Approximate Data Fusion Algorithm for Internet of Things Based on Probability Distribution

Xiao-qiang Wu(✉), Lan Wu, and Liyan Tu

College of Mechanical Engineering, Inner Mongolia University for the Nationalities, Inner Mongolla, Tongliao 028043, China
wxqimun@163.com

Abstract. In the context of big data, data fusion in the perception layer of the Internet of Things is extremely necessary. Fusion data can reduce the amount of data traffic in the network, avoid wasting network resources and bring great convenience to users' observation and analysis. Aiming at the high computational complexity of the data fusion algorithm at the current, an approximate data fusion algorithm for the perception layer of the Internet of Things based on the probability distribution is proposed in this paper. Firstly, the data fusion model of the perception layer of the Internet of Things and the probability distribution model of the node data are analyzed. And then, disturbances are applied to the node data to achieve the purpose of concealing the collected data. Finally, the approximate fusion of data in the sensing layer is achieved by collecting the probability distribution of the data. The experimental results verify the effectiveness of the fusion algorithm and test the influence of the algorithm parameters on the fusion effect, which provides a reference for the engineering implementation of the algorithm.

Keywords: Big data · Internet of Things · Perception layer · Data fusion · Probability distribution

1 Introduction

Internet of Things technology is an extension of Internet technology. The Internet started from the end of the 20th century with the purpose of academic exchanges. It has gradually developed into a global information exchange platform [1, 2]. The Internet affects our society, economy and culture. Due to the pursuit of intelligence, people want to add various devices and articles to the Internet. In this context, people proposed the concept of the Internet of Things. The emergence of the Internet of Things has caused many problems, such as the problem of multi-sensor data fusion in sensor networks, information connectivity problems of multiple devices and items, data communication problems between sensors, and equipment monitoring and control problems.

A key issue that needs to be solved in the data fusion of the Internet of Things awareness layer is the contradiction between limited network resources and security requirements [3, 4]. To address this issue, scholars have proposed a variety of security data fusion schemes from different perspectives. The secure information aggregation

S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, pp. 137–144, 2019.
https://doi.org/10.1007/978-3-030-19086-6_15

protocol first presents the data fusion results of the data fusion node, and then uses efficient sampling and interaction verification to ensure that the fusion value is an approximation of the true value. However, this protocol requires high reliability data values and consumes high node resources [5]. The secure data aggregation and verification protocol uses a key sharing scheme to distribute keys to nodes in the cluster. At the same time, nodes in the cluster partially sign the average value of the calculated data in the cluster. This scheme can verify the integrity of the fused data, but the amount of calculation is large [6]. Luo proposed efficient and secure data aggregation, using fuzzy algorithms and model codes to eliminate redundant information sensed by sensor nodes and perform corresponding data fusion operations. This method helps to improve the confidentiality of data and the energy efficiency of sensor nodes [7]. Ganeriwal and Srivastava proposed a reputation based framework sensor networks trust model for wireless sensor networks when researching data fusion technology [8].

The above security scheme either increases the number of interactions between nodes or poses excessive challenges to the nodes' computation and storage resources. Therefore, an approximate data fusion algorithm for the perception layer of the Internet of Things based on the probability distribution is proposed. The chapters of the manuscript are arranged as follows. The first part is the introduction. The second part introduces the model of data fusion in the perception layer of the Internet of Things. The third part proposes the approximate fusion algorithm of sensory data. The fourth part analyzes the performance of the fusion algorithm. The fifth part is a simulation experiment. The last part is the conclusion.

2 Internet of Things Sensing Layer Data Fusion Model

The sensing layer is composed of various sensors or sensor networks and controllers. Its function is similar to that of human beings. It can acquire natural signals. It is mainly used to collect and process signals to a certain extent, form information or identify objects. The core technologies of this layer include sensor technology, computer control technology, and radio frequency technology. The core products involved include sensors, sensor networks, and controllers.

Multi-sensor data fusion is also called information fusion. For information fusion, it is difficult for researchers to give a consistent and comprehensive definition. The overall data produced by multiple sensors is denser than the information it has in its various components. The multi-sensor data fusion method relates to the quality and efficiency of information fusion, which is mainly reflected in the fusion algorithm. Therefore, the core problem in the research of information fusion technology is to study the fusion algorithm. Due to the diversity and complexity of information, information fusion methods must have certain parallel processing capabilities and robustness. In general, non-linear mathematical methods that are fault-tolerant, adaptable, memory-capable, and parallel-processing are all fusion algorithms. At present, the more common data fusion algorithms can be divided into two categories. They are data fusion method based on stochastic theory and artificial intelligence based data fusion method.

2.1 Network Models and Security Assumptions

Assume that the perception layer of the Internet of Things is large and the number of nodes is large. The node is fixed in position after it is deployed and will not be moved again. The network topology structure is a cluster tree hybrid structure whose model is shown in Fig. 1.

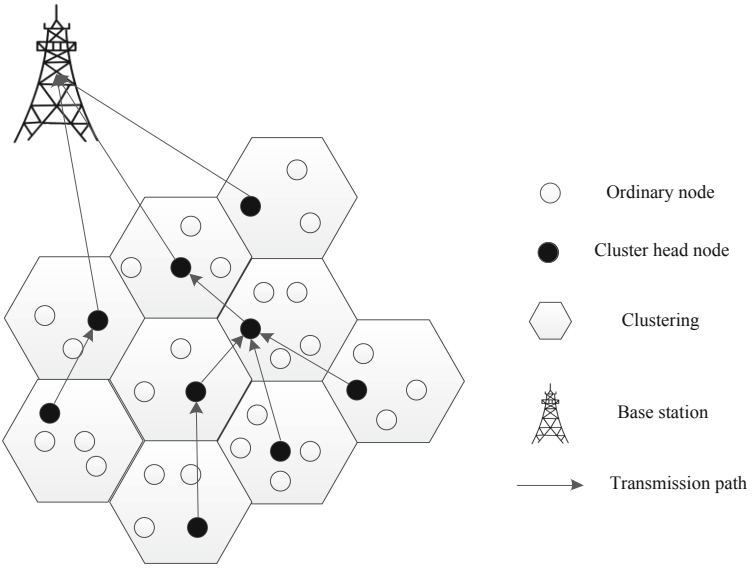

Fig. 1. Network topology

The central node of the entire network is called a base station. The base station summarizes all the information. Each small area in the network forms a cluster. The nodes in the cluster responsible for data fusion aggregation are called cluster heads. A tree structure is formed between the cluster heads, and the result of the fusion is uploaded to the base station in a multi-hop manner.

2.2 System Parameter Design and Workflow

The data security fusion algorithm consists of two parts: data encryption and data fusion and ID compression fusion. Table 1 shows the meaning of symbols in the system.

Table 1. System parameters and meanings

Symbol	Representative meaning
S_i	The node numbered i
$Seed_i$	Shared secret random number seed shared by node i and base station
p_i	Data collected by node i
$f(x)$	The function that generates the session key, parameter x is a random number seed
key_i	Session key of Node i
$cipher_i$	The data collected by node i is encrypted in cipher text
Q	Large prime Q for modulus
Tab_i	List of child nodes of cluster head i
$Count$	Number of nodes in the cluster
$Length$	ID number compression coded packet length
$Number$	ID number compression coding group number

3 Approximate Fusion Method Based on Probability Distribution

The data collected by the nodes consists of the actual values of the data and measurement errors. The measurement error is usually small. Therefore, if the attacker directly intercepts the measured values, it can be approximated as the actual data. The main idea of the proposed algorithm is to use the method of expanding the measurement error to perturb the original data, so that the data after the disturbance is invalid to the attacker because the error is too large. All nodes in the network share rule bits: The original collected data is added to a variable that is uniformly distributed, and the result is transmitted to the cluster head. The cluster head receives the data of the nodes in the cluster and adds them directly. Because the intermediate process does not need to be decrypted, the algorithm satisfies the additive homomorphism feature, and the cluster head receives less data. If the attacker intercepts the fusion result of the cluster head and simply replaces the disturbance data superimposed by all the nodes with a uniformly distributed mathematical expectation, there is still a large error. Because the base station has data of the entire network, in the case of a large amount of data, more accurate results can be obtained by replacing the disturbance data with mathematical expectation.

The network consists of a base station and N clusters. The cluster heads are denoted by $A_1, A_2 \ldots, A_N$, respectively. Before the data transmission starts, the algorithm flow is described according to the node type.

3.1 Ordinary Leaf Node Fusion

Collect raw data, denoted as v_i. Generate random variable X_i, X_i is uniformly distributed $U(0, R)$, where R is the system given parameters. Calculate the fusion result according to the following formula

$$cp_i = v_i + X_i \tag{1}$$

The fusion result is passed to the cluster head.

3.2 Cluster Head Data Fusion

Let the cluster head receive data as $p_1, p_2, \ldots, p_i$. Calculate the sum of ciphertext received by cluster heads

$$sum_i = \sum_{j=1}^{n_i} p_i \tag{2}$$

The data received by each cluster head is

$$p_i = Enc(sum_i | num_i, key_i) \tag{3}$$

Each cluster head data is broadcast to the base station through multiple hops.

3.3 Base Station Data Fusion

According to the ID number and corresponding symmetric key, Data is decrypted based on the decryption function. The sum of the ciphertexts of all nodes is

$$sum = \sum_{i=1}^{N} sum_i \tag{4}$$

The sum of the original data collected by the node estimated by the base station is

$$total = sum - \sum_{i=1}^{k} num_i \tag{5}$$

Therefore, the actual average value of node data can be calculated by

$$average = total/N\_work \tag{6}$$

4 Performance Analysis

Approximate fusion method based on probability distribution uses ambiguous methods to encrypt data. For node S_i, the original data collected is v_i, the generated disturbance data is X_i, and the encrypted data is

$$p_i = v_i + X_i \tag{7}$$

The security of cryptographic algorithms lies in the degree of confusion between ciphertext and plaintext. When the number of nodes is large, the maximum value of the parameter R that the disturbance data obeys the uniform distribution is usually larger than the average value of the original collected data. Let

$$R = \mu + 3\sigma \tag{8}$$

At this time, data security is the highest. If the system security requirement is not very high, R can be appropriately reduced to reduce system errors.

5 Experiment Analysis

Assume that the network consists of N_group clusters. The number of child nodes in each cluster head is equal to C_Num and both are 100. The variables collected at different nodes in the same cluster are independent of each other and subject to the same normal distribution. The systematic error threshold is β, the perturbation data $X \sim U(0, R)$ added by each node in the network, the normal distribution image selects 0.025 points. Search for the appropriate R value based on the parameters. The goal of the search is that the base station can estimate the approximate value of the error threshold by statistical means. Through the way of controlling variables, study the influence of different parameters on the value of R. The experimental results are shown in Figs. 2, 3 and 4.

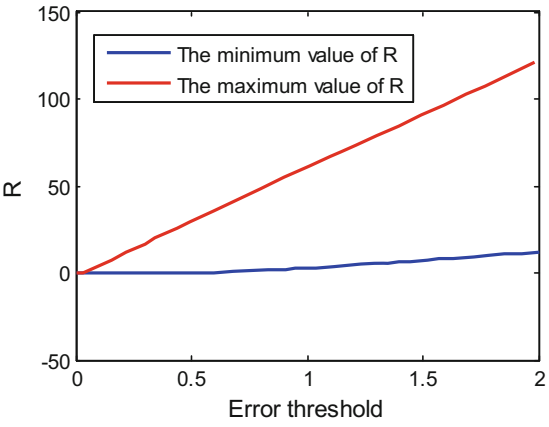

Fig. 2. Relationship between error threshold and *R*

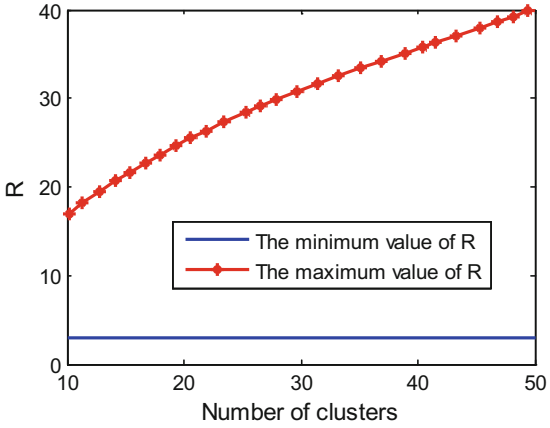

Fig. 3. Relationship between the number of clusters and *R*

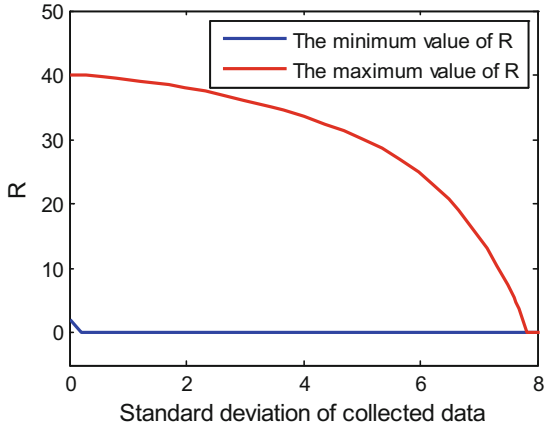

Fig. 4. Relationship between standard deviation and *R*

The parts between the two curves in Figs. 2, 3 and 4 are all the range of R that meets the requirements. On the basis of satisfying the requirements, increasing R increases the security of the system and decreases R to reduce the system error. Therefore, the R value needs to be set by the user according to the actual situation. As can be seen from Fig. 2, the maximum and minimum values of R increase as the error threshold increases. From Fig. 3, it can be seen that when the number of nodes in the cluster is fixed, the maximum value of R increases as the number of clusters increases, and the minimum value is independent of the number of clusters. As can be seen from Fig. 4, both the maximum and minimum values of R decrease as the standard deviation of the acquired data increases, and when the standard deviation is too large, the R value that meets the requirement may not be found.

6 Conclusion

An approximate fusion algorithm based on the probability distribution for the Internet of Things is proposes in the manuscript. The accuracy of the fusion results is related to the size of the network. The more nodes in the network, the higher the accuracy of the fusion results is. Theoretical analysis and simulation experiments have proved its feasibility, energy saving and safety. However, although this fusion algorithm can realize the data security fusion, it does not form a complete protocol. It still needs to be studied in terms of key distribution and key update.

Acknowledgements. Inner Mongolia National University Research Project (NMDYB1729). Inner Mongolia Autonomous Region Science and Technology Innovation Guide Project in 2018: KCBJ2018028.

References

1. Daza, L., Misra, S.: Beyond the Internet of Things: everything interconnected: technology, communications and computing. IEEE Wirel. Commun. **24**(6), 10–11 (2018)
2. Xiao, F., Miao, Q., Xie, X., et al.: Indoor anti-collision alarm system based on wearable Internet of Things for smart healthcare. IEEE Commun. Mag. **56**(4), 53–59 (2018)
3. Wang, M., Perera, C., Jayaraman, P.P., et al.: City data fusion: sensor data fusion in the Internet of Things. Int. J. Distrib. Syst. Technol. **7**(1), 15–36 (2015)
4. Gong, B., Wang, Y., Liu, X., et al.: A trusted attestation mechanism for the sensing nodes of Internet of Things based on dynamic trusted measurement. China Commun. **15**(2), 100–121 (2018)
5. Kalpakis, K., Dasgupta, K., Namjoshi, P.: Efficient algorithms for maximum lifetime data gathering and aggregation in wireless sensor networks. Comput. Netw. **42**(6), 697–716 (2003)
6. Wang, X., Mu, Y., Chen, R.: An efficient privacy-preserving aggregation and billing protocol for smart grid. Secur. Commun. Netw. **9**(17), 4536–4547 (2016)
7. Luo, W., Hu, X.: An efficient security data fusion protocol in wireless sensor network. J. Chongqing Univ. Posts Telecommun. Nat. Sci. Ed. **21**(1), 110–114 (2009)
8. Ganeriwal, S., Balzano, L.K., Srivastava, M.B.: Reputation-based framework for high integrity sensor networks. ACM Trans. Sens. Netw. **4**(3), 1–37 (2008)