Robust Truth Discovery Scheme Based on Mean Shift Clustering Algorithm

Jingxue Chen, Jingkang Yang, Juan Huang, Yining Liu

Guangxi Key Laboratory of Trusted Software
Guilin University of Electronic Technology, China
1015725207@qq.com, beliefshigu@163.com, 985865656@qq.com, lyn7311@sina.com

Abstract

Data conflict is an inevitable problem in data collection due to the limitations in the real world. How to determine the most reliable data among many collected data is a problem worth studying. Truth discovery is a widely used resolution to integrate heterogeneous data. Existing truth discovery schemes mainly aggregate data from all sources to obtain the truth. However, it is intuitive that exceptional data should be excluded before aggregation, and how to identify exceptional data is also difficult. To tackle this problem, we use the mean shift clustering algorithm to remove exceptional data and obtain the truth value. Experiment results indicate our scheme performs well in various conditions.

Keywords: Truth discovery, Mean shift clustering algorithm, Aggregation, Data conflict

1 Introduction

With the development of big data and sensors [1-2], data collection becomes more and more important. A large amount of data is collected for statistical analysis, prediction results, training models and other purposes [3]. However, in the process of data collection, the data from different sources for the same object have heterogeneity due to the difference in the environment and equipment used by the observers. Hence, how to discover "truth" from these conflicting data is a big problem that must be solved in the process of data collection [4-6].

Truth discovery is a widely used solution to resolve the confliction problem [7-9]. Yin et al. [10] first proposed the concept of truth discovery. They process all the collected data by employing the relationships between data sources and provided data to obtain truth values. Many truth discovery schemes have been proposed to solve conflicts in data integration by aggregating data from different sources to obtain the truth value.

In the real world, data provided from some sources are noisy and far from the "truth" due to equipment failure, extreme conditions and even malicious attacks [11]. Such exceptional data should be excluded so that the truth value is not affected by them. However, the existing methods are mainly to obtain truth by aggregating data from all sources [12-13]. Some methods assign different weights to different data sources to reduce the impact of exception data on the truth value, which does not completely solve the above problem.

In this paper, the truth discovery process is divided into two sub-tasks: removing exceptional data and aggregating normal data. The distribution of exceptional data and normal data are different in feature space. The clustering algorithm can divide all the data in feature space into different clusters [14]. Data with similar features can be divided into the same cluster. By doing this, normal data and exceptional data can be divided into different clusters. In real truth discovery tasks, the exceptional data is various, that is, it may be divided into multiple clusters. Therefore, the number of obtained clusters cannot be predicted before clustering. Mean shift clustering algorithm [15] is used to accomplish the clustering task in this paper because it does not need to set the number of clusters.

The main contributions of our work are summarized as follows:

(1) We propose a novel truth discovery scheme with robustness, which is able to eliminate the affection of exceptional data completely. The proposed scheme divides the truth discovery task into two sub-tasks to remove exceptional data and aggregate normal data separately.

(2) We employ the mean shift clustering algorithm to divide normal data and exceptional data. This is a completely new truth discovery scheme, and to the best of our knowledge, this is the first time to use a clustering algorithm in a truth discovery scheme.

(3) We use a real-world dataset to evaluate the performance of our scheme and baselines. Experimental results indicate that our scheme outperforms baselines in various conditions and

*Corresponding Author: Juan Huang; E-mail: 985865656@qq.com
DOI: 10.53106/160792642021072204011

achieves robustness.

The structure of this paper is organized as follows. Section 2 reviews the existing truth discovery schemes. In Section 3, we describe our proposed method. Section 4 presents our experimental results compared with baseline schemes. Finally, we conclude our work in Section 5.

2 Related Works

The problem of data confliction has been widely studied and many resolutions have been proposed [16-18]. Bleiholder et al. [19] designed an information integration system, which uses multiple conflict handing functions to integrate data. Yin et al. [10] proposed a scheme called Truth Finder to find "truth" among conflicting data, which uses the relationships between sources and data they provide. Jiang et al. [20] take the consequences of incorrect values into consideration. They proposed a framework to exclude incorrect values and select the value that minimizes the expected cost of errors as the "truth" by calculating the conditional probability. Li et al. [11] proposed a CHR framework. In [11], each source has its own weight, which is updated by iteration. The truth value is obtained by calculating the weighted average of the data provided by all sources.

The most common principle of performing truth discovery schemes [21-22] recently mainly includes two stages, weight update and truth update. These schemes are inspired by CHR [11] and can be summarized as that at the initial point of the truth discovery algorithm, an assumed ground truth value is randomly assigned, weight update stage and truth update stage are updated until convergence is achieved. However, all of these schemes do not exclude exceptional data. This common truth discovery algorithm is regarded as the baseline and inspired by all the above excellent schemes, we propose our robust truth discovery scheme based on the mean shift clustering algorithm.

3 Method

In this section, we describe the problem formulation at first. Then we introduce the mean shift clustering algorithm, which is the tool employed to resolve the problem. The framework and process of the proposed scheme are described at last.

3.1 Problem Formulation

In the truth discovery system, there are K participants to observe an object, which has M properties. Each participant submits a collection containing M observations for the M properties of the object. Due to the difference in equipment and conditions, the collections from different participants are heterogeneous. In this case, K heterogeneous observations can be obtained for each property of an object while only one truth value needed. The purpose of truth discovery is to output one collection of M observations for an object, which is regarded as the truth of the object.

In the real world, some exceptional data might be mixed in with the heterogeneous data due to equipment failure, extreme conditions and even attacks from the adversary [11, 23]. Exceptional data should not affect the truth value. The final "truth" should be calculated from normal data.

To resolve the conflicts of heterogeneous data, we divide the truth discovery process into two sub-tasks: removing exceptional data and aggregating normal data. The features of exceptional data are markedly different from those of normal data, it is possible to remove exceptional data from heterogeneous data. The proposed scheme employs a clustering algorithm to cluster the normal data and exceptional data into different clusters. The truth value is aggregated from the cluster of normal data. By doing this, a truth value independent of exceptional data can be obtained.

3.2 Background of Mean Shift Clustering Algorithm

Mean shift is a widely used clustering algorithm [24-28], in which a circular sliding window is set to complete the clustering of data. The center point of the sliding window is constantly updated to the mean value of the points inside the sliding window. In each iteration, the sliding window moves to a denser region until it converges. The main advantage of the mean shift algorithm is that we do not need to pre-specify the classes of the data. In the process of truth discovery, we usually cannot predict the number of types of exceptional data, thus the mean shift algorithm is more suitable in the truth discovery scheme.

The main process of mean shift clustering is shown in Figure 1. Simply, a circle with a fixed radius serves as a sliding window to find dense areas of data points. In each iteration, the sliding window moves in the direction of higher data density. The center point of the sliding window is constantly updated to the mean value of the points inside the sliding window. When the convergence condition is satisfied, the sliding window stop moving and the data points visited during this process are grouped into a cluster. Other sliding windows repeat this process until all the data points are divide into different clusters.

The steps of the mean shift algorithm are as follows:

1. Determine the radius R of the circular sliding window.

2. Start sliding the window from a randomly selected center point c.

3. Visit all the points within the radius R from c to form a set D and group these points into a cluster C.

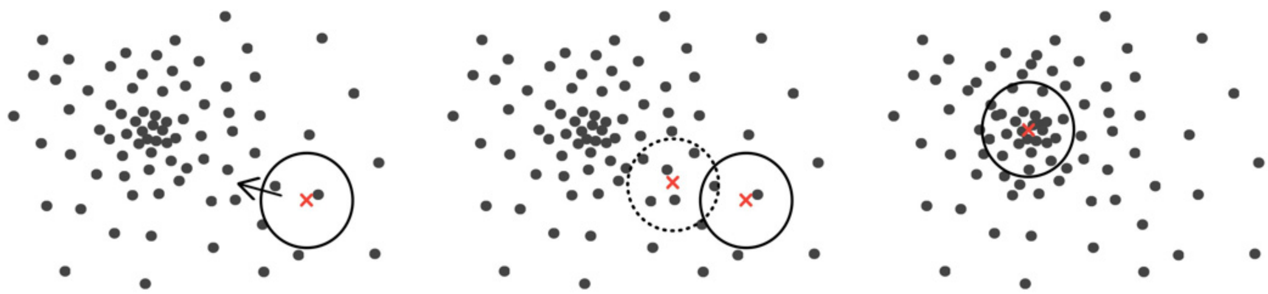

Figure 1. The process of mean shift clustering algorithm

Then add 1 to a label P_i^j for the i -th point belongs to this cluster, which represents the j -th class. P_i^j will be used for the clustering of the last step.

4. Compute the vectors from all the points in D to c and add them up to get the mean shift vector S.

5. Move c in the direction of S and the distance of $|S|$.

6. Repeat steps 3, 4, and 5 until the length of S is reduced to the convergence condition, all the points visited during this process are grouped into cluster C, whose center point is c.

7. Set a threshold μ. When S reaches the convergence condition, if there exist a cluster C' with center point c' and the distance between c' and c is less than μ, merge C' and C. Otherwise, take C as a new cluster.

8. Repeat step 2, 3, 4, 5 and 6 until all points are visited.

9. For each point that is grouped into multiple clusters, group it into the cluster with the biggest label P_i^j, that is, the i -th point with multiple clusters is grouped into j -th cluster finally when P_i^j is the biggest label.

Mean shift vector. The key of mean shift cluster algorithm is to update the center point according mean shift vector. We first show the basic form of the mean shift vector:

$$S = \frac{1}{q}\sum_{x_i \in S_q}(x_i - c) \tag{1}$$

where c denotes the current center point, S_q denotes the set of points whose distance from c is less than the radius R, q is the size of S_q.

In the basic form of the mean shift vector, the farther point from center point within the radius R get the higher weight. We expect that the closer point gets higher weight for center point. Therefore, we use the Gaussian kernel function to calculate the mean shift vector:

$$S = \frac{\sum_{i=1}^{q} x_i G(\|\frac{x_i - c}{h}\|^2)}{\sum_{i=1}^{q} G(\frac{x_i - c}{h})} - c \tag{2}$$

where $G(\cdot)$ denotes Gaussian kernel function, h denotes the bandwidth of Gaussian kernel function.

Center point update. The center point is update by adding the mean shift vector calculated by Eq.(2). New center point is obtained by:

$$c_{new} = c + S \tag{3}$$

where c_{new} denotes the new center point.

3.3 Robust Truth Discovery Scheme

The framework of our scheme consists of three modules: data collecting module, data clustering module and truth value filter.

(1) Data collecting module is responsible for collecting source data from each participant and assign an ID to each participant.

(2) Data clustering module divides the source data into clusters using a mean shift clustering algorithm and calculates the center points of each cluster, where n is unknown before clustering.

(3) Truth value filter traverses all the clusters and filters out the exceptional data, and then obtains the truth value of the participants.

Algorithm 1. Mean Shift Clustering Algorithm

Input: Data points from K sources: $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$.

Output: n clusters $\{C^{(1)}, C^{(2)}, \ldots, C^{(n)}\}$ and their center points $\{c^{(1)}, c^{(2)}, \ldots, c^{(n)}\}$.

1. Initialize the radius R;
2. **repeat**
3. Randomly select a point from all the unvisited points as center point c;

repeat
  Place the sliding window with c as the center and R as the radius;
  Count the number of points within the sliding window as q;
  for $i := 1$ to q **do**
    $P_i^j := P_i^j + 1$;
  end for
  Calculate mean shift vector S according to Eq.(2);
  Update center point c according to Eq.(3);
until Convergence criterion satisfied;
Group all the points visited in this iteration into cluster C;
if $\exists$ cluster C', c' is the center point of C', and $|c' - c| < \mu$ **do**
  Merge C' and C;
until All the points are visited;
for each point grouped into multiple clusters **do**
  Group it into the cluster with the biggest label;
end for
return $\{C^{(1)}, C^{(2)}, \ldots, C^{(n)}\}$ and $\{c^{(1)}, c^{(2)}, \ldots, c^{(n)}\}$.

The process of truth discovery is illustrated in Figure 2. Firstly, an object is determined by the system and K participants are responsible for observing the truth of this object. Secondly, the data collecting module assigns an ID to each participant to distinguish them and then collect all the observation collections for an object from these K participants. These source data are submitted to the data clustering model and clustered into n clusters by mean shift clustering algorithm. The output of data clustering module, as well as the input of truth value filter, are n center points of the n clusters and the ID of collections in each cluster. Truth value filter counts the number of collections for each cluster. The largest cluster is taken as the normal data and the remaining clusters are removed. The center point of the largest cluster is taken as the truth of the object. The aggregation and clustering processes are simultaneous, the center points, as well as the aggregations of each cluster, can be obtained at the end of clustering. Truth value filter only needs to filter out the aggregation of the largest.

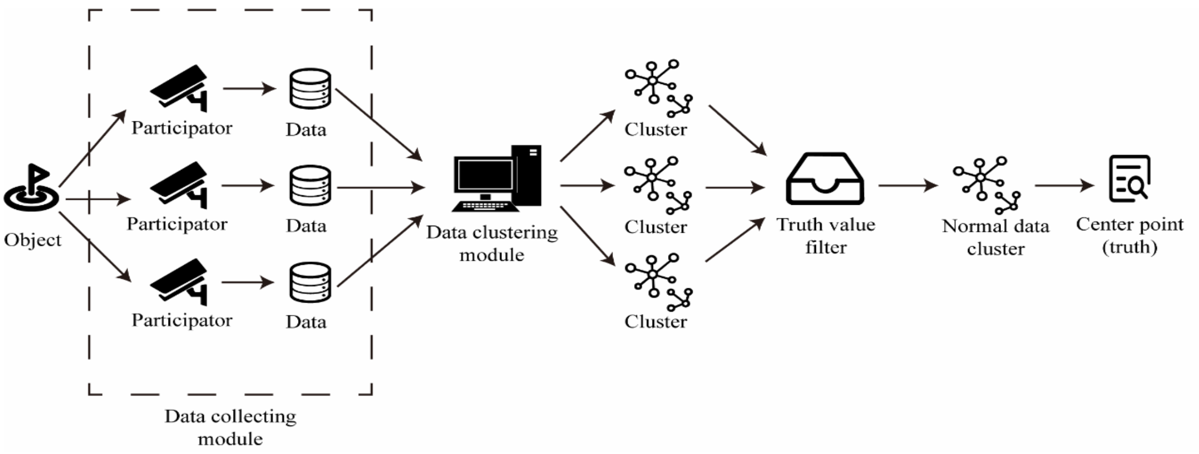

Figure 2. The process of proposed truth discovery scheme

4 Experiments

We first introduce the real-world dataset in Section 4.1 and the performance measures in Section 4.2. In Section 4.3, we present the baseline methods of truth discovery. The experimental results are shown in Section 4.4.

4.1 Robust Truth Discovery Scheme

In experiments, we use the Berkeley Earth temperature dataset. The dataset is divided into three categories: source data, intermediate data and output data. Source data contains heterogeneous types of properties. Specifically, there are several time-series temperature data with each site. In intermediate data category, data is collapsed so that there is only one time-series temperature per site. The data collection of adjusted is provided as the output data, which has been corrected bias.

In this dataset, each time-series temperature data is regarded as a collection from a participant and the value of each timestamp is regarded as an observation. Our target is to discover the truth temperature data of each site, that is, to merge the several time-series temperature data of each site and calculate one time-series truth value. Then the obtained truth value is compared with ground truth to calculate the prediction error.

We select the source data and output data to conduct our experiment. The source data is divided into two classes: with exceptional data and without exceptional data. For the first class, we take temperature data of 44,945 sites, and for the second class, we take 23,521 sites, as the test set. The corresponding output data is taken as the ground truth.

4.2 Performance Measures

We adopt three indicators to measure the proposed scheme and baselines:

Mean Absolute Error (MAE): This indicator can measure the mean absolute distance between the obtained truth value and the ground truth, which indicate how far apart they are. The define of MAE is

$$MAE = \frac{1}{l}\sum_{i=1}^{l} |(y_i - \hat{y}_i)| \tag{4}$$

where y_i is the ground truth, $\hat{y}_i$ is the obtained truth value, l is the number of test samples.

Mean Square Error (MSE): This indicator can measure the mean square distance between the obtained truth value and the ground truth. MSE is more sensitive to outliers. The define of MSE is

$$MSE = \frac{1}{l}\sum_{i=1}^{l} (y_i - \hat{y}_i)^2 \tag{5}$$

R-squared: This indicator reflects the fitting degree of the model. The closer it is to 1, the better the measured model fitting is. The define of R-squared is

$$R-squared = 1 - \frac{\sum_{i=1}^{l}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{l}(y_i - \overline{y}_i)^2} \tag{6}$$

where $\overline{y}$ is the mean value of all the test samples, l is the number of test samples.

4.3 Baseline Schemes

The performance of the proposed scheme is compared with two statistical methods and a widely used truth discovery method.

Mean: Mean method takes the mean value of all the time-series temperature data as the truth value.

Median: Median method takes the median of each timestamp, then puts them into a time-series as the truth value.

Conflict Resolution on Heterogeneous Data (CRH) framework: This scheme is widely used in truth discovery systems [8-9, 12-13], which is proposed to resolve the conflicts of heterogeneous data and discover the truth value by adjusting the weights of participants dynamically. In CRH, an initial estimate of "truth" is given and each participant is assigned a weight at the beginning, then the obtained truth value and weights are updated iteratively until the convergence criterion is satisfied. The weight of each participant is updated by

$$w_k = \log\left(\frac{\sum_{k=1}^{K}\sum_{m=1}^{M} d(x_m^k, x_m^*)}{\sum_{m=1}^{M} d(x_m^k, x_m^*)}\right) \tag{7}$$

where K is the number of participants, M is the number of properties of an object, $d(x_m^k, x_m^*)$ is a distance function denoting the distance between the source data x_m^k and the truth x_m^*. In our experiment, M is the number of timestamps, the distance function is $d(x_m^k, x_m^*) = (x_m^k - x_m^*)^2$.

The truth value of new iteration is updated by

$$x_m^* = \frac{\sum_{k=1}^{K} x_m^k \cdot w_k}{\sum_{k=1}^{K} w_k} \tag{8}$$

where w_k is the weight of the k-th participant, which is calculated in Eq.(7), where $k \in [1, K]$ and $m \in [1, M]$.

4.4 Experimental Results

We evaluated the proposed scheme and baselines with the measures in Section 4.2. The operations are implemented in Python with the Intel(R) Core(TM) i5-8400 CPU 2.80 GHz. We set the sliding window parameter to 0.8. We first conduct experiments under normal condition, in which the data is raw including normal data and outliers caused by non-malicious acts. The results are shown in Figure 3. Our scheme performs well with various participants.

In order to evaluate the robustness of our scheme and baselines, we conduct experiments with exceptional data. In this paper, exceptional data is divided into two types. One type is caused by non-malicious acts, such as equipment failure, extreme conditions and record mistake. Another is caused by malicious attacks. For the former, we extract outlier data from the dataset. The results are shown in Figure 4. For the latter, we take the method in [11] to simulate malicious attacks. Different levels of Gaussian noise are injected into normal data, which is taken as the second type of exceptional data. The results are shown in Figure 5.

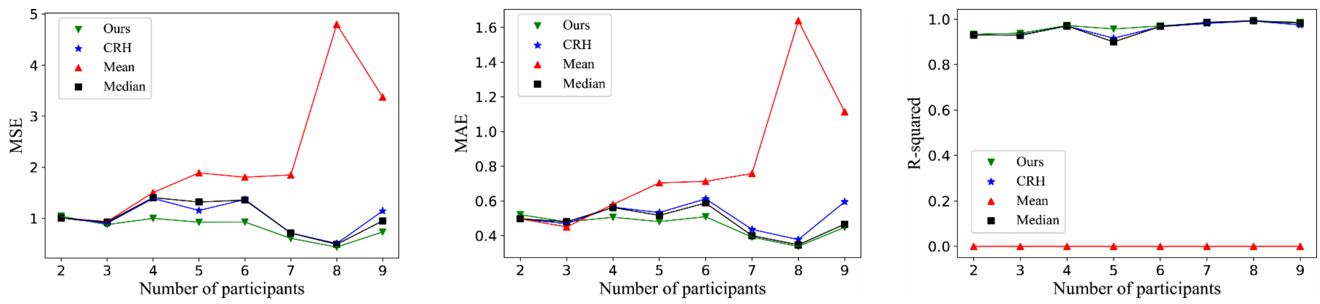

Figure 3. The performance comparisons between proposed scheme and baselines in normal condition

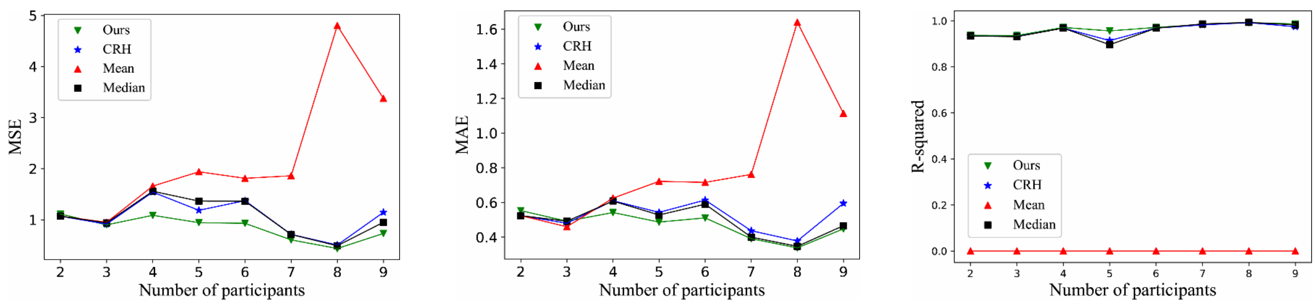

Figure 4. The performance comparisons between proposed scheme and baselines with outliers

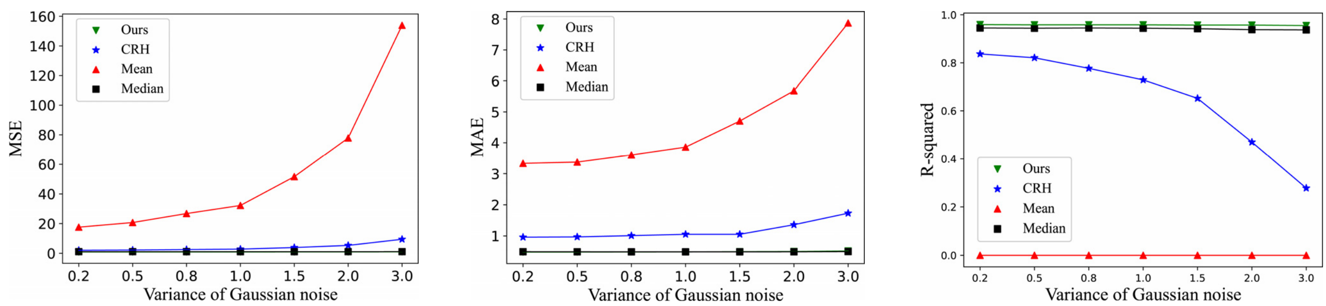

Figure 5. The performance comparisons between proposed scheme and baselines with different levels of Gaussian noise

The experimental results show that our scheme outperforms baselines under both of the two types of exceptional conditions, which demonstrate robustness is achieved in our scheme.

5 Conclusion

In this paper, we proposed a novel truth discovery scheme based on mean shift clustering algorithm, which divided the truth discovery process into two sub-tasks: removing exceptional data and aggregating normal data. To this end, we utilized a mean shift clustering algorithm to divide exceptional data and normal data into different clusters. After filtering, the normal data cluster was retained and the center of it was taken as the final "truth". We conducted a series of experiments on a real-world dataset to evaluate the performances of our scheme and baselines. The results showed that our scheme outperforms the baselines both in normal condition and in tackling outliers. In addition, our scheme performed well under malicious attacks and achieves robustness.

Acknowledgments

This work was partly supported by Natural Science Foundation of China under grant Nos.62072133 and 61662016, partly supported by Key projects of Guangxi Natural Science Foundation under grant No. 2018GXNSFDA281040, partly supported by The Science and Technology Major Project of Guangxi Province under grant No. AA18118039-3, partly supported by study abroad program for graduate student of Guilin University of Electronic Technology under grant No. GDYX2019008, partly supported by Innovation Project of Guangxi Graduate Education under Grant Nos. YCSW2020153, YCSW2019142 and 2018YJCX50, partly supported by GUET Excellent Graduate Thesis Program under grant No. 18YJPYSS12.

References

[1] W. Wu, N. Xiong, C. Wu, Improved clustering algorithm based on energy consumption in wireless sensor networks, *IET Networks*, Vol. 6, No. 3, pp. 47-53, May, 2017.

[2] Y. Liu, N. Xiong, Y. Zhao, A.-V. Vasilakos, J. Gao, Y. Jia, Multi-layer clustering routing algorithm for wireless vehicular sensor networks, *IET Communications*, Vol. 4, No. 7, pp. 810-816, April, 2010.

[3] T. Broens, A. Van Halteren, M. Van Sinderen, K. Wac, Towards an application framework for context-aware m-health applications, *International Journal of Internet Protocol Technology*, Vol. 2, No. 2, pp. 109-116, February, 2007.

[4] Y. Zheng, H. Duan, C. Wang, Learning the truth privately and confidently: Encrypted confidence-aware truth discovery in mobile crowdsensing, *IEEE Transactions on Information Forensics and Security*, Vol. 13, No. 10, pp. 2475-2489, October, 2018.

[5] Y. Du, Y.-E. Sun, H. Huang, L. Huang, H. Xu, Y. Bao, H. Guo, Bayesian co-clustering truth discovery for mobile crowd sensing systems, *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 2, pp. 1045-1057, February, 2020.

[6] Y. Wang, K. Wang, C. Miao, Truth Discovery against Strategic Sybil Attack in Crowdsourcing, *2020 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, pp. 95-104.

[7] P. Sun, Z. Wang, Y. Feng, L. Wu, Y. Li, H. Qi, Z. Wang, Towards personalized privacy-preserving incentive for truth discovery in crowdsourced binary-choice question answering, *2020 39th IEEE Conference on Computer Communications*, Toronto, Canada, 2020, pp. 1133-1142.

[8] C. Zhang, L. Zhu, C. Xu, K. Sharif, X. Liu, PPTDS: A privacy-preserving truth discovery scheme in crowd sensing systems, *Information Sciences*, Vol. 484, pp. 183-196, May, 2019.

[9] C. Zhang, L. Zhu, C. Xu, K. Sharif, X. Du, M. Guizani, LPTD: Achieving lightweight and privacy-preserving truth discovery in CIoT, *Future Generation Computer Systems*, Vol. 90, pp. 175-184, January, 2019.

[10] X. Yin, J. Han, P. S. Yu, Truth discovery with multiple conflicting information providers on the web, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 6, pp. 796-808, June, 2008.

[11] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, J, Han, Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation, *2014 ACM SIGMOD international conference on Management of data*, Snowbird, Utah, USA, 2014, pp. 1187-1198.

[12] C. Miao, W. Jiang, L. Su, Y. Li, S. Guo, Z. Qin, H. Xiao, J. Gao, K. Ren, Cloud-enabled privacy-preserving truth discovery in crowd sensing systems, *2015 13th ACM Conference on Embedded Networked Sensor Systems*, Seoul, South Korea, 2015, pp. 183-196.

[13] C. Miao, L. Su, W. Jiang, Y. Li, M. Tian, A lightweight privacy-preserving truth discovery framework for mobile crowd sensing systems, *2017 36th IEEE Conference on Computer Communications*, Atlanta, GA, USA, 2017, pp. 1-9.

[14] S.-Z. Selim, M.-A. Ismail, K-means-type algorithms: A generalized convergence theorem and characterization of local optimality, *IEEE Transactions on pattern analysis and machine intelligence*, Vol. PAMI-6, No. 1, pp. 81-87, January, 1984.

[15] Y. Cheng, Mean shift, mode seeking, and clustering, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 17, No. 8, pp. 790-799, August, 1995.

[16] J. Bleiholder, F. Naumann, Data fusion, *ACM computing surveys (CSUR)*, Vol. 41, No. 1, pp. 1-41, January, 2009.

[17] X.-L. Dong, F. Naumann, Data fusion: resolving data conflicts for integration, *Proceedings of the VLDB Endowment*, Vol. 2, No. 2, pp. 1654-1655, August, 2009.

[18] X.-L. Dong, D. Srivastava, Big data integration. *2013 29th IEEE international conference on data engineering (ICDE)*, Brisbane, Australia, 2013, pp. 1245-1248.

[19] J. Bleiholder, F. Naumann, Conflict handling strategies in an integrated information system, *Mathematisch-Naturwissenschaftliche Fakultät II*, Institut für Informatik, 2006.

[20] Z. Jiang, A decision-theoretic framework for numerical attribute value reconciliation, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 7, pp. 1153-1169, July, 2012.

[21] G. Xu, H. Li, S. Liu, M. Wen, R. Lu, Efficient and privacy-preserving truth discovery in mobile crowd sensing systems, *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 4, pp. 3854-3865, April, 2019.

[22] C. Zhang, C. Xu, L. Zhu, Y. Li, C. Zhang, H. Wu, An efficient and privacy-preserving truth discovery scheme in crowdsensing applications, *Computers & Security*, Vol. 97, Article No. 101848, October, 2020.

[23] W. Guo, B. Tondi, M. Barni, A Master Key Backdoor for Universal Impersonation Attack against DNN-based Face Verification, *Pattern Recognition Letters*, Vol. 144, pp. 61-67, April, 2021.

[24] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 24, No. 5, pp. 603-619, May, 2002.

[25] G. Gan, C. Ma, J. Wu, Data clustering: theory, algorithms, and applications, *Society for Industrial and Applied Mathematics*, 2020.

[26] D. Comaniciu, P. Meer, Mean shift analysis and applications, *1999 7th IEEE International Conference on Computer Vision*, Kerkyra, Greece, 1999, pp. 1197-1203.

[27] R.-T. Collins, Mean-shift blob tracking through scale space, *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, 2003, pp. II-234.

[28] Q. Shang, A Dynamic Resource Allocation Algorithm in Cloud Computing Based on Workflow and Resource Clustering, *Journal of Internet Technology*, Vol. 22, No. 2, pp. 403-411, March, 2021.

Biographies

Jingxue Chen received the B.S. degree in software engineering from Qufu Normal University, Jining, China, in 2018. She is currently working toward the master's degree with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China. Her research interests include privacy-preserving and truth discovery.

Jingkang Yang received the B.S. degree in Mechanical Design-Manufacture and Automation from Lanzhou Jiaotong University, Lanzhou, China, in 2018. He is currently working toward the master's degree with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China. His interests include privacy-preserving and machine learning.

Juan Huang received her M.E. degree in software engineering, University of Electronic Science and Technology of China, China, in 2013. Now she is researching assistant with Finance Section, Guilin University of Electronic Technology, Guilin, China. Her research interests include the software engineering and data privacy.

Yining Liu received the B.S. degree in applied mathematics from Information Engineering University, Zhengzhou, China, in 1995, the M.S. degree in computer software and theory from the Huazhong University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in mathematics from Hubei University, Wuhan, in 2007. He is currently a professor with School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, China. His research interests include the information security protocol and data privacy.