9, 11, 20, 23]. Being different from the traditional approaches such as majority voting and mean, truth discovery estimates object truths by taking the source reliabilities into consideration. Most truth discovery methods follows the principle that sources which frequently provide trustworthy observations are reliable, and observations that are supported by reliable sources are trustworthy. Guided by this principle, truth discovery was first proposed by Yin et al. [18]. Since then, various truth discovery methods are proposed to solve various aspects of truth discovery problems. In terms of data type, [3, 7, 14, 15] are developed to process categorical data while [12, 13, 21] are specifically designed for continuous data. In order to resolve conflicts among heterogeneous data, Li et al. [10] proposed a general truth discovery method in which various distance functions can be plugged in to capture the difference between observations. Most existing truth discovery methods assume that each object has only one property. This assumption is relaxed in [5, 6, 16, 22] and these methods are develop to solve multi-truth truth discovery problem. In [3, 4, 17], the authors assume that the sources are not independent. These methods identify object truths by considering the copying relationship among sources.

The existing truth discovery methods can be generally categorized into three ways to formulate the truth discovery task. The *iterative based methods* [3, 7, 18] treat object truths and source weights dependently, and the object truths and source weights are updated iteratively until algorithms converge. The *probabilistic graphical model based methods* [21, 22] model the object truth, source weight and observation as random variables. The dependencies among the random variables are usually captured by a Bayesian network, then the object truths are inferred by probabilistic inference techniques. The *optimization based methods* [10, 12, 13] formulate the truth discovery as an optimization task where the object truths and source reliabilities are modeled as variables. An objective function that involves these variables needs to be optimized and the optimal object truths that minimize the objective function are selected as truths.

The truth output of most truth discovery methods can be categorized into the following two methods. The *scoring method* [18, 19] assigns a trustworthy score to each observation. Then it requires a post decision making process to select the observation as the truth of an object based on the scores of the observations. Usually the observation with the highest score for an object is selected as the truth. The *labeling method* [10, 12, 21] directly assigns a label or a truth to an object. The labeling method is especially helpful when the truth discovery deals with continuous data. In this scenario, the truth of an object might not be observed by any source. For example, the ground truth of temperature of Auckland on June

5th is 26.3. The observations provided by three sources claiming the temperature are 25.6, 26.1 and 26.5, and the ground truth is not among the observations. Thus, the scoring method may fail to work in this case.

There is some work that share similarities with ours. In [15], source reliabilities are modeled as latent variables, its EM based solution can incorporate a small set of ground truths to help truth inference. But it is limited to work with categorical only. The work that is closet to ours is the semi-supervised truth discovery SSTF [19]. SSTF is originally designed for categorical data, and it uses a graph based semi-supervised technique, label propagation, to propagate the trustworthiness of ground truths to the observations. SSTF is limited that it uses scoring technique to output object truths. Therefore, it requires the ground truths are among observations, which is not suitable for the truth discovery applications in which the data is continuous. SSTF also uses a predefined similarity function to capture the relations among observations. This similarity function is application specific and usually hard to define in practice. In contrast, the proposed method OpSTD is specially designed for semi-supervised truth discovery over continuous data and the setting of OpSTD is much simpler. The experiments in Section 5 also demonstrates that OpSTD outperforms SSTF to find continuous object truths.

3 Semi-Supervised Truth Discovery on Continuous Data

In this section, we will formulate the problem of the semi-supervised truth discovery for continuous object truths first. Then the framework and an optimization based method are presented.

3.1 Problem Formulation

We first define the notations for the truth discovery problem, which are also summarized in Table 1.

Definition 1. ***Object, Source and Observation****: An* ***object****, o, is a thing or an event that has a continuous property. A* ***source****, s, is an information provider which can observe and report the property value of object o. An* ***observation****, $v_o^s \in \mathbb{R}$, is the continuous property value of object o reported by source s.*

Definition 2. ***Ground truth and estimated truth****: The* ***ground truth****, $\bar{v}_o^* \in \mathbb{R}$, of object o is the fractal truth that correctly describes the property value of o. It is usually unknown a priori. The* ***Estimated truth****, $v_o^* \in \mathbb{R}$, of object o, is the estimated most trustworthy information describing the property value of o, it is the output of a given truth discovery method.*

Definition 3. ***Source Weight****: The* ***source weight****, $w^s \in \mathbb{R}^+$, reflects the reliability of source s. The information provided by sources with high source weights is usually more trustworthy and closer to the truth.*

Notation	Description
O	Set of all the objects
O_u	Set of objects whose ground truths are unknown
O_g^s	Set of objects whose ground truths are available
O_u^s	Set of objects observed by s, and the objects' ground truths are unknown
O_g	Set of objects observed by s, and the objects' ground truths are available
S	Set of all the sources
S_o	Set of sources that observe object o
V	Set of all the observations
V_u^*	Set of estimated truths for objects in O_u
W	Set of all the source weights
v_o^s	The observation for object o reported by source s
w^s	Weight of source s
$\bar{v_o^*}$	The ground truth of object o
v_o^*	The estimated truth of object o

Table 1: Notations

In this paper, we study the semi-supervised truth discovery for continuous object truths, in which we use some partially available ground truths to supervise the truth discovery process. Let S be the set of all the sources and O be the set of all the objects. We split O into two sets O_g and O_u where O_g and O_u are disjoint and $O_g \cup O_u = O$. O_g is the set of objects whose ground truths are available, and O_u is the set of objects whose ground truths are unknown. Usually $|O_g| << |O_u|$. Next, we formally define the truth discovery task.

Problem Definition. Given the observations V where $V = \{v_o^s\}_{o \in O, s \in S}$ and a set of ground truths $\{\bar{v_o^*}\}_{o \in O_g}$, semi-supervised truth discovery for continuous object truths aims at resolving conflicts among multi-source data and estimating the truths $V_u^* = \{v_o^*\}_{o \in O_u}$ with the help of available ground truths.

3.2 The OpSTD Framework

In this section, we present the OpSTD framework. We formulate the semi-supervised truth discovery as an optimization problem. The intuitions are (1) objects that provide observations closer to the ground truth can be inferred as reliable sources; and (2) the observations reported by reliable sources should be close to the estimated truth. Based on this intuition, we use ground truths to guide the source weight estimation that can in turn impact on the truths estimation for the objects whose ground truths are unknown. Following, we present the optimization framework that can incorporate the available ground truths for truth discovery.

$$\begin{aligned} &\min_{V_u^*, W} f(V_u^*, W) = \sum_{o \in O_u} \Big\{ \sum_{s \in S_o} w^s (v_o^* - v_o^s)^2 \Big\} + \theta \sum_{o \in O_g} \Big\{ \sum_{s \in S_o} w^s (\bar{v_o^*} - v_o^s)^2 \Big\} \\ &\qquad \sum_{s \in S} \exp(-w^s) = 1 \end{aligned} \tag{1}$$

In Equation (1), S_o is the set of sources that observe object o. In the first term $\sum_{o \in O_u}\{\sum_{s \in S_o} w^s(v_o^* - v_o^s)^2\}$, for source s, $(v_o^* - v_o^s)^2$ models the estimated error made by s on the observation for object o, and it computes the discrepancy between the observation provided by sources and the estimated object truths. This term itself estimates the source weights and object truths in an unsupervised manner. In order to minimize f, the optimization process will assign high weights to sources which make small estimated errors. Similarity, if the estimated error is large, it will assign a low weight to w^s to minimize the error's contribution in the objective function.

The second term $\sum_{o \in O_g}\{\sum_{s \in S_o} w^s(\bar{v_o^*} - v_o^s)^2\}$ introduces supervision into the objective function to supervise source weight and object truth estimation process. For a source s, $(\bar{v_o^*} - v_o^s)^2$ models the discrepancy between the ground truth and the source's observation for object o. It is the real error made by s for object o. To minimize the objective function, it penalizes the unreliable sources and assigns low weights to them if the real error is large. θ is a hyper parameter which balances these two terms in the objective function. Combining these two terms makes the proposed framework semi-supervised. The source weights are determined by both estimated errors and real errors, and the ground truths supervises object truths and source weights estimation. This will be further discussed in Section 3.3.

The constraint function, $\sum_{s \in S} \exp(-w^s) = 1$ is required mathematically to constrain the source weights between 0 and 1. Otherwise the source weights can be set to $-\infty$ to minimize the objective function.

3.3 The Iterative Solution

The object truths and source weights shall be learned jointly to minimize the objective function, and the optimal values learned after the optimization process will be selected as the object truths and source weights. In order to minimize the objective function f, we choose to use block coordinate descent [1] in which it iteratively updates one set of variables while fixing the other set to keep reducing the value of f until reaching convergence. There are two steps involved to minimize function f. Step one is to update the estimated truths V_u^* while fixing the source weights W. Step two is to update the source weights W while fixing the estimated truths V_u^*. These two steps can be mathematically formulated by Formulas (2) and (3). Next we discuss in details on how to derive the rules to update source weights and estimated truths.

$$V_u^* \leftarrow \arg\min_{V_u^*} f(V_u^*, W) \tag{2}$$

$$W \leftarrow \arg\min_{W} f(V_u^*, W) \qquad s.t. \qquad \sum_{s \in S} \exp(-w^s) = 1 \tag{3}$$

Object truth update rule: In this step, we update the set of estimated object truths V_u^* while fixing W. By setting $\frac{df_W(V_u^*)}{dv_o^*} = 0$ for the object $o \in O_u$, we get the following update rule for each estimated object truth:

$$v_o^* = \frac{\sum_{s \in S_o} w^s v_o^s}{\sum_{s \in S_o} w^s} \tag{4}$$

Source weight update rule: We use Lagrange multiplier approach to solve Formula (3). The Lagrangian can be formulated as:

$$\mathcal{L}(W, \lambda) = f(V_u^*, W) + \lambda(\sum_{s \in S} \exp(-w^s) - 1) \tag{5}$$

where λ is the Lagrange multiplier. By setting $\frac{d\mathcal{L}(W,\lambda)}{dw^s} = 0$, from the constraint we can derive that

$$\lambda \exp(-w^s) = \sum_{o \in O_u^s} (v_o^* - v_o^s)^2 + \theta \sum_{o \in O_g} (\bar{v}_o^* - v_o^s)^2 \tag{6}$$

where O_u^s and O_g^s are both observed by source s, but their ground truths are unknown and available respectively. Combined with the constraint equation $\sum_{s \in S} \exp(-w^s) = 1$, we can compute the Lagrange multiplier as:

$$\lambda = \sum_{s \in S} \Big\{ \sum_{o \in O_u^s} (v_o^* - v_o^s)^2 + \theta \sum_{o \in O_g^s} (\bar{v}_o^* - v_o^s)^2 \Big\} \tag{7}$$

Plugging Equation (7) back to Equation (6), we can derive the source weight update rule in Equation (8).

$$w^s = -\log\Big(\frac{\sum_{o \in O_u^s} (v_o^* - v_o^s)^2 + \theta \sum_{o \in O_g^s} (\bar{v}_o^* - v_o^s)^2}{\sum_{s \in S} \Big\{ \sum_{o \in O_u^s} (v_o^* - v_o^s)^2 + \theta \sum_{o \in O_g^s} (\bar{v}_o^* - v_o^s)^2 \Big\}} \Big) \tag{8}$$

Discussion: From Equation (8) of the source weight update rule, we can see that a source has higher weight if it makes few errors among all the sources. Specifically, the errors are determined by the estimated errors and real errors, and the proportion can be adjusted by controlling θ. If we increase θ, the source weight will be computed mostly by the real errors. In the extreme case where $\theta = \infty$, the term $\sum_{o \in O_u^s} (v_o^* - v_o^s)^2$ is ignored and the source weight is totally determined by objects with the ground truths. Conversely, if we decrease θ, the source weight will be computed mostly by the estimated errors. If $\theta = 0$, this is equivalent to the truth discovery in an unsupervised setting where the ground truths do not contribute to the truth discovery process and we estimate source weights solely from the observations. In Section 5.3, we will experimentally show the effect of θ to the performance of OpSTD.

From Equation (4) we can see that the estimated object truth is computed by weighted aggregation in which all the observations for object $o \in O_u$ contribute to the estimated truth, but the contribution is discounted by the weights of the sources which provide these observations. As a result, the estimated truth will be close to the observations from sources with high weights. Furthermore, the source weights are partially computed by ground truths as in Equation (8). Thus, the ground truths also impact the truths estimation for the objects whose ground truths are unknown.

The algorithm flow of the OpSTD is summarized in Algorithm 1. First, the source weights are initialized. If no prior knowledge is available about the reliabilities of the sources, the source weights can be initialized uniformly, i.e. $w^s = -\log(\frac{1}{|S|})$. Otherwise the source weights can be changed accordingly to reflect the initial belief of the source reliability. Then the algorithm iteratively update object truths and source weights by Equations (4) and (8) until convergence.

Algorithm 1: OpSTD Algorithm Flow

```
Input  : Observations $V$, ground truths $V_l^*$ for $O_l$
Output : Inferred object truths $V_u^*$
Initialize source weights;
repeat
    for $o \in O_u$ do
        Update $v_o^*$ by Equation (4);
    end
    for $s \in S$ do
        Update $w^s$ by Equation (8);
    end
until Convergence
return $V_u^*$
```

4 Theoretical Analysis

In this section, we theoretically analyze the convergence property of the OpSTD algorithm and its time complexity.

4.1 Convergence Analysis

We prove the following theorem to show the convergence of OpSTD algorithm, and it is valid to use block coordinate descent to minimize the objective function given in Equation (1).

Theorem 1. *The iterative process in OpSTD algorithm converges, and the optimal solutions, V_u^* and W, is a stationary point for the objective function in Equation (1) to attain minimum.*

Proof. There are two blocks of variables, V_u^* and W, involved in the objective function f. We use $\mathcal{Y}$ to denote the union of the two blocks of variables, i.e. $\mathcal{Y} = \{V_u^*, W\}$. Let the size of $\mathcal{Y}$ be l where $l = |V_u^*| + |W|$. Then the optimization problem can be rewritten as:

$$\text{minimize} \quad f(y), \qquad \text{s.t.} \quad y \in \mathcal{Y}$$

According to [1], let $\{y^z\}$ be the sequence generated by the following rule:

$$y_i^{z+1} = \operatorname*{arg\,min}_{\xi \in \mathcal{Y}_i} f(y_1^{z+1}, \ldots, y_{i-1}^{z+1}, \xi, y_{i+1}^z, \ldots, y_l^z) \qquad \text{for} \qquad i = 1, 2, \ldots, l$$

where z is the current iterate index, then every limit point of y^z is a stationary point and $f(\{y^z\})$ is the global minimum of f if f satisfies the following two conditions:

1. f is continuously differentiable over $\mathcal{Y}$.
2. For each $y_i \in \mathcal{Y}_i$, $f(y_1, y_2, \ldots, y_{i-1}, \xi, y_{i+1}, \ldots, y_l)$, viewed as a function of ξ while the other variables are fixed, attains a unique minimum $\bar{\xi}$ over $\mathcal{Y}_i$, and is monotonically non-increasing in the interval from y_i to $\bar{\xi}$.

Next, we show the objective function f satisfies the two above conditions in the following two scenarios:

- Scenario 1: Update V_u^* while fixing W. In this case, $f_W(V_u^*)$ is a combination of quartic functions $w^s(v_o^* - v_o^s)^2$ where $w^s > 0$. Hence, $f_W(V_u^*)$ is strictly convex and continuously differentiable and attains a unique minimum.
- Scenario 2: Update W while fixing V_u^*. In this case, $f_{V_u^*}(W)$ is a combination of linear functions w.r.t w^s, which is affine, strictly convex and continuous differentiable. In addition, the exponential function is strictly convex, the constraint in the objective function is also strictly convex. Thus, $f_{V_u^*}(W)$ is continuously differentiable and attains a unique minimum while fixing V_u^*.

Therefore, Algorithm 1 converges when f attains its minimum $f(y^z)$, and $\{V_u^*, W\} = \{y^z\}$ is the stationary point. □

4.2 Time Complexity Analysis

We analyze the time complexity of OpSTD algorithm by analyzing the computational complexity of each iteration in Algorithm 1. In the object truth update step, each object can be observed by up to $|S|$ sources. The cost of updating object truths is $O(|O_u| \times |S|)$ since this step computes the sum of observations weighted by source weights. In the source weight update step, each source can observe up to $|O|$ objects. The cost of updating source weight is $O(|O| \times |S|)$ since this step computes the squared error between each source's observation and truths. Therefore, the computational complexity of each iteration in OpSTD algorithm is $O(|O| \times |S|)$. In the truth discovery application, there are at most $|O| \times |S|$. Hence, the computational complexity of each iteration is also linear with the number of observations.

5 Experiments

In this section, we experimentally compare the proposed method with the state-of-art truth discovery methods on both real and synthetic datasets. All the experiments are conducted on a PC with Intel i7 processor and 16 GB RAM.

5.1 Experiment Setup

In this subsection, we describe the baseline methods, datasets and performance metrics used to evaluate OpSTD.

Baseline Methods OpSTD is compared with the following state-of-art truth discovery methods:

- GTM [21]: A probabilistic graph model based method for resolving conflicts on continuous data.
- CRH [10]: Finding truth of heterogeneous data by using various loss functions.
- SSTF [19]: A semi-supervised truth discovery method adopts graph based semi-supervised learning method to learn object truths with a small set of ground truths.
- Mean: This method does not consider source reliabilities. It simply aggregates the object truth by taking the mean of the observations.

Datasets We use two real world datasets and one synthetic dataset to evaluate the proposed method. The ground truths of all the datasets are available for evaluation.

- **Weather** [3]: It contains daily weather information for 30 cities over 6 months. The daily temperature property for each city is adopted in the experiments.
- **Stock** [3]:It records data for 1000 stocks collected from 55 sources over 21 working days in 2011. The open price property for each stock is adopted in the experiments.
- **Gas Price**: In order to compare OpSTD against SSTF, we generate a synthetic dataset that is suitable for SSTF. In this dataset, regular gas prices of 500 gas stations in US from Gasbuddy are collected for one day as ground truth. We generate 30 sources with different reliabilities. For each object, we select a random source with high reliability and let it provide ground truth as the observation to the object. The observations provided by the rest of the sources are generated by adding different levels of Gaussian noise based on their reliabilities to the ground truth. Thus, the ground truth of each object is among the observations and it satisfies the condition of SSTF.

[3] http://lunadong.com/fusionDataSets.htm

Method	Dataset					
	Weather		Stock		Gas Price	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
OpSTD	**0.7274**	**1.1546**	**0.0038**	**0.0002**	**0.2264**	**0.0781**
SSTF	N/A	N/A	N/A	N/A	0.2613	0.1057
GTM	0.8196	1.5074	0.0044	0.0004	0.2502	0.0946
CRH	0.7829	1.4518	0.0046	0.0004	0.2525	0.0987
Mean	0.9524	2.2517	0.0128	0.004	0.3156	1.1514

Table 2: Accuracy Comparison

Performance Metrics The data in these datasets are continuous, the difference between estimated truth and ground truth can be measured by their numerical distance. Therefore, we use the following two metrics to evaluate the accuracy of OpSTD:

- Mean Absolute Error (MAE): it measures the mean of the overall absolute error between estimated truth and ground truth.
- Root Mean Square Error (RMSE): it measure the root of the mean squared error between estimated truth and the ground truth.

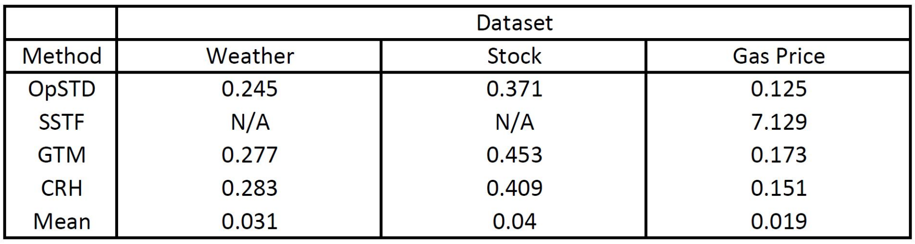
Both MAE and RMSE measure the discrepancy between estimated truth and ground truth. The lower the measure, the more accurate the method is. Being different from MAE, RMSE penalizes heavily on large errors.

We use running times of OpSTD to evaluate its efficiency.

5.2 Performance Comparison

In this section, we report the performance evaluation for OpSTD against the baseline methods on the three datasets. For weather and stock datasets, since the ground truths are not among the observations, it does not satisfy the condition of SSTF, SSTF is not able to estimate object truths for these two datasets. For each dataset, we randomly choose 20% objects and use the ground truths of these objects in the truth discovery process, the ground truths of the rest objects are used for evaluation. For all the baseline methods, we use the best parameters that results in the best performance.

Accuracy Comparison The experiment results conducted on the three datasets in terms of accuracy are summarized in Table 2. As shown in the table, OpSTD consistently achieves the best accuracy in terms of MAE and RMSE. Among all the methods, Mean performs worst because it simply takes the average of observations for each object as truth, which does not take source reliabilities into consideration. Compared with GTM and CRH, OpSTD's error is reduced ranging from 7% - 14% in terms of MAE and 17% - 50% in terms of RMSE over the three datasets. The reason is that these two methods explore an unsupervised approach which does not use ground truths in the truth discovery process. Therefore, their errors are larger compared to our method. OpSTD also outperforms the semi-supervised method SSTF. Note that SSTF's accuracy is even lower than GTM and CRH even if it uses ground truths to estimate object truths. This is because its algorithm is designed for handling categorical data and it runs poorly on continuous data scenarios.



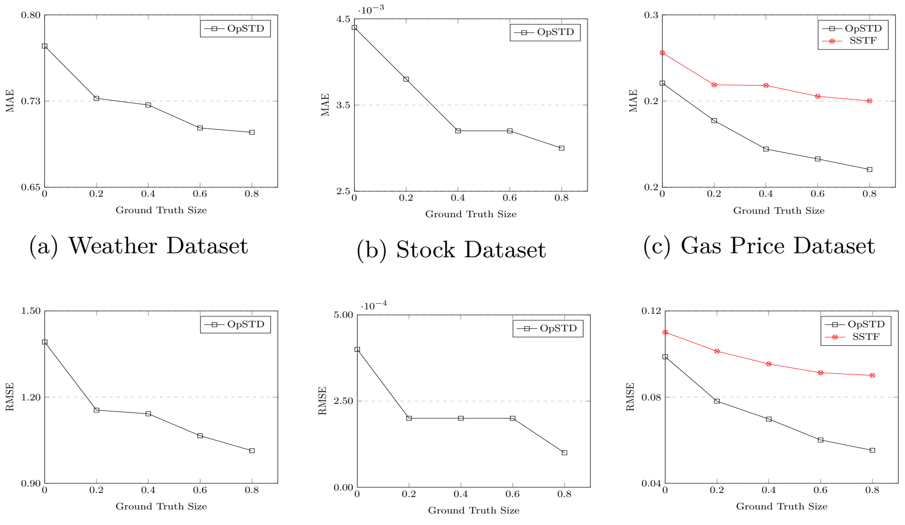
	Dataset		
Method	Weather	Stock	Gas Price
OpSTD	0.245	0.371	0.125
SSTF	N/A	N/A	7.129
GTM	0.277	0.453	0.173
CRH	0.283	0.409	0.151
Mean	0.031	0.04	0.019

Table 3: Running Times (Second(s))

Efficiency The experiment results conducted on the three datasets in terms of running times are summarized in Table 3. From this table we can see that Mean achieves the optimal efficiency. This is because Mean ignores source reliabilities estimation and it outputs mean of observations as truths directly. Among the baseline methods, OpSTD runs about 10% faster than GTM and CRH over the three datasets. The reason is that OpSTD uses 20% ground truths as its input and it estimates the truths for the rest 80% objects, while GTM and CRH discovers truths for the whole dataset. Compared with SSTF, OpSTD runs 57 times faster, which demonstrates the superiority of OpSTD for truth finding with ground truths.

5.3 Sensitivity Analysis

We study the effect of ground truth size and θ on our method. We first evaluate the effect of ground truth size to the accuracy of OpSTD. The θs are fixed at 20, 15 and 35 for weather, stock and gas price datasets, respectively. We vary the available ground truth size over the whole dataset from 0 to 0.8 with the step of 0.2. Since the accuracy of SSTF is also sensitive to the ground truth size, we also use different ground truth sizes to test SSTF on gas price dataset in this experiment. The experiment result is plotted in Figure 1. From Figure 1, on one hand, we can see that both MAE and RMSE are high for all the three datasets when ground truth size is 0. This is the case when no ground truth is used in our method and its accuracy is the same as CRH. As the ground truths are involved in our truth discovery process, the errors begin to drop; on the other hand, we



(a) Weather Dataset

(b) Stock Dataset

(c) Gas Price Dataset

(d) Weather Dataset

(e) Stock Dataset

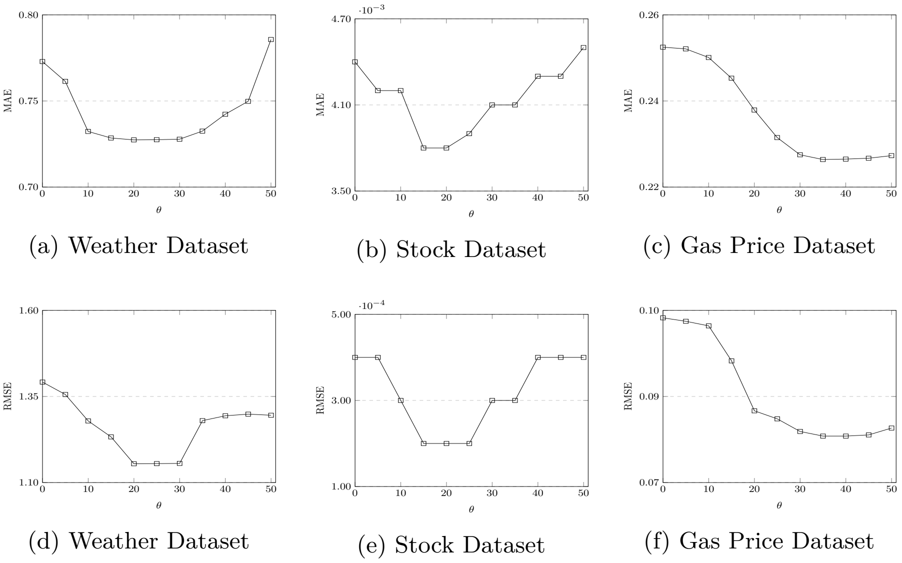
(f) Gas Price Dataset

Figure 1: Effects of Ground Truth Size to MAE and $RMSE$

can also see that the errors are inverse proportional to the size of ground truths. This demonstrates that the ground truth indeed benefits the truth estimation in OpSTD. On the other hand, from Figures 1(c) and 1(f) we can find that OpSTD outperforms SSTD in terms of MAE and RMSE under all ground truth sizes. This shows that OpSTD can utilize ground truths better for truth discovery tasks with continuous object truths.

The effect of θ to the accuracy of our method is plotted in Figure 2. In this experiment, we fix ground truth size at 0.2 and vary θ from 0 to 50. From this figure we can see the errors begin to decrease when θs begin to increase from 0 and reach the optimal error very soon. Being different from the ground truth size, as we keep increasing θ, the errors begin to increase after it reaches the optimal ones. The reason is that as we increase θ, the real errors become significant and it dominates the estimated errors in Equation (8). This may cause the estimated source weights overfit the objects whose ground truths are available, but less general to the rest 80% objects whose object truths are estimated. Given different datasets having different distribution and characteristics, θ is sensitive to OpSTD and we use the best θ to achieve the optimal performance.

In summary, ground truth, even a small set of ground truth, are beneficial for truth discovery. By effectively incorporating ground truths into our method, the accuracy can be improved significantly. When ground truth size is small, theta is sensitive to different datasets and can be tuned to achieve optimal results.



(a) Weather Dataset

(b) Stock Dataset

(c) Gas Price Dataset

(d) Weather Dataset

(e) Stock Dataset

(f) Gas Price Dataset

Figure 2: Effects of θ to MAE and $RMSE$

6 Conclusion

In this paper, we investigate semi-supervised truth discovery method for continuous object truths. We formulate the truth discovery problem as an optimization task in which object truths and source weights are modeled as variables, and the ground truths is formulated as a regularization term to reinforce the source weights. An iterative solution is developed to estimate object truths and source weights and its convergence property and time complexity are analyzed. We also conduct a series of experiments to demonstrate that the proposed method outperforms the existing truth discovery methods in terms of both accuracy and efficiency.

References

1. Bertsekas, D.P.: Nonlinear programming. Athena scientific Belmont (1999)
2. Cho, J.H., Swami, A., Chen, R.: A survey on trust management for mobile ad hoc networks. IEEE Communications Surveys & Tutorials 13(4), 562–583 (2011)
3. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. Proceedings of the VLDB Endowment 2(1), 550–561 (2009)
4. Dong, X.L., Berti-Equille, L., Srivastava, D.: Truth discovery and copying detection in a dynamic world. Proceedings of the VLDB Endowment 2(1), 562–573 (2009)
5. Fang, X.S.: Truth discovery from conflicting multi-valued objects. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 711–715. International World Wide Web Conferences Steering Committee (2017)
6. Fang, X.S., Sheng, Q.Z., Wang, X., Ngu, A.H.: Smartmtd: A graph-based approach for effective multi-truth discovery. arXiv preprint arXiv:1708.02018 (2017)

7. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 131–140. ACM (2010)
8. Lee, Y.W., Pipino, L.L., Funk, J.D., Wang, R.Y.: Journey to data quality. The MIT Press (2009)
9. Li, M., Sun, X., Wang, H., Zhang, Y., Zhang, J.: Privacy-aware access control with trust management in web service. World Wide Web 14(4), 407–430 (2011)
10. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data. pp. 1187–1198. ACM (2014)
11. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM Sigkdd Explorations Newsletter 17(2), 1–16 (2016)
12. Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., Han, J.: On the discovery of evolving truth. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 675–684. ACM (2015)
13. Meng, C., Jiang, W., Li, Y., Gao, J., Su, L., Ding, H., Cheng, Y.: Truth discovery on crowd sensing of correlated entities. In: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems. pp. 169–182. ACM (2015)
14. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 877–885. Association for Computational Linguistics (2010)
15. Pasternack, J., Roth, D.: Latent credibility analysis. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1009–1020. ACM (2013)
16. Pochampally, R., Das Sarma, A., Dong, X.L., Meliou, A., Srivastava, D.: Fusing data with correlations. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data. pp. 433–444. ACM (2014)
17. Qi, G.J., Aggarwal, C.C., Han, J., Huang, T.: Mining collective intelligence in diverse groups. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1041–1052. ACM (2013)
18. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering 20(6), 796–808 (2008)
19. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th international conference on World wide web. pp. 217–226. ACM (2011)
20. Zhang, J., Tao, X., Wang, H.: Outlier detection from large distributed databases. World Wide Web 17(4), 539–568 (2014)
21. Zhao, B., Han, J.: A probabilistic model for estimating real-valued truth from conflicting sources. Proc. of QDB (2012)
22. Zhao, B., Rubinstein, B.I., Gemmell, J., Han, J.: A bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment 5(6), 550–561 (2012)
23. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: is the problem solved? Proceedings of the VLDB Endowment 10(5), 541–552 (2017)