IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2018, accepted January 3, 2019, date of publication February 15, 2019, date of current version April 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897794

Estimate Information Fusion Weight of WSNs Nodes Based on Truth Discovery Optimization Method Among Conflicting Sources of Data

KEJIANG XIAO[1,2,3], ZHIWEN CHEN[3], AND CHUNHUA YANG[3]

[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2]School of Computer Science, Nankai University, Tianjin 300350, China
[3]School of Information Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Zhiwen Chen (zhiwen.chen@csu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61803391, in part by the China Postdoctoral Science Foundation under Grant 2018T110844 and Grant 2017M620355, and in part by the Major Program of the National Natural Science Foundation of China under Grant 61890932.

ABSTRACT In practical wireless sensor networks (WSNs)-based applications, often only a few nodes are active while most of the others are activated occasionally due to limited resource of the WSNs. Thus most sensors only provide a few observations and only a few sensors make many observations, which often cause long-tail issue to undermine information fusion performance. So we present a confidence-aware information fusion scheme named CAIF to solve such a problem. In particular, we first make a quantitative study on the long-tail data phenomenon and the relationship between node-target distance and node sensing capability, which can provide a guideline to improve node weight estimation error caused by long-tail data. Then, we propose a truth discovery-based method in WSNs via incorporating node-target distance into the truth discovery optimization solution framework to infer the sensor node's fusion weight. In order to adapt to the distribution characteristic of the WSNs, we propose a distributed implementation to estimate the sensor nodes' weights via confidence level and node-target distance to further improve the fusion performance. Besides, the iterative process shown in CAIF converges to a stationary point of the optimization problem and its time complexity is linear with respect to the total number of observations. Finally, we conduct extensive experiments on real data to validate and evaluate CAIF. The experimental results demonstrate the superior performance of our method over existing solutions in terms of root-mean-square error and accuracy.

INDEX TERMS Wireless sensor networks, information fusion, confidence level, weight optimization.

I. INTRODUCTION

Wireless sensor networks (WSNs) are gaining popularity in diverse fields because they can support multiple applications [1]–[3]. Different from traditional networks, a WSN enjoys its own layout and resource constraints, such as limited energy, short communication range, limited process-ing and so on. Besides, sensor nodes with limited computing resources are also small and inexpensive compared with traditional sensors nodes. What is more, such constraints depends on specific applications and monitored environments. Such environments play an important role in determining the network size, network topology, and deployment strategy. To measure the same physical quantity, each sensor node is characterized by its own performance (weight) and power usage. WSNs are usually deeply integrated with dynamic physical environments where uncertainties are ubiquitous [4], which result in conflicting observation for the same detection target among the sensor nodes. Thus it is important to distinguish the most trustworthy information by utilizing information fusion techniques from multiple sensor nodes of conflicting information in WSNs. This is a non-trivial problem due to three major challenges.

The associate editor coordinating the review of this manuscript and approving it for publication was Jin-Liang Wang.

A. WEIGHT ESTIMATION

Information conflicts have been researched in database area for many years [5] and some approaches are also proposed to process such conflicts for information fusion. Among them, majority voting method [6] is usually utilized to solve information conflicts for categorical data. This method treats information owning the highest occurrences as the truth

information and enjoys some robust-ness also in the case of unequal quality sensors. While mean or median [7] can also be regarded as the truth information and for continuous values. The shortcoming of such methods mentioned above is that all the sources (sensor nodes) are equally reliable and have the same weight. In complicated world information fusion, fusion weight estimation plays is important to discover the correct information from conflict-ing data, especially when there exist sensor nodes providing low quality information, such as faulty sensors that keep producing data. However, the user does not know which sensor node is more reliable and which piece of information is correct in advance.

B. DISTRIBUTED ESTIMATION [2]

WSN has distributed characteristic and often consists of many sensors deployed in the practical surveillance area. Each sensor node makes a local observation about underlying physical phenomenon, quantizes its observations, and transfers the data back to a fusion center/base station (fusion node). The goal of the sensor network design is to estimate such physical phenomenon as accurately as possible under the limited network resource. Thus it is necessary to design distributed estimation framework to overcome this challenge.

C. LONG-TAIL DATA [8]

The long-tail of multi-source data means that most monitoring targets obtain a few claims from a small number of sensors and only a few targets obtain many claims from a lot of sensors. In practical WSNs-based applications, there often only a few sensor nodes are active while most of others are activated occasionally, which often causes some nodes with very few observations. The number of observations made by the sensor nodes typically exhibits long-tail problem, that is: 1) most of the sensor nodes only provide information about one or two items; 2) there are only a few sensor nodes that can make many observations. Long-tail phenomena are ubiquitous in real world applications, which bring obstacles to the task of information trustworth-iness estimation in the process of information fusion.

While in WSNs-based information fusion, the node fusion weight plays an important role in improving fusion performance, many truth discovery methods was proposed to estimate such weight and infer truth without any supervision [9]. In particular, the observations about the same targets can be acquired from a variety of sources. Under such circumstance, information conflicts often are generated inevitably. Thus how to identify the truths among conflicting data from multiple sensor nodes becomes an important problem. Here the truth information of a monitoring target is the most trustworthy one from all possible candidate observations sensed by sensors. To distinguish such truths, weighted aggregation of the multi-source information is conducted based on the estimated node weight. As for truth discovery problems, if a set of assertions claimed by multiple sources (sensor nodes) are given, each claimed value is labeled as true or false and the reliability of each sensor node is computed. Node weight estimation is the most important characteristic of truth discovery. The current main research about truth discovery focuses on iteratively computing and updating the trustworthiness of a source sensor as a function of the belief in its claims, and then the belief score of each claim as a function of the trustworthiness of the source sensors asserting it. Thus the sensor node weight estimation and truth discovery steps are closely linked: the sensor nodes with true information more often will be allocated higher weight, and the informa-tion given by sensor nodes with higher weight will be considered as truth information.

Thus for the mentioned above challenges, a truth estimation problem is considered in the paper, and a truth discovery base Confidence Aware Information Fusion (CAIF) approach is presented here to estimate such truths. The long-tail data issue is first analyzed in WSNs and the relationship between node-target distance and sensing performance is also explored. These can provide guideline to overcome the long-tail issue. Then based on such discovery, the sensor node-target distance is incorporated into the truth discovery framework to estimate the sensor nodes' information fusion weight which can improve the fusion performance. In order to overcome physical environments' uncertainties, we combine both node-target distance and confidence level of observation error to deal with long-tail data and further improve the information fusion performance. What is more, due to constrained-resource of the WSNs, we design a distributed node weight estimation scheme via considering both node-target distance and confidence level, which make CAIF be suitable to the distributed characteristic of the WSNs. The contributions of this paper are summarized as follows.

1)We propose a truth discovery based method to solve long-tail data phenomenon in WSNs, which incorporate sensor node-target distance into the truth discovery algorithm optimization solution framework to infer the sensor node's information fusion weight.

2)We propose a distributed implementation in WSNs to estimation the sensor nodes' weights via considering both confidence level and node-target distance. In particular, we plug the node-target distance into squared loss of the errors to compute its confidence interval. The optimized weight is obtained via minimizing the confidence interval's upper bound, which can process the long-tail data better. Besides, the iterative process shown in the proposed method converges to a stationary point of the optimization problem and its time complexity is linear with respect to the total number of observations.

The rest of this paper is organized as follows. Related work is reviewed in Section II. We give an overview about CAIF in Section III. Quantitative study for long-tail data is introduced in Section IV. Section V presents an confidence-aware information fusion approach named CAIF. In Section VI, we evaluate the performance of CAIF and analyze the results. Finally, we give conclusive remarks in Section VII.

II. RELATED WORK

In practical monitoring scenarios, many WSNs based applications utilize information fusion coming from various sensor nodes to improve surveillance performance. A survey of the current work [10] has been performed for information fusion and event detection in WSNs. Zhao *et al.* [11] constructed an efficient framework that can exploit multiple available information to analyze and improve the performance of the WSN-based indoor localization system. Multiple pieces of information are fused to derive the relationship to the target position and eliminate the error. The information fusion in WSNs is divided into three levels: data level fusion, feature level fusion and decision level fusion. A mechanical fault diagnosis method based on multi-level hierarchical information fusion in WSNs is proposed in [12] to meet the real-time transmission of a large number of vibration signals when applied to mechanical fault diagnosis. However, these works assume all the sensor nodes are equally reliable, and thus the votes from different sensor nodes are uniformly weighted during fusion process, which can not reflect the real monitoring environ-ments.

While motivated by the importance but lack of knowledge in nodes weight, we found that many truth discovery approaches [13] have been proposed to estimate source weight without any supervision in the field knowledge discovery. In particular, the source weight can only be inferred based on the data. The source weight estimation and truth finding steps are tightly combined through the following principle: The sources that provide true information more often will be assigned higher weight degrees, and the information that is supported by reliable sources will be regarded as truths. For example, Wang [14] proposed a new model that solves a bi-dimensional estimation problem to jointly estimate the correctness and theme relevance of claims as well as the source reliability and theme awareness of sources. A multi-dimensional estimation problem is solved in [15] to jointly estimate the correctness and mood neutrality of claims as well as the reliability and mood sensitivity of sources. A new time sensitive truth discovery scheme is proposed in [16]. In this method, the source responsiveness and the claim lifespan are incorporated into a analytical framework, by which a maximum likelihood estimation problem is solved to determine both the truth information and source weight. Xu *et al.* [17] proposed an efficient and privacy-preserving truth discovery scheme in crowd sensing systems. Specifically, it utilized the additive homomorphic privacy-preserving data aggregation and super-increasing sequence techniques to achieve both high performance and strong privacy protection. In order to take advantage of a joint inference on data with heterogeneous types for truth discovery, a conflict resolution scheme [18], [19] is proposed to resolve conflicts among multiple sources of heterogeneous data types. It modeled the problem using an optimization framework where truths and source reliability are defined as two sets of unknown variables. The objective is to minimize the overall weighted deviation between the truths and multi-source observations where each source is weighted by its reliability. The works mentioned above are not applied in WSNs and also do not consider the limitations of WSNs. These works have been applied in the database community for years. The approaches mentioned are mainly centralized algorithms and also do not consider long-tail problem.

In recent yeas, there are also some truth discovery based methods. For example, Li *et al.* [20] focus on the probabilistic model and formulate it as a geometric optimization problem. Based on a sampling technique and a few other ideas, the first $(1 + \epsilon)$-approximation solution is achiev-ed. What is more, a novel distributed truth discovery frame-work is proposed in [21], which can effectively and efficiently aggregate conflicting data stored across distributed servers, with the differences among the objects as well as the importance level of each server being considered. Because confidence interval contains richer information, Liu *et al.* [22] propose a novel approach called TruthDiscover to determine the most trustworthy object in Linked Data with a scale-free property. More specifically, TruthDiscover consists of two core components: Priori Belief Estimation for smoothing the trustworthiness of sources by leveraging the topological properties of the Source Belief Graph, and Truth Computation for inferencing the trust-worthiness of source and trust value of an object. Besides Zheng *et al.* [23] propose a new system architecture enabling encrypted truth discovery in mobile crowd-sensing, which focus on general and realistic mobile crowd-sensing scenarios with varying levels of user participation, and the security design is built on the confidence-aware truth discovery approach for its state-of-the-art accuracy in such scenarios. Along the whole workflow, the sensory data and reliability degrees of users, as well as the inferred truths of the requester, are kept private. But these works also do not consider the limitations of WSNs.

Long tail data phenomenon is ubiquitous in WSN-based applications. For example, Iyer *et al.* [24] shown that in the wireless sensor networks the efficiency of cross-layer QoS performance of routing algorithms with MAC losses has a long tail, because it is similarly observed in Power Law. Besides, in practical applications, in order to obtain stringent accuracy requirements for target monitoring while maximizing network lifetime in WSN-based applications, only a few sensors are incessantly active while most of others are activated occasionally, which also lead to long tail issue. A typical example is sleep scheduling mechanism [25] in WSNs, which brings the result that most sensors only provide a few observations and only a few sensors make many observations [26], [27], which often causes some nodes with very few observations. As for the long-tail phenomenon, a confidence-aware truth discovery method is proposed in [8] to automatically estimate truths from conflicting data with long-tail issue. The proposed method not only estimates source reliability, but also considers the confidence interval of the estimation, so that it can effectively reflect real source reliability for sources with various levels of participation. Similarly to [8], Xiao *et al.* [28] also proposed a novel truth

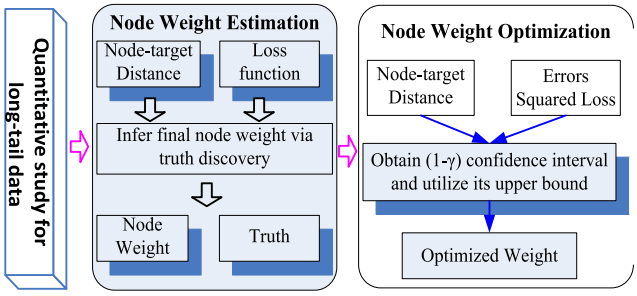

FIGURE 1. An overview of CAIF scheme.

discovery method to construct confidence interval estimates as well as identify truths. But these works are not also applied in WSNs and do not consider the distributed characteristic of the WSNs.

III. AN OVERVIEW OF CAIF SCHEME

In the paper, we present a confidence-aware information fusion approach named CAIF. We first make quantitative study on long-tail data to provide guideline for improving fusion performance. Then in order to solve long-tail issue and estimate fusion weight, node-target distance and confidence level of observation error are incorporated into truth discovery framework. Besides, we also propose cent-ralized and distributed implementation respectively in WSN to estimate the nodes' weights. In particular, an overview include three parts described as follows as shown in Fig.1.

A. QUANTITATIVE STUDY

The long-tail phenomenon is first studied by utilizing real data and such phenomenon is serious. This will undermine the fusion performance [8]. Through the research on relationship between node capability and node-target distance, we find that node-target distance is inversely proportional to node's capability, which can overcome long-tail problem in a certain when utilizing such weight for information fusion.

B. NODE WEIGHT ESTIMATION

We provide a node-target distance based truth discovery algorithm to estimate node weight in a centralized way for fusion information in WSNs, which is computed by minimizing overall weighted deviation between the truths, the sensor nodes' observations and node-target distance to solve the long-tail data problem.

C. NODE WEIGHT OPTIMIZATION

We provide a distributed node weight estimation method to optimize about our proposed method in *WSN*. In particular, we incorporate the node-target distance into the confidence interval of observations' error. The optimized weight is obtained via minimizing such confidence interval's upper bound to process the long-tail data.

IV. QUANTITATIVE STUDY FOR LONG-TAIL DATA

In this section, we present real world information fusion applications where the long-tail phenomenon can be observed. Although the long-tail phenomenon is not rare in information fusion tasks, it does not receive enough attention yet. Next, we first make quantitative study on long-tail phenomenon, and then the relationship among detection capability and sensor-target distance is researched, which will provide guideline to improve information fusion performance via relieving long-tail issue.

A. LONG-TAIL DATA PHENOMENON

In WSN based applications, for the same object or event, different sensors nodes may report differently due to many factors, such as the quality of the sensors and ubiquitous uncertainties in practical monitoring environments. Truth discovery techniques can be useful for information fusion to improve the quality of sensor data integration by inferring the sensor nodes' quality. In many practical WSN-based applications, in order to achieve energy efficiency, only a few sensor nodes are active while most of others are activa-ted occasionally, which causes the long-tail data phenome-non: most sensor nodes provide few observations and only a small proportion of the sensor nodes can provide a large number of observations. Note that sensor nodes usually include communication module, sensing module and calculation module. The sensor nodes' energy usage is mainly concentrated on the CPU and the wireless transceiver. Under the normal circumstances, wireless communication consumes relatively more energy [29]. While the member sensor nodes which join in the information process will need to sense and communicate with a Fusion Center/ Base Station. Thus keeping only a fraction of all the nodes active can reduce communication and sensing energy.

A representative example of long-tail phenomenon is the construction of indoor floor plans [30]. Indoor floorplans are usually needed by indoor localization and are typically in the form of a building blueprint for a specific indoor environment. If an Indoor floorplan is given, we can define the indoor environment. This research topic has recently drawn a growing interest since it potentially can support a wide range of location-based applications. The goal of indoor floor plan problem is to develop an automatic floor plan construction system that can infer the information about the building Indoor floorplan from the readings of inertial sensors (e.g., accelerometer, gyroscope, and compass). Here we are interested in one specific task of floor plan construction, i.e., to estimate the distance between two indoor points (e.g., a hallway segment, which is a straight or curved path in a building that is adjacent to rooms. If one person walks through the hallway, the person will pass the rooms along the hallway sequentially, which means how rooms are arranged along the hallway.). The estimated distances given by different sensor nodes are inevitably different due to the varieties in their walking patterns and the quality of sensor nodes.

Because long-tail phenomenon is a general scenario, some distributions can be used to describe such phenom-enon. Thus the technical detail and indication of relevance about long-tail data can be expressed by following distributions: 1) The histograms in terms of the number of sensor observations

can be fit into an exponential distribution, which can be expressed by formula $\varphi_{power-law} = \varphi_{\min}(1-\kappa)^{-1/(\delta-1)}$, if given a source of uniform random numbers κ in the range $0< \kappa <1$, the formula for generating random numbers $\varphi_{power-law}$ can be obtained from continuous distributions; δ is scaling parameter of the discrete power law. 2) The data about the number of sensors and number of the observations can be fit into power law distribution, which can be expressed by formula $\varphi_{\exp onential} = \varphi_{\min} - \ln(1-\kappa)/\Xi$, the random numbers $\varphi_{\exp onential}$ can be obtained from continuous distribu-tion when the uniform random numbers κ are in the range $0 < \kappa < 1$, Ξ is the exponential parameter, $\varphi_{\min}$ is the minimum value of $\varphi_{\exp onential}$. 3) The data about the number of sensors and the number of observations can be fit into log-normal distribution, which can be expressed by $\varphi^1_{\log-normal} = \exp(\Gamma \sin\vartheta)$, $\varphi^2_{\log-normal} = \exp(\Gamma\cos\vartheta)$, where $\Gamma = \sqrt{2\sigma^2\ln(1-\kappa_1)}$, $\vartheta = 2\pi\kappa_2$, κ_1 and κ_2 are the the uniform random numbers, σ^2 is varia-nce. From the formula, we can find that there is no simple closed-form expression to generate a single random. The formula can generate two independent log-normally distributed random numbers $\varphi^1_{\log-normal}$ and $\varphi^2_{\log-normal}$. In order to demonstrate the long-tail phenomenon clearer, we further fit Indoor Floorplan into power law function, a typical long-tail distribution. Experimental results show that most sensors only provide a few claims and only a few sensors make many claims, and the fitting curves closely match the observations, which is a strong evidence of long-tail problem in WSNs.

B. SENSOR-TARGET DISTANCE AND DETECTION CAPABILITY

Wisconsin SensIT experiment data [31] is utilized to perform vehicle detection via information fusion. For the data set extraction, a k-Nearest Neighbor classifier was used to label each 0.75s data segment from each separate node as a detection or non-detection,which is coming from each separate sensor node. Two types of vehicles (aav and dw) are utilized in the experiment and two features are utilized for such detection classification which are the distance between the vehicle and sensor node and the acoustic signal energy for that given time respectively. The sensor and sensor cluster readings are classified into the vehicle types. The trace data is the vehicle runs including aav3, dw3 aav4, dw4; aav5, dw5; aav6, dw6; aav7, dw7; aav8, dw8; aav9, dw9. Among them, aav3 and dw3 were used for training, the rest are used for detection classification. The events in these runs were identified manually and we can obtain the event labeling for each run. According to experimental analysis, we can make conclusion that the sensor's sensing performance will become weak when the distance increases [31]. According to the relationship between sensing node- target distance and measurement quality, we can find that node-target distance can reflect the real weight of the sensor nodes at a certain, although the relationship of sensing node-target distance on measurement quality has been recently analyzed in detection problems. Thus we consider incorporate node-target distance into truth discovery framework to estimate node weight and improve information fusion performance which is undermined by long-tail issue.

Therefore, according to the section A and B mentioned above, we can make conclusions as follows: 1) Long-tail data phenomenon is ubiquitous in WSN-based applications which will undermine fusion performance for lower precise of node weight estimation. 2) sensor nodes' sensing performance will become weak when the distance increases, which can reflect the real weight of the sensor nodes at a certain. Thus we consider incorporate node-target distance into truth discovery framework to estimate node weight and improve information fusion performance which is undermined by long-tail issue.

V. CONFIDENCE-AWARE INFORMATION FUSION

In this section, we first formalize the problem and propose a centralized and distributed confidence-aware information fusion approach respectively. Then its convergence and time complexity are analyzed.

A. PROBLEM FORMULATION

We assume there exist N sensor nodes in the WSN considered. Each node is indexed with n, which denotes the n-th sensor node in the WSNs. At the same time, each node has M kinds of feature observations which are indexed with m. The m-th observation feature obtained by the sensor node n is represented as $o_m^{(n)}$. The truth for the observation feature m is denoted as $o_m^{(*)}$, the weight for each sensor node n is w_n. Table I summarizes the important notations used in this paper. With these notations, the problem can be formalized as follows.

Suppose there are observations set $\{\vec{o}_n\}_{n=1}^N$, where the vector $\vec{o}_n = [o_1, o_2, \ldots, o_M]$ represents different observations of the n^{th} node. The expected output are the truths $o^{(*)} = \{o_1^{(*)}, o_2^{(*)}, o_3^{(*)}, \ldots, o_m^{(*)}, \ldots, o_{M\times N}^{(*)}\}$ and nodes' weights $\nabla = \{w_1^{(*)}, w_2^{(*)}, w_3^{(*)}, \ldots, w_n^{(*)}, \ldots, w_N^{(*)}\}$. The truth of the observation feature is vector about the characteristic of monitoring target (eg., highest temperature of every day) which are extracted from the observations sensed by sensor nodes. Thus the observation feature is used to describe the object; and a sensor node describes the place where information about objects' properties can be collected. The intuitions behind the proposed method are that the truths should be close to the observations given by the reliable sensor nodes, and the nodes with smaller node-target distance has bigger weights. Based on these intuitions, we formulate the problem by

$$\begin{aligned}\min_{o^{(*)},\nabla} f(o^{(*)},\nabla) &= \alpha \times \sum_{n=1}^{N}\left(w_n \sum_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n)})\right) \\ &\quad +\beta \times \sum_{n=1}^{N} w_n d_n \\ s.t.\ \sum_{n=1}^{N} \exp(-w_n) &= 1, \nabla \in D \end{aligned} \tag{1}$$

where α and β are hyper parameters that balance between the two terms in the objective function. w_n is the weight of

the node n. d_n is the distance between the sensor node n and the monitoring target.

We are trying to search for the values for two sets of unknown variables $o^{(*)} = \{o_1^{(*)}, \ldots, o_m^{(*)}, \ldots, o_M^{(*)}\}$ and $\nabla = \{w_1^{(*)}, \ldots, w_n^{(*)}, \ldots, w_N^{(*)}\}$, which correspond to the collection of truths and source weights respectively, by minimizing the objective function $f(o^{(*)}, \nabla)$. There are two types of functions that need to be plugged into this framework. 1) *Loss function*. Λ_{m} refers to a loss function defined based on the distance between the truth and observation.This function measures the distance between the observation $o_m^{(n)}$ and the truth $o_m^{(*)}$. This loss function should output a high value when the observation deviates from the truth and a low value when the observation is close to the truth. 2) *Regularization function.* $\sum_{n=1}^{N} \exp(-w_n)$ reflects the distributions of nodes' weights∇. To constrain the nodes' weights into a certain range, we need to specify the regularization of ∇ and the domain $D(\nabla \in D)$. If each sensor node's weight w_n is unconstrained, then the optimization problem is unbounded. That is because we can simply take w_n to be $-\infty$.

Algorithm 1 Centralized Confidence-aware Node Weight Estimation

Input: Observations from N sensor nodes$\{\vec{o}_1,\vec{o}_2,\vec{o}_3,\ldots,\vec{o}_m\ldots,\vec{o}_{N\times M}\}$
Sensor node-target distance d_n obtained by location algorithm
Output: Truths $o^{(*)} = \{o_m^{(*)}\}_{m=1}^{M}$,nodes' weights $\nabla = \{w_1, \ldots, w_n, \ldots, w_N\}$

```
Initialize the truths o^(*)
repeat
    Compute and update nodes' weights ∇ according to Eq. (15)
    for n ← 1 to N do
        for m ← 1 to M do
            Update the truth of the m-th observation o_m^(*) according to Eq.(3) based on current estimation of nodes' weights.
        end for
    end for
until Convergence criterion is satisfied.
return o^(*) and ∇
```

B. PROPOSED SOLUTION

The benefits of adopting this optimization-based formulation are: 1) It encodes the idea of truth discovery. 2) It allows us to incorporate constraints and prior knowledge about sensor nodes weights. 3) In the following, we will show that this formulation can be linked with MAP (Maximum a posteriori) estimation which gives an efficient incremental solution. In this optimization problem as shown in Eq.(1), two sets of variables are involved, sensor nodes' weights ∇ and aggregated results $o^{(*)}$. To solve this problem, we adopt coordinate descent, in which one set of variables are fixed in order to solve for the other set of variables as shown in the algorithm 1.

TABLE 1. Parameter meaning.

Symbol	Parameter meaning
N	Number of the sensor nodes
n	The n-th sensor node
d_n	The distance between the monitoring target and the node n
ε	The error of the weighted combination
m	The m-th observation feature obtained by each node
$o^{(*)}$	Set of all the properties observations' truths
w_n	The weight of the sensor node n
Λ_m	Loss function
λ	a Lagrange multiplier
1-γ	confidence interval
∇	Set of all the sensor nodes' weights
$o_m^{(n)}$	The m-th target's properties observations obtained by the sensor nodes
$o_m^{(*)}$	The m-th truth of target's properties observations obtained by the sensor nodes
α, β	Weight balance parameter

1) CASE 1: TRUTH CALCULATION

In this case, node weights ∇ are fixed, we update the truth by minimizing the objective function as follows.

$$o_m^{(*)} \leftarrow \arg\min_{o} f(o^{(*)}, \nabla) \tag{2}$$

As shown in Eq. (2), the truth computation step (Eq. (2)) depends on the loss function. We respect the characteristics of each feature observation and utilize different loss functions to describe different notions of deviation from the truths. Thus, truth computation will differ among various data types.When the nodes' weights ∇ are fixed, we take the derivative of Eq. (2) with respect to $o_m^{(*)}$ so as to infer the truth. Then we get the following result.

$$o_m^{(*)} = \sum\nolimits_{n=1}^{N} w_n \times o_m^{(n)} / \sum\nolimits_{n=1}^{N} w_n \tag{3}$$

2) CASE 2: WEIGHT CALCULATION

In the process of weight calculation, we first fix all sets of truths, and then calculate the sensor nodes' weights based on the difference between the truths and the observations sensed by the sensor nodes.

$$\nabla \leftarrow \arg\min_{\nabla} f(o^{(*)}, \nabla), \quad \text{s.t.} \sum\nolimits_{n=1}^{N} \exp(-w_n) = 1 \tag{4}$$

According to Lagrange multipliers method, the optimization problem can be solved. The Lagrangian of Eq. (1) is computed as follows.

$$L(\{w_n\}_{n=1}^{N}, \lambda) = \alpha \times \sum\nolimits_{n=1}^{N} w_n \sum\nolimits_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n)}) + \beta \times \sum\nolimits_{n=1}^{N} (w_n \times d_n) + \lambda(\sum\nolimits_{n=1}^{N} \exp(-w_n) - 1) \tag{5}$$

where λ is a Lagrange multiplier. Let the partial derivative of Lagrangian with respect to w_n be 0, we get following formula.

$$\alpha \times \sum_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n)}) + \beta \times d_n = \lambda \exp(-w_n) \quad (6)$$

From the constraint that $\sum_{n=1}^{N} \exp(-w_n) = 1$, we can derive that

$$\lambda = \alpha \times \sum_{n'=1}^{N} \sum_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n')}) + \beta \times \sum_{n'=1}^{N} d_{n'} \quad (7)$$

We can then derive the update rule for each node's weight by plugging Eq. (7) into Eq. (6) as shown in Eq. (8).

$$w_n = -\log\left(\frac{\alpha \times \sum_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n)}) + \beta \times d_n}{\alpha \times \sum_{n'=1}^{N} \sum_{m=1}^{M} d_m(o_m^{(*)}, o_m^{(n')}) + \beta \times \sum_{n'=1}^{N} d_{n'}}\right) \quad (8)$$

where n denotes the index of a sensor node. This update rule shows that a node's weight is higher when its observations are more often close to the truths.

Algorithm 2 Distributed Confidence-Aware Weight Estimation

Input: Observations from N sensor nodes$\{\vec{o}_1, \vec{o}_2, \vec{o}_3, \ldots, \vec{o}_m \ldots, \vec{o}_{N\times M}\}$
Sensor node-target distance d_n obtained location algorithm

Output: Truths$o^{(*)} = \{o_m^{(*)}\}_{m=1}^{M}$, nodes' weights $\nabla = \{w_1, \ldots, w_n, \ldots, w_N\}$

```
Initialize the truths o^(*)
repeat
    Each sensor node compute its weight w_n via its history observations and initialization truth as shown in Eq. (15).
    Each node broadcasts a packet to its neighbors and the packet contain each node's residual energy value and ε.
    for n ← 1 to N do
        for m ← 1 to M do
            Each node compute Γ_n = w_n × o_m^(n) via its history observa- tions and weight
            Update the truth of the m-th observation o_m^(*) via Eq. (3) based on the current estimation of nodes' weights.
        end for
    end for
until Convergence criterion is satisfied.
return o^(*) and ∇
```

However, because there exists ubiquitous noise, so the sensor node-target distance d_n can not also accurately reflect the real weight of the sensor nodes. As we have observed the long-tail phenomenon in Section III, most of the sensor nodes have very few observations. We assume all the sensor nodes obtained their observations independently. Errors, which are differences between the observations and the truths, may occur for every sensor node. The variance of the error distribution reflects the reliability degree of this node: if a node is unreliable, the errors it makes occur frequently and have a wide spectrum in general, so the variance of the error distribution is big. Gaussian distribution is utilized to describe the errors, which is widely adopted in many fields such as crowd sensing [8]. For each sensor node, its error follows a Gaussian distribution with mean 0 and variance σ^2. Since we have the sensor node independence assumption, the errors that nodes make are independent too. We can then compute the distribution for the error of the weighted combination in Eq. (3) as follows.

$$\varepsilon \sim N(0, \sum_{n=1}^{N} w_n^2 \sigma_n^2 \Big/ \sum_{n=1}^{N} w_n^2) \quad (9)$$

The shape of the distribution is determined by variance for a Gaussian distribution. If the variance is small, then the distribution has a sharp and high central peak at the mean, which indicates a high probability that errors are close to 0. Thus we want the variance of the ε to be as small as possible. Usually the theoretical σ_n^2 is unknown for each sensor node. Inspired by sample variance and sensor node-target distance d_n, the following estimator can be used to estimate the real variance σ_n^2 by Eq. (10).

$$\hat{\sigma}_n^2 = \frac{1}{d_n} \times \frac{\sum_{m\in N_n} (o_m^n - o_m^{*(0)})^2}{|N_n|} \quad (10)$$

where $o_m^{*(0)}$ is initial truth (such as the mean, median or mode of the observations), $|N_n|$ is the number of observations made by the sensor node n. Another interpretation of Eq. (10) is that σ_n^2 represents the mean of the squared loss of the errors which are made by the sensor node n.

Then, we adopt $(1-\gamma)$ confidence interval for σ_n^2, where γ, also known as significant level, is usually a small number such as 0.05. This also applies to missing data. As we illustrate above, the difference between o_m^n and $o_m^{*(0)}$ follows a Gaussian distribution $\mathbf{N}(0, \sigma_n^2)$. Since the sum of square of standard Gaussian distribution has chi-squared distribution, we have:

$$\frac{\sum_{m\in N_n} (o_m^n - o_m^{*(0)})^2}{\sigma_n^2} = \frac{d_n \times |N_n| \times \hat{\sigma}_n^2}{\sigma_n} \sim \chi^2(d_n \times |N_n|) \quad (11)$$

Thus, we have the following formula.

$$P\left(\chi^2_{(1-\gamma/2), d_n\times|N_n|} < \frac{d_n \times |N_n| \times \hat{\sigma}_n^2}{\sigma_n^2} < \chi^2_{\gamma/2, d_n\times|N_n|}\right) = 1-\gamma \quad (12)$$

which gives the $(1-\gamma)$ confidence interval of σ_n^2 as follows:

$$\left[\frac{\sum_{m\in N_n} (o_m^n - o_m^{*(0)})^2}{\chi^2_{(1-\gamma/2), d_n\times|N_n|}}, \quad \frac{\sum_{m\in N_n} (o_m^n - o_m^{*(0)})^2}{\chi^2_{\gamma/2, d_n\times|N_n|}}\right] \quad (13)$$

The confidence level is more informative according to Eq. (13). Although two sensor nodes with different numbers of observations, which is caused by missing data or sleep scheduling mechanism [25], may have the same $\hat{\sigma}_s^2$, the confidence interval of σ_n^2 for these two nodes can be significantly different.

The upper bound of the confidence interval is a biased estimator on σ_n^2, but the bias is big only on nodes with few observations. As the number of claims from a node increases, the bias drops. We can substitute the unknown variance σ_n^2 in Eq.(10) by this upper bound and rewrite the optimization problem Eq. (4) as shown in Eq. (14).

$$\begin{aligned} &\min_{w_n} \sum_{n=1}^{N} w_n \sum_{m \in N_n} (o_m^n - o_m^{*(0)})^2 \Big/ \chi^2_{\gamma/2, d_n \times |N_n|} \\ &s.t. \sum_{n=1}^{N} \exp(-w_n) = 1, \nabla \in D \end{aligned} \tag{14}$$

This optimization problem is convex, so the global minimum guarantees that we can find the best weight assignment under this scenario. The closed form solution is given as follows.

$$w_n \propto \chi^2_{\gamma/2, d_n \times |N_n|} \Big/ \sum_{m \in N_n} (o_m^n - o_m^{*(0)})^2 \tag{15}$$

According to Eq. (15), we can find that the node's weight is inversely proportional to the upper bound of the $(1-\gamma)$ confidence interval for its real variance. Bedsides the chi-squared probability value will dominate the weight when a node only provides very few observation. If a sensor node can provide sufficient observations, the chi-squared' probability value is close to $|N_n|$ and has small bias on the estimator. Thus, the proposed method automatically adjusts weights for nodes with different numbers of observations.

Therefore, truth discovery based method in our paper can better solve long-tail data phenomenon in WSNs, that is because it incorporate sensor node-target distance and confidence level into the truth discovery algorithm optimization solution framework to infer the sensor node's information fusion weight.While even if a well defined probabilistic framework would be much better at representing uncertainty and fusing uncertain information sources with differing confidences, it may lead to an inappropriate weight assignment for most of sensor nodes with less claims, and further cause inaccurate truth computation because of ubiquitous long-tail problem. Although Dempster-Shafer evidential methods has a better theoretical foundation, it also can not better solve long-tail problem and obtain inaccurate fusion weight.

C. DISTRIBUTED IMPLEMENTATION IN WSN

At the beginning, each feature observation m initializes its truth $o_m^{(*)}$. Then, the computing process is based on the two cases of updates. *Case one:* each node's weight w_n can be computed via its history observations and initialization truth at the local node as shown in Eq. (15). *Case two*: each node compute $\Gamma_n = w_n \times o_m^n$ via its history observations and node weight computed by case one. Then each node broadcasts a packet to its neighbors and the packet contain each node's weight w_n and Γ_n. After that each type sensor's truth can be computed by Eq. (3). In this way, the first iteration computing is completed via the case one and case two. The iteration process will continue until convergence criterion is satisfied. Thus the sensor node's weight can be computed in a distributed way via the algorithm 2.

D. ALGORITHM ANALYSIS

In this section, in order to prove the convergence of the problem, we first analyze the convexity property of the objective function with respect to each variables set. And then the time complexity of our proposed scheme is discussed.

Theorem 1: The iterative process shown in Algorithm 2 converges to a stationary point of the optimization problem.

Proof: Because the truths are fixed, the optimization problem Eq. (1) has only weight variables ∇. Another variable t_n is introduced and let $t_n = \exp(-w_n)$ in order to prove the convexity of Eq. (1). Thus the optimization problem can be rewritten in terms of t_n:

$$\begin{aligned} \min_{\{t_n\}_{n=1}^N} f(t_n) &= \alpha \times \sum_{n=1}^{N} -\log(t_n) \sum_{m=1}^{M} \Lambda_m(o_m^{(*)}, o_m^{(n)}) \\ &\quad + \beta \times \sum_{n=1}^{N} -\log(t_n) \times d_n \\ \text{s.t.} \sum_{n=1}^{N} t_n &= 1 \end{aligned} \tag{16}$$

According to Eq. (16), its constraint is linear in t_n. The objective function is a linear combination of negative logarithm functions and thus it is convex. Thus, Eq. (1) is convex while fixing all the truths, such that a unique minimum for w_n can be achieved with the update rule in Eq. (6).

Since the weights are fixed, Eq. (1) is a summation of quadratic functions with respect to $o_m^{(*)}$. It is convex since these quadratic functions are convex and summation operation preserves convexity. As a result, a unique minimum for $o_m^{(*)}$ can be achieved with the update rule in Eq.(15). Based on the property of the block coordinate descent, every limit point of the objective function is a stationary point if the objective function can get a unique minimum during each iteration. Based on the above analysis, the objective function is convex with respect to each variables set. Thus a unique minimum can be obtained at each iteration. As the condition holds, this theorem holds. Besides, the time complexity of the CAIF method is linear with respect to the total number of claims, i.e. $O(|o|)$, where $|o|$ is the input size of the proposed method. If the aforementioned iterative procedure is adopted, the time complexity of the CAIF method is then changed to $O(|o| \times t)$, where t is the number of iterations. Note that before information fusion, according to the node weight computed by the algorithm 2, fusion node and member nodes will be selected similar to the method in [2], which can form clusters. When the sensor nodes joining information fusion, member sensor nodes in the cluster first make local decision and transmit the local decision at each sample interval to fusion node. Fusion node makes a final decision at each sample interval.

VI. EXPERIMENTS

To evaluate the performance of our proposed method, we have conducted extensive experiments using the real world data set, which show that CAIF method is efficient and outperform traditional methods when integrating node-target distance and confidence level of observation error. Next, we first describe experimental methodology and related settings, and then experimental results and related analysis are introduced.

A. EXPERIMENTAL METHODOLOGY AND RELATED SETTINGS

1) DATA SETS

In order to obtain stringent accuracy requirements for target detection or classification while extending network lifetime as much as possible in the military vehicles monitoring [31], only a few sensor nodes are activated incessantly while most of sensor nodes are occasionally active. According to the experiment of the reference in [32], an average of 10 active sensor nodes can cover all critical locations in SensIT experiment. Thus we can construct a truncated dataset from the original SensIT dataset [31] including the sensing data of 23 nodes, which exits long-tail phenomena and is utilized to demonstrate effectiveness of the proposed method. In particular, the sensor nodes are deployed along a road and each node contains an acoustic, seismic and infrared sensor. The vehicles pass through the network 20 times along the road and provides ground truth through the GPS trajectory.The sensor readings are classified into the vehicle types which are the claims. The truncated dataset contains a total trace length of 30830 time intervals and each 0.75s data segment from each separate node is one time interval. One claim can be obtained in an time interval. Ground truth are manually assigned to each interval by a human operator to ensure high accuracy of the class labels. The real sensor data and ground truth are used as the input to a trace-driven WSNs simulation running on a computer. Besides, we assume each node is a low power mote-class device, such as Mica2 mote [33]. Although the radio communication is often loss in WSNs, we focus on information fusion performance, and assume communication is reliable.

2) PERFORMANCE MEASURES

In this experiment, we focus on continuous sensor data. To evaluate the performance of our method, we adopt the following three measures: *RMSE (Root-Mean-Square Error), Accuracy*and *Energy consumption.*

RMSE: It is computed as the percentage of the approach's output that are different from the ground truths. In the paper, we use *RMSE*as the performance measure of an approach. RMSE can be computed by Eq.(17) as follows.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(o_m^{(*)} - o_m^n)^2}{M \times N}} \tag{17}$$

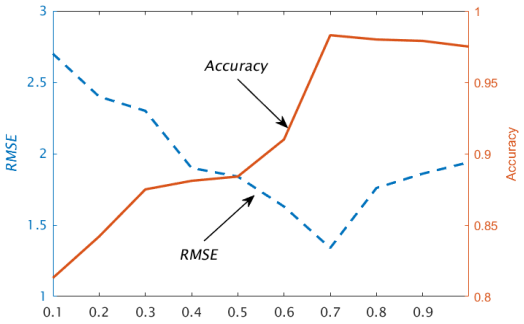

FIGURE 2. RMSE and Accuracy with different α.

For this measure, the lower the value of *RMSE*, the closer the methods estimation value is to the ground truths and thus the better the estimation performance.

Accuracy: It is defined as the ratio of the correct number of detection samples to the total number of detection samples. Thus Accuracy can be computed by Eq.(18)

$$Accuracy = \frac{\wp}{X} \tag{18}$$

where $\wp$ is the total correctly target detection samples and X is the total target detection samples. A detection sample means a CPA feature, which is variance of one time interval. Besides, the metric of energy usage is also adopted. Such energy consumption is determined by active node sampling time and transmission energy as defined in [34]. For the proposed method, iterative procedure is applied and the significance level α is set as 0.05.

3) BASELINE METHODS

In our CAIF method, weighted median is utilized for continuous data, that is because it is efficient and robust in noisy environments. Weight assignment is calculated by minimizing the upper bound of the confidence interval so that the difference in node wight is emphasized. The proposed approach is compared with traditional methods that cover a wide variety of ways to resolve conflicts. These approaches include CATD, CRH, Median, DisM and Mean method. Note that CAIF method means a distributed implementation algorithm while CAIF-C mehtod is a centralized implementation of CAIF in the following experiment.

CATD [8], [28] method (Confidence-aware Truth Discovery): It is a statistical method that has been proposed for long-tail phenomenon in truth discovery, where confidence interval is incorporated in source node weight estimation.

CRH [18], [19] method (Conflict Resolution on Heterogeneous Data): it is a framework that infers the truths from multiple sources with different data types, such as continuous and categorical data.

Median method [7]: Median calculates the median of all observations on each type sensor data of each object as the final output.

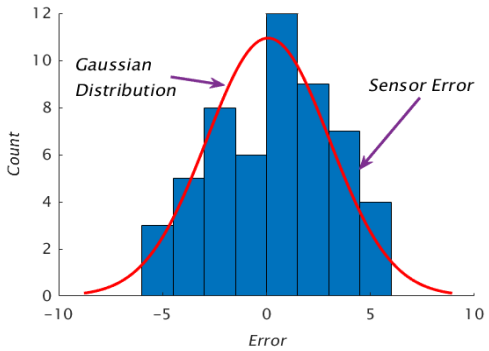

FIGURE 3. Source node error distributions.

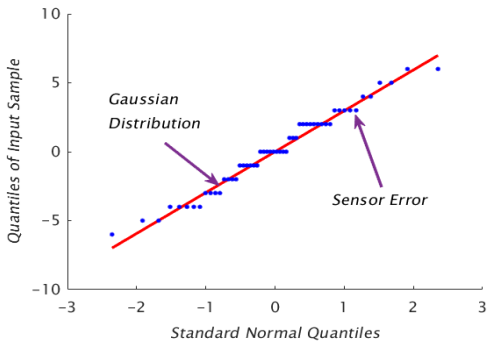

FIGURE 4. Source node error distributions.

Mean method: the truth for each type sensor is the mean of the claims. This is traditional way of resolving conflicts in categorical data without source node weight estimation and assume all the sources are equally reliable and have the same weight.

DisM method [31], [35]: Each sensor node's fusion weight is estimated via sensor-target distance and the fusion node make information fusion via such weight. Sensor-target distance often reflects each source sensor's sensing performance in a certain.

B. EXPERIMENTAL RESULTS AND RELATED ANALYSIS

1) $\alpha\beta$ SELECTION AND ASSUMPTION VALIDATION

The parameter α and β defined in Eq. (1) are analyzed via using the empirical history data. As shown in Fig.2, in the analysis process, we set different α ($\alpha = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$) to compute RMSE and accuracy for continuous sensor data respectively. According to the experiment on real data set, we find that when $\alpha = 0.7$, RMSE is lower and accuracy is higher than other settings of α as shown in Fig.2. Because $\alpha + \beta = 1$, then $\beta = 0.3$. Thus the value of α and β are set to 0.7 and 0.3 respectively.

Besides, because we have Gaussian assumption on source error distribution [8] in section V-B, normality tests is conducted via utilizing the truncated dataset to validate this assumption. Fig.3 shows the error distributions of sensor nodes from truncated dataset. Gaussian distributions are fitted and the mean is approximate 0. In order to further validate error distributions, a well-known graphical technique Q-Q plot is used to conduct normality testing. As shown in Fig. 4, data points are plotted against a theoretical Gaussian distribution (the line in the plot) and an approximate straight line indicates strong normality. Figs. 3 and 4 proves that the observation errors of the sensor nodes are indeed Gaussian distributed. Besides, practical convergence property of the proposed method CAIF is also analyzed. The proposed method converges as discussed in Section IV-C. In order to show the convergence in practice, *RMSE*on continuous data with different iterations are computed. we find that *RMSE* of CAIF can obtain stable quickly during the 10-th iteration.

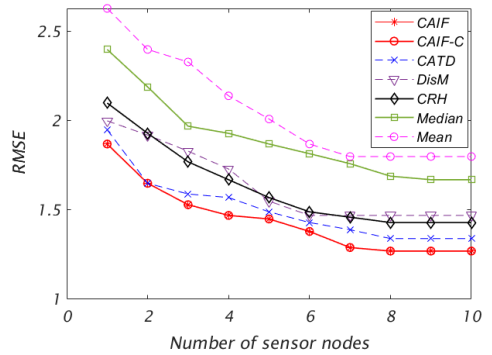

FIGURE 5. RMSE with different number of selected nodes.

2) RMSE AND ACCURACY WITH DIFFERENT NUMBER OF NODES

In this section, we make comparison our method's RMSE and accuracy with that of CRH, CATD, Median, DisM and Mean method under different number of sensor nodes. As shown in Fig. 5 and Fig. 6, CAIF has lower *RMSE* and higher accuracy than that of CRH, CATD, Median, DisM and Mean method, which demonstrate that CAIF method's estimation is closer to the ground truth and has better information fusion performance than that of all the comparison methods. That is because CAIF considers both the node-target distance and observation error's confidence level comprehensively to overcome long-tail issue better. While because method utilizes the confidence level of the observation error to solve the long-tail problem at a certain, CATD has lower *RMSE* and higher accuracy than that of other comparison methods. Besides, all the methods' *RMSE* decrease and *accuracy* increase with the increase of the number of the sensor nodes. That is because multi-source fusion can improve the performance of the information fusion.

3) SENSOR WEIGHT WITH DIFFERENT GROUPS

In this section, we make comparison our method's weight with that of CRH, CATD and DisM method under different groups. As different methods adopt various weight computation, the source weights are normalized into the range [0], [1] by dividing the maximum weight to make a fair comparison. In order to illustrate the problem brought by the

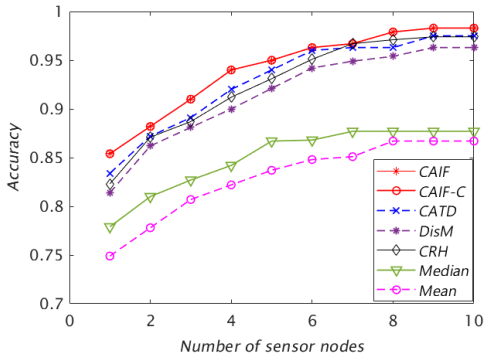

FIGURE 6. Accuracy with different number of selected nodes.

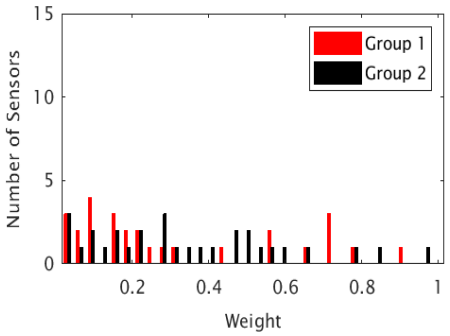

FIGURE 7. Node weight of DisM with different group.

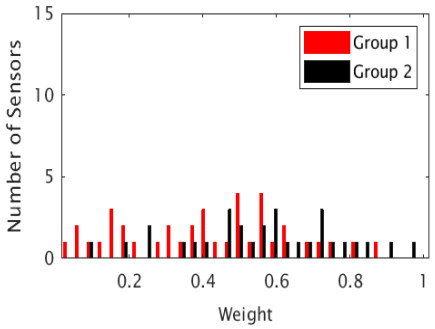

FIGURE 8. Node weight of CRH with different group.

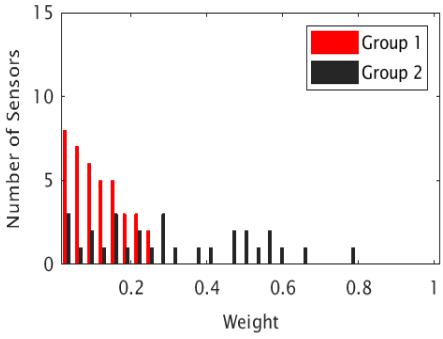

FIGURE 9. Node weight of CATD with different group.

long-tail phenomenon, sensor nodes are divided into two groups: Group 1 contains nodes with less than five claims and Group 2 contains nodes with five or more claims. This threshold is set so that the ratio of group sizes is not too extreme. Intuitively, Group 1 nodes should have small weights, because each of them provides only few observations. Group 2 nodes may have large weights or small weights depending on the sensor nodes' reliability. Figs. 7∼10 shows the weight distributions of these two groups of nodes for CATD, DisM, CRH and CAIF method respectively. Here these two baselines are chosen because the other truth discovery baselines are designed for continuous data only, thus the weights learned by those methods on numerical claims are not representative. The problem for CRH is that the weight distribution of Group 1 nodes is polarized. The number of Group 1 nodes which have weights as high as 1 stands out. Each Group 1 node only makes a few observations, and if the claims are correct, then its accuracy is high, so CRH assign a large corresponding node's accuracy is low, so it is assigned a small weight. Although CRH have reasonable node weight estimation on big sensor nodes, the inaccurate estimation on large amount of small nodes discounts their performance. Thus CRH method ignores the difference between big and small sources, and assign source weights purely based on accuracy without considering the sample size. While the problem for DisM is similar to CRH method and its weight distribution of Group 1 nodes is also polarized. The number of Group 1 sensors which have weights as high as 1 stands out. Each Group 1 sensor only makes a few observations, and if the sensor is close to the monitoring target, then its accuracy is high, so DisM method also assigns a large corresponding sensor's accuracy is low, so it is assigned a small weight. Although DisM method also have reasonable node weight estimation on small sensor nodes, the inaccurate estimation on large amount of big nodes discounts their performance. Thus DisM method ignores the difference between big and small source sensors, and assign nodes' weights based on sensor-target distance. CAIF and CATD method are aware that when the claims made by a small node happen to be accurate, it does not confirm that this small sensor node is reliable; and for big sources, the bias on node weight estimation is low. While because CAIF incorporates the node-target distance into the confidence level which can better solve long-tail issue than CATD, as shown in Fig. 10, we can see that Group 1 nodes have relatively low weights. For Group 2 nodes, some of them have low weights whereas others have big weights. Thanks to the accurate node weight estimation, the proposed CAIF method provides more accurate truths. Besides, as for the centralized implementation CAIF-C of CAIF, although the radio communication is often loss in WSNs, we focus on information fusion performance, and assume communication is reliable. Thus the performance of CAIF-C is similar to CAIF.

4) ACTIVE NODES AND ENERGY CONSUMPTION

In the section, the energy usage of our methods (CAIF and CAIF-C) are compared with other traditional methods (CATD, CRH, Mean, Median, and DisM). According to the discus-sion in [32], in order to save energy, the experiment

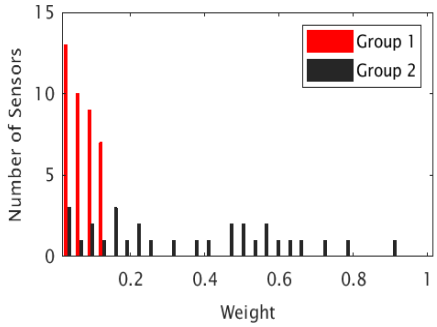

FIGURE 10. Node weight of CAIF with different group.

with 10 nodes awake at all times can meet system accuracy requirements. Thus all the methods' average active nodes are 10. Because CAIF is a distributed method, its energy usage for radio is 0.177J and its total energy usage is 26.558J. While other methods are centralized, their energy usage for radio is 3.197J and total energy usage are 29.578J respectively. Thus the distributed implementation of CAIF has lower energy usage than other methods.

VII. CONCLUSION

In order to conduct effectively information fusion in WSNs' applications, one important problem is to identify the node information fusion weight among conflicting sources of data. But it is usually unknown which one is more reliable a priori. Moreover, long-tail data phenome-non is ubiquitous in WSN-based applications which will undermine fusion performance for lower precise of node weight estimation. Specially, we show that node-target distance and confidence level can reflect the node's capability to estimate fusion weight precisely, which can solve long-tail issue in a certain. Thus a confidence-aware information fusion scheme named CAIF is presented to solve such problem via combing node-target distance and confidence interval in a distributed way. The proposed method can converge to a stationary point of the optimization problem and its time complexity is linear with respect to the total number of observations. Because of the accurate node weight estimation, the proposed CAIF method provides more accurate truths. Experimental results also demonstrate the superior performance of CAIF over existing solutions in terms of RMSE and accuracy with different number of nodes. In future work, we will make theoretical analysis about the selection of parameter α and select more metrics to make comparison with traditional methods.

REFERENCES

[1] A. W. Nagpurkar and S. K. Jaiswal, "An overview of WSN and RFID network integration," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 497–502.

[2] K. Xiao, R. Wang, T. F. J. Li, and P. Deng, "Divide-and-conquer architecture based collaborative sensing for target monitoring in wireless sensor networks," *Inf. Fusion*, vol. 36, pp. 162–171, Jul. 2017.

[3] K. Xiao, R. Wang, H. Deng, L. Zhang, and C. Yang, "Energy-aware scheduling for information fusion in wireless sensor network surveillance," *Information Fusion*, vol. 48, pp. 95–106, Aug. 2019.

[4] K. Xiao, J. Li, and C. Yang, "Exploiting correlation for confident sensing in fusion-based wireless sensor networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4962–4972, Jun. 2018.

[5] Z. Jiang, "A decision-theoretic framework for numerical attribute value reconciliation," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 7, pp. 1153–1169, Jul. 2012.

[6] F. Bauchot, J.-Y. Clement, G. Marmigere, and P. Secondo, "Voting method," U.S. Patent 7 789 306 B2, Sep. 7, 2010.

[7] R. Moessner and S. Theis, "Outlier correction with a median method," U.S. Patent 8 185 347 B2, May 22, 2012.

[8] Q. Li *et al.*, "A confidence-aware approach for truth discovery on long-tail data," *VLDB Endowment*, vol. 8, no. 4, pp. 425–436, Dec. 2014.

[9] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang, "Mining collective intelligence in diverse groups," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1041–1052.

[10] Y. Singh and U. Chugh, "Clustering, information fusion and event detection in wireless sensor networks: A review," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2015, pp. 566–571.

[11] Y. Zhao, X. Li, S. Zhang, T. Meng, and Y. Zhang, "Practical performance analysis for multiple information fusion based scalable localization system using wireless sensor networks," *Sensors*, vol. 16, no. 9, p. 1346, 2016.

[12] B Tang, B Deng, and L. Deng, "Mechanical fault diagnosis method based on multi-level fusion in wireless sensor networks," *Zhendong Ceshi Yu Zhenduan/J. Vibrat, Meas. Diagnosis*, Apr. 2016, pp. 92–96.

[13] Y. Li *et al.*, "A survey on truth discovery," *ACM SIGKDD Explor. Newslett.*, vol. 17, no. 2, pp. 1–16, 2016.

[14] D. Wang, J. Marshall, and C. Huang, "Theme-relevant truth discovery on twitter: An estimation theoretic approach," in *Proc. AAAI*, 2016, pp. 1–9.

[15] J. Marshall and D. Wang, "Mood-sensitive truth discovery for reliable recommendation systems in social sensing," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 167–174.

[16] C. Huang, D. Wang, and N. Chawla, "Towards time-sensitive truth discovery in social sensing applications," in *Proc. IEEE 12th Int. Conf. Mobile Ad Hoc Sensor Syst.*, Oct. 2016, pp. 154–162.

[17] G. Xu, H. Li, C. Tan, D. Liu, Y. Dai, and K. Yang, "Achieving efficient and privacy-preserving truth discovery in crowd sensing systems," *Comput. Secur.*, vol. 69, pp. 114–126, Aug. 2017.

[18] 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 *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 1187–1198.

[19] Y. Li *et al.*, "Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 1986–1999, Aug. 2016.

[20] S. Li, J. Xu, and M. Ye, "Approximating global optimum for probabilistic truth discovery," in *Proc. Int. Comput. Combinat. Conf.*, 2018, pp. 96–107.

[21] Y. Wang, F. Ma, L. Su, and J. Gao, "Discovering truths from distributed data," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 505–514.

[22] W. Liu, L. J. Liu, B. Wei, H. Duan, and W. Hu, "A new truth discovery method for resolving object conflicts over linked data with scale-free property," in *Proc. Knowl. Inf. Syst.*, 2018, pp. 1–31.

[23] Y. Zheng, H. Duan, and C. Wang, "Learning the truth privately and confidently: Encrypted confidence-aware truth discovery in mobile crowdsensing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 10, pp. 2475–2489, Oct. 2018.

[24] V. Iyer, S. S. Iyengar, R. Murthy, B. Hochet, V. Phoha, and M. B. Srinivas, "Multi-hop scheduling and local data link aggregation dependant Qos in modeling and simulation of power-aware wireless sensor networks," in *Proc. Int. Conf. Wireless Commun. Mobile Comput., Connecting World Wirelessly*, Jun. 2009, pp. 844–848.

[25] Y. Feng, X. Bai, N. Dang, and S. Wang, "A sleep scheduling mechanism based on power law distribution for mobile delay tolerate networks," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Sep. 2015, pp. 481–490.

[26] Z. Zhang *et al.*, "A short review on sleep scheduling mechanism in wireless sensor networks," in *Proc. Int. Conf. Heterogeneous Netw. Quality Rel. Secur. Robustness*, 2017, pp. 66–70.

[27] W. Liu, Y. Shoji, and R. Shinkuma, "Logical correlation-based sleep scheduling for WSNs in ambient-assisted homes," *IEEE Sensors J.*, vol. 17, no. 10, pp. 3207–3218, May 2017.

[28] H. Xiao *et al.*, "Towards confidence in the truth: A bootstrapping based truth discovery approach," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1935–1944.

[29] T. Huang, H. Chen, L. Cui, and S. Li, "An effective discriminator for differentiating the root causes of packet transmission failures in indoor WSNs," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 7, p. e3135, Jul. 2017.
[30] G. Shen, Z. Chen, P. Zhang, T. Moscibroda, and Y. Zhang, "Walkie-Markie: Indoor pathway mapping made easy," in *Proc. Presented 10th Symp. Netw. Syst. Design Implementation (NSDI)*, 2013, pp. 85–98.
[31] M. Duarte and Y. H. Hu, "Vehicle classification in distributed sensor networks," *J. Parallel Distrib. Comput.*, vol. 64, no. 7, pp. 826–838, Jul. 2004.
[32] M. Keally, G. Zhou, G. Xing, and J. Wu, "Exploiting sensing diversity for confident sensing in wireless sensor networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1719–1727.
[33] *MICAz_Datasheet*. [Online]. Available: https://wenku.baidu.com/view/fbb2d58ca0116c175f0e486c.html?2018.7
[34] V. Shnayder, M. Hempstead, B. Chen, G. W. Allen, and M. Welsh, "PowerTOSSIM: Efficient power simulation for TinyOS applications," in *Proc. ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2004, pp. 126–137.
[35] L. Liu, A. Ming, H. Ma, and X. Zhang, "A binary-classification-tree based framework for distributed target classification in multimedia sensor networks," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 594–602.

KEJIANG XIAO received the M.S. degree in computer science and technology from Chongqing University, Chongqing, China, in 2011, and the Ph.D. degree in computer science and technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2015. Since 2015, he has been with the Information and Communication Company of State Grid Hunan Electric Power Company, Changsha, China. His research interests include wireless sensor networks and information fusion.

ZHIWEN CHEN was born in Yongzhou, China. He received the B.E. degree in electronic information science and technology and the M.Sc. degree in electronic information and technology from Central South University, China, in 2008 and 2012, respectively, and the Ph.D. degree in electrical engineering and information technology from the University of Duisburg-Essen, Germany, in 2016. His research interests include model-based and data driven fault diagnosis.

CHUNHUA YANG received the M.S. degree in automatic control engineering and the Ph.D. degree in control science and engineering from Central South University, Changsha, China, in 1988 and 2002, respectively. From 1999 to 2001, she was a Visiting Professor with the University of Leuven, Leuven, Belgium. Since 1999, she has been a Full Professor with the School of Information Science and Engineering, Central South University. From 2009 to 2010, she was a Senior Visiting Scholar with the University of Western Ontario, London, ON, Canada. Her research interests include modeling and optimal control of complex industrial processes, fault diagnosis, and intelligent control systems.