A Distributed Anomaly Filtering Algorithm for Heterogeneous Data Based on City Computing

Shiwei Wang
Future Network Innovation Center
Beijing University of Technology
Beijing,China
+86 15733253135
edward_gavin@emails.bjut.edu.cn

Yangyang Li
National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data, China Academy of Electronics and Information Technology
Beijing, China

Xiaobin Xu*
Future Network Innovation Center
Beijing University of Technology
Beijing, China
xuxiaobin@bjut.edu.cn

Guijie Yue
National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data, China Academy of Electronics and Information Technology
Beijing, China

ABSTRACT
In modern cities, numerous urban perception devices collect and release urban data all the time, but urban data may become abnormal due to environmental interference or artificial tampering. In view of the problem that urban data will face data anomalies, this paper designs a distributed gauss membership anomaly data filtering algorithm, and defines a set of extraction protocols suitable for heterogeneous data. Simulation results show that this algorithm can filter abnormal data in real time, improve the efficiency of urban computing and reduce the cost of network.

CCS Concepts
• **Computing methodologies→Distributed algorithms.**

Keywords
Urban computing; the internet of things; big data; abnormal filtering; mobile edge computing.

1. INTRODUCTION
Urban computing is based on computer technology, which solves some problems existing in cities by constantly acquiring, integrating and analyzing various heterogeneous data in cities. Literature [1], [2] describes the extensive application of the Internet of things in urban transportation, environment and other fields, but also points out that in urban data collection, there are many types of sensing devices and complex data transmission channels. Literature [3] discussed that attackers injected false data into the network, which had a bad impact on terminal decision-making. In particular, if no one destroys the data in the process of data transmission, the data will also face the problem of data anomaly [4], [5]. On the one hand, these false or abnormal data will affect the quality of the collected data and lead to the wrong decision of the urban system. On the other hand, they will occupy the valuable network and computing resources of urban computing, causing negative impact. Therefore, urban data security has become an important proposition in urban computing [6].

The capacity of urban computing is precious and limited, so it is particularly important to allocate resources reasonably. Mobile edge computing sinks the cloud computing center to the edge of cloud, making the cloud computing center closer to the demand side of the resource. In the face of the scenario with large data throughput, the distributed processing method can effectively reduce the overall transmission delay of the network and improve the processing efficiency of the system [7], [8].

Now, urban anomaly data processing has been applied in many aspects of urban computing: Literature [9] explored the abnormal data analysis of 160 million taxi trip records in New York City; Literature [10] explored the abnormal data detection of large-scale traffic data; Literature [11] explored the monitoring of abnormal data in wireless sensor networks. Literature [9], [10] explored the problem of abnormal data processing for a single application of the city, Literature [11] explores the problem of abnormal data processing in local sensor networks. However, a single application scenario data solution may rely on the unique characteristics of the application data, resulting in a non-generic approach. At the same time, the abnormal data solution in the local network can't satisfy the "multi-data, multi-task" scenario at the city scale, so this paper proposes a distributed anomaly filtering algorithm for heterogeneous data at the city level.

The main work of this paper is as follows: An extraction protocol for heterogeneous data is designed for application layer. In this protocol, each data source device sends all data to the network at the same time. Different applications can get the required type data by setting the offset value according to the actual needs, so as to realize the quick acquisition of heterogeneous data; Based on the mobile edge computing architecture, a distributed anomaly data filtering algorithm is designed. Fuzzy set is used to represent each kind of data and the membership function of abnormal data is calculated. A single node filters out abnormal data and marks suspicious data by setting thresholds for suspicious data and abnormal data. After receiving the data of multiple nodes, the suspicious data is further analyzed based on the data of multiple nodes and the abnormal data is filtered.

ICCAI '20, April 23–26, 2020, Tianjin, China

ACM ISBN 978-1-4503-7708-9/20/04…$15.00
DOI: https://doi.org/10.1145/3404555.3404636

The rest of this paper is structured as follows: section 2 introduces relevant work; Section 3 introduces the heterogeneous data extraction protocol for city computing; Section 4 presents the filtering algorithm of abnormal data; Section 5 gives the results and analysis of the simulation experiment. Section 6 summarizes the whole paper and points out the key points for the next step.

2. RELATED WORK

The key point of abnormal filtering algorithm is to realize the recognition of abnormal data or outlier. Regarding the identification of abnormal points, literature [12] established the feature space of attributes, and calculated the outlier distance of data points to judge the degree of anomalies of data points. This method has good detection effect and wide applicability, but the time complexity of this algorithm is too high, the real-time recognition ability is weak, and can't meet the needs of real-time distributed scene. Literature [13] propose a one-class support Tucker machine (OCSTuM) and an OCSTuM based on tensor Tucker factorization and a genetic algorithm called GA-OCSTuM, this method can retain the structural information of data while improving the accuracy and efficiency of anomaly detection. But this method can only detection raw data in one step. Literature [14] present a tunable algorithm for distributed outlier detection in dynamic mixed-attribute data sets, but this method consumes a lot of memory and is relatively inefficient. In literature [15], [16], the method of density clustering is used to capture the outliers. However, the method based on clustering is limited to the selection and number of clustering clusters, and each clustering model is only applicable to specific data types. Literature [17] proposed an outlier recognition algorithm based on gaussian statistics, but the method based only on statistics, which relies excessively on prior knowledge, could not deal well with non-prior cases.

In order to solve the above problems, and considering the spatial-temporal correlation of data, this paper proposes a distributed anomaly filtering algorithm based on gaussian membership deployed on sensing nodes. By analyzing the membership degree of data and the membership degree of data difference, the algorithm is jointly responsible for filtering the abnormal and invalid information. At the same time, the mobile edge calculation is used to further filter the abnormal data by combining multiple nodes, so as to further improve the filtering level by analyzing the abnormal data from a higher dimension. At the same time, combining with the filtering algorithm proposed in this paper, this paper proposed the matching city data extraction protocol.

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

3. HETEROGENEOUS DATA EXTRACTION PROTOCOL

3.1 Data Definition in City Computing

In urban computing, the number of device terminals and data types obtained is large, so the transmission message can be formulated uniformly to reduce the pressure of system processing and identification.

Assuming that there are n types of data that might be included in the city calculation, then each type of device can obtain m types of data.

In this paper, fixed-length encoding is used to encode n data types, and the encoding length L is obtained according to definition 1. Starting with the binary number 0, it corresponds to the city data type in a self-incrementing form, forming a unique id number for each data type.

Definition 1

$$L = log_2 n \quad (1)$$

As for the data content, due to the different length of city data content, this paper adopts the method of hard coding to directly set the data length for each type of data at the perception node.

3.2 Data Packet Structure

The contents of the packet are shown in Table 1, where the data type id and data content are required to match the added packets. Packets are generated and uploaded as binary streams.

Table 1. Data packet field

Field	Optional/Required	Describe
Data type id	Required	City data type id
data content	Required	data content
Device id	Required	Sensing device id
Device type	Optional	Device type
Service type	Optional	Environment, transportation, etc.

3.3 Data Extraction Method

There are many possible applications of urban computing, each requiring several types of data. Each application record needs to record the device id of the data source and the data content within each data type under that device id.

The data extraction process is shown in Figure 1. Codecs provide the ability to transform and analyze data streams. Codecs analyze data streams in the form of data slicing, decompose heterogeneous data streams into multi-class data fragments, and extract urban data information from them.

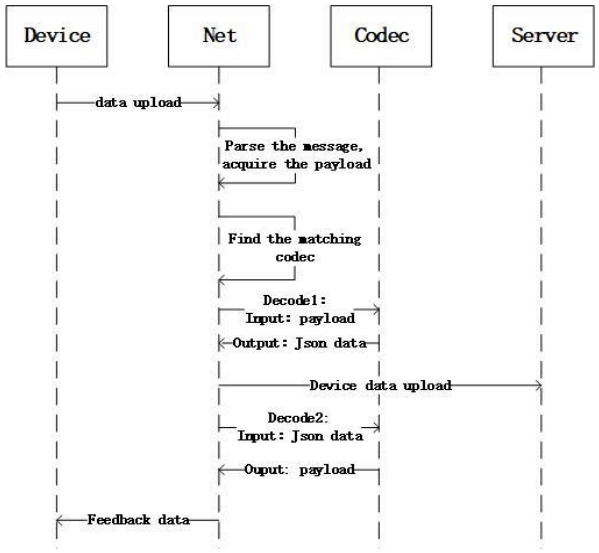

Figure 1. Data processing flow.

4. ABNORMAL DATA FILTERING ALGORITHM

In large-scale city data filtering, each device is required to have the ability to filter data independently. In order to solve the

problem of equipment independent filtering, a data filtering scheme based on gaussian membership analysis is proposed. This algorithm calculates the gaussian membership degree of the city data and the gaussian membership degree of the change of the data difference. The gaussian membership degree of data describes the membership degree of data corresponding to the whole from the perspective of data distribution, and the gaussian membership degree of data difference change discusses the possibility of continuous data when data change from the perspective of data difference change distribution. The joint membership degree of the data reflects its trust relation and membership relation to the whole. The higher the membership, the higher the data reliability. At the same time, combining with the spatiotemporal correlation of data, this paper proposes the abnormal data filtering based on mobile edge cloud, which further improves the ability of data mining and analysis.

4.1 Membership Analysis of Single Node Abnormal Data

In practical applications, many urban physical data can be described by gaussian distribution or approximate gaussian distribution [17]. According to probability theory, the description of gaussian distribution is shown in definition 2.

Definition 2 A normal distribution is an important probability distribution. Its probability density function is:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-u)^2}{2\sigma^2}\right) \quad x \in (-\infty, +\infty) \qquad (2)$$

In this formula, u and σ are constants, and σ>0, then *x* follows the gaussian distribution, *X~N (u, σ^2)*, X is normal random variable.

For the city data satisfying X~N (u, σ^2), we can preprocess the city data, calculate the overall data distribution probability density function $P(x)$, and calculate the data gaussian membership $Y(x)$ according to definition 3.

Definition 3 This definition describes the mapping from probability density function to data gaussian membership.

$$\mathrm{Y}(x) = \mathrm{P}(x) * \sqrt{2\pi}\sigma \qquad (3)$$

In this formula, $P(x)$ is the probability density function of the gaussian distribution corresponding to the city data, and is the variance of the gaussian distribution corresponding to it. In the definition, data values that deviate too far from their historical distribution will result in smaller data gaussian membership.

Similarly, according to the difference of data change, the gaussian distribution of data difference change can also be established to obtain the gaussian membership degree $D(x)$ of data difference.

The joint membership of the data is $P(x)$, $A(x) = Y(x) * D(x)$, $A(x) \in [0, 1]$. For the joint membership degree of data, we can further differentiate, mark and filter the data by defining the threshold T_{at} of abnormal membership degree and T_{st} of suspicious membership degree. For $A(x) \in [0, T_{at})$ mark it as abnormal data for filtering; For $A(x) \in [T_{at}, T_{st})$ mark it as suspicious data; For $A(x) \in [T_{at}, T_{st})$ mark it as reasonable data.

The data filtering process is shown in Figure. 2. After the data passes through the filter, it is divided into abnormal data, suspicious data and reasonable data. Upload suspicious data and reasonable data, and filter out abnormal data directly locally.

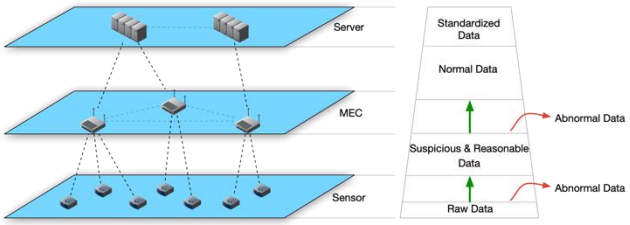

Figure 2. Perception node and MEC network model on urban computing.

4.2 Abnormal Data Filtering Based on Mobile Edge Cloud

The filtering algorithm based on gaussian membership can only well solve the problem of data filtering in the normal data range. However, in case of special circumstances, mobile edge computing can help urban computing to consider the problem of data anomalies from a higher dimension, as shown in Figure. 3.

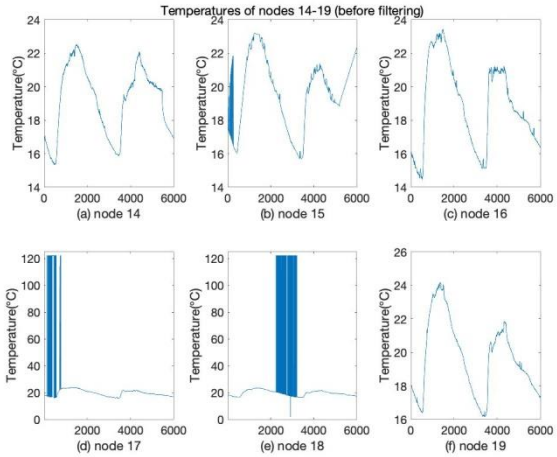

Figure 3. The original data.

In the mobile edge layer, data of multiple nodes can be analyzed jointly to mine the potential of data anomalies. If the data uploaded by many nodes is suspicious, it can be attributed to abnormal changes in the environment. For example, if the temperature drops suddenly, the temperature data collected by all nodes will drop sharply, then such data is normal data. By setting the suspicious number threshold *m*, when the number of nodes reporting suspicious changes is greater than the suspicious number threshold, the suspicious change of data is considered as normal change. If it is less than this threshold, the suspicious change of the data is considered as abnormal change and is filtered.

5. EXPERIMENTS

Based on the above algorithm design, this section mainly tests the filtering ability of the algorithm. In this article, we will in a real data set on the basis of experimental design, as the chart shows, in the range of the whole city, there are many sensing nodes, collect a variety of heterogeneous data, the city of Beijing University of Technology campus layout the six nodes, node to once every 30 s frequency acquisition of data, data collection of temperature, humidity, illumination city, etc. Experimental environment: MATLAB R2017a development, processor Intel(R) Core (TM) i5-4690mq, memory 8 GB, operating system for Windows10.

In this experiment, our algorithm will simulate six heterogeneous sensors on the campus of Beijing university of technology. The data collected by the six heterogeneous sensors all contain temperature data, which is generated based on the real data set of Berkeley Intel laboratory.[18] Our simulation experiment extracts the temperature data from the complex data of six heterogeneous nodes, and adopts the distributed real-time data filtering algorithm based on the mobile edge cloud architecture to identify and filter the abnormal data.

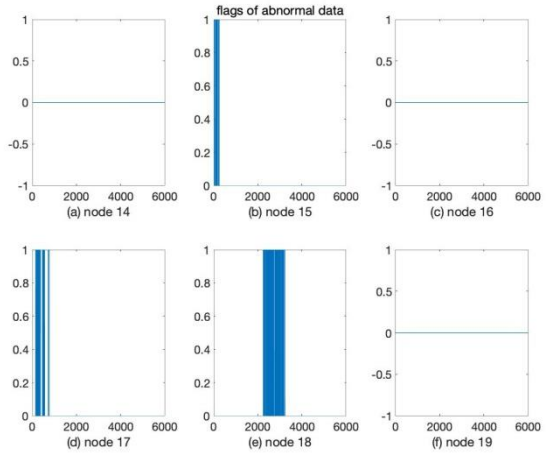

Figure 4. Marked Abnormal data.

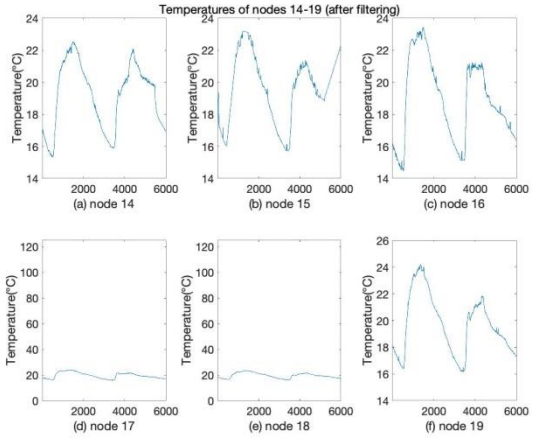

Figure 5. Filtered data.

In this experiment, the abnormal membership threshold T_{at} was set as 0.005, the suspicious membership threshold T_{st} was set as 0.01, and the suspicious number threshold m was set as 2. Figure. 3 shows the distribution of the original data. It can be seen that the data has a large fluctuation and the data anomalies are relatively obvious. Figure. 4 shows the distribution of abnormal data. It can be seen that the distribution of abnormal data is relatively concentrated. Figure. 5 shows the filtered data. By comparing Figure. 4 and Figure. 5, after algorithm processing, the data has a better recovery effect.

Simulation results show that the algorithm has good ability to filter abnormal data.

6. CONCLUSION AND FUTURE WORK PROSPECT

This paper proposes a data filtering scheme based on gaussian membership analysis. By calculating the gaussian membership degree of the data, the data is filtered preliminarily. At the same time, based on the mobile edge cloud, this paper proposes a multi-node combined filtering method, which further improves the ability of data mining and analysis. Simulation results show that the algorithm has good ability to filter abnormal data. But the algorithm needs enough prior information, and the time complexity of preprocessing is high, so it needs further optimization.

7. ACKNOWLEDGMENTS

This research is supported in part by National Key Research and Development Project (Grant No. 2017YFC0820506), the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC), the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences(No.LSU-KFJJ-2018-06), the International Research Cooperation Seed Fund of Beijing University of Technology(No.2018B41).

8. REFERENCES

[1] Atzori, L., Iera, A., and Morabito, G. 2010. The Internet of Things: A survey. *Computer Networks*, 54(15):2787-2805.

[2] Henrik, B., Niels, O. B., Tobias, F., Kaj, G., Mikkel, B. K., Paul, L., and Markus, W. 2013. On heterogeneity in mobile sensing applications aiming at representative data collection. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication* (UbiComp '13 Adjunct). Association for Computing Machinery, New York, NY, USA, 1087–1098. DOI=https://doi.org/10.1145/2494091.2499576

[3] Farid, L., Ahc ène, B., Rahim, K., Reinhardt, E., and Massinissa S. 2016. Faulty Data Detection in Wireless Sensor Networks Based on Copula Theory. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies (BDAW '16)*. Association for Computing Machinery, New York, NY, USA, Article 29, 1–7. DOI=https://doi.org/10.1145/3010089.3010114

[4] Chen, P.S., Lin, S.C., Sun, C.H. 2015. Simple and effective method for detecting abnormal internet behaviors of mobile devices. Inf. Sci. 321, 193–204 (2015)

[5] Pin, Z., Jing, X., Muazu, H., and Wen-Min, M. 2015. Access control research on data security in Cloud computing. 873-877. 10.1109/ICCT.2015.7399964.

[6] Yu, Z., Licia, C., Ouri, Wolfson., and Hai, Y. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans.* Intell. Syst. Technol. 5, 3, Article 38 (September 2014), 55 pages. DOI=https://doi.org/10.1145/2629592

[7] Jing, Z., et al. 2017. An evolutionary game for joint wireless and cloud resource allocation in mobile edge computing. *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, Nanjing, 2017, pp. 1-6. doi: 10.1109/WCSP.2017.8170956.

[8] Meng-Ting, L., Richard, Y., Ying-Lei, T., et al. 2018 Computation Offloading and Content Caching in Wireless

Blockchain Networks with Mobile Edge Computing. *IEEE Transactions on Vehicular Technology*, 2018:1-1.

[9] Zhang, J. 2012. Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC. *Acm Sigkdd International Workshop on Urban Computing*. ACM, 2012.

[10] Lam, P., Wang, L., Ngan, H. Y. T., et al. 2017. Outlier Detection in Large-Scale Traffic Data by Naïve Bayes Method and Gaussian Mixture Model Method. *Electronic Imaging*. 2017.

[11] Ghorbel, O., Ayadi, A., Loukil, K., et al. 2017. Classification data using outlier detection method. In Wireless sensor networks. *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 2017.

[12] Han-Chen, S., Ling-Da, W., Ying-Mei, W. 2005. Dispersed Featured Point Detection. *Acta Simulata Systematica Sinica*, 2005.

[13] Deng, X., Jiang, P., Peng, X., and Mi, C. 2019. An Intelligent Outlier Detection Method with One Class Support Tucker Machine and Genetic Algorithm Toward Big Sensor Data in Internet of Things. In *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4672-4683, June 2019.

[14] Otey, M.E., Ghoting, A., and Parthasarathy, S. 2006. Fast Distributed Outlier Detection in Mixed-Attribute Data Sets. *Data Min Knowl Disc* 12**,** 203–228 (2006). https://doi.org/10.1007/s10618-005-0014-6

[15] Shukla, M. and Kosta, Y. P. 2016. Empirical analysis and improvement of density-based clustering algorithm in data streams. *2016 International Conference on Inventive Computation Technologies (ICICT).*, Coimbatore, 2016, pp. 1-4.

[16] Yamei, L. and Renwu, Y. 2019. A Distributed Outlier Detection Algorithm Based on Density Clustering. *Computer & Digital Engineering*, 2019.

[17] Xiao, D.Q., Feng, J.Z., Zhou, Q., and Yang, B. 2008. Gauss Reputation Framework for Sensor Networks. J. Commun. 2008, 29, 47–53.

[18] http://db.csail.mit.edu/labdata/labdata.html