Edge Computing Based Real-Time Streaming Data Mining Method for Wireless Sensor Networks

Zhong-xing Huang[1], Xiao-li Ren[2(✉)], Zai-ling Zhou[1], He Zhu[1], and Zhi-li Lin[1]

[1] Guangzhou Metro Design and Research Institute Co., Ltd., Guangzhou 510000, China
[2] Zhongye Design Co., Ltd., Guangzhou 510000, China

Abstract. Traditional data mining techniques are difficult to be directly applied to wireless sensor networks because of the multidimensional and multilayered characteristics of wireless sensor networks. Based on the theory of edge computing, the framework of distributed data mining workflow in wireless sensor networks is optimized, and the flow of distributed data mining in wireless sensor networks is demonstrated. Finally, the design requirements of data mining methods are realized.

Keywords: Edge computing · Wireless sensor network · Data mining

1 Introduction

Due to the large scale and random deployment of WSN, the communication environment, limited energy supply and high failure rate are often impaired, which makes WSN knowledge mining face many severe challenges. A lot of dynamic data will be generated in sensor network applications. In order to ensure the operation effect of the network, the edge computation is used to analyze the network data and extract the knowledge, and the real-time stream data is mined. Therefore, the development of wireless sensor network data mining technology, is essentially to promote the development of real-time intelligent wireless sensor networks. In the literature [6], some scholars put forward a method of integrated feature clustering, which uses matrix representation and convolutional neural network to extract and fuse features, and uses multi-source data structure combined with missing data interpolation method to achieve high accuracy of data mining. In the literature [7], aiming at the data processing of soil, two data mining algorithms, multiple adaptive regression and gene expression programming, are proposed to construct a data model, train the data model and collect features, thus realizing the deep mining of complex data. Traditional data mining can not be directly applied to wireless sensor networks because of its centralization, heavy computation and emphasis on transaction data processing. Therefore, this paper proposes a real-time stream data mining method based on edge computing for wireless sensor networks. By analyzing the data characteristics of wireless sensor networks, the frequent itemsets of data sets are obtained, the data

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 417, pp. 144–155, 2022.
https://doi.org/10.1007/978-3-030-94554-1_13

management architecture is constructed by using sensor nodes, the data is compressed to reduce the traffic, and the data mining process is optimized by using the edge computing theory.

2 Real-Time Streaming Data Mining Method for Wireless Sensor Networks

2.1 Composition of Wireless Sensor Network Real-Time Streaming Data Structure

Real-time stream data feature mining in wireless sensor networks is to extract application-oriented, acceptable and accurate data models and patterns from the continuous and fast data streams in sensor networks. In the process of data mining, the data can not be stored and must be processed in time. Data mining algorithm must be effective and fast processing of high-speed data. Traditional data mining algorithms are good at processing and analyzing static datasets, but not suitable for processing large, high dimensional and distributed data generated by wireless sensor networks. Based on this, we first classify the features of real-time stream data in wireless sensor networks as follows (Table 1):

Table 1. Characteristics of real-time streaming data in wireless sensor networks

Name	Traditional data	Wireless sensor network data
Processing architecture	Focus	Distribution
Data type	Static state	Dynamic
Memory usage	Unlimited	Restricted
Processing time	Unlimited	Restricted
Computing power	High	Low
Energy	Unlimited	Limited
Data stream	Static	Successive
Data length	Limited	Infinite
Response time	Non real time	Real time
Update rate	Low	high
Number of scans	Many times	Single time

Data mining is based on the features of different classes of data. The task of real-time stream frequent patterns mining in wireless sensor networks is mainly carried out under the condition of limited computing and storage resources. According to the mining results, we mine the maximal frequent itemsets, closed frequent itemsets, complete frequent itemsets and Twk frequent itemsets for real-time stream data in wireless sensor networks. Furthermore, the random mining algorithm in frequency range based on

relative error count can divide the WSN data into probability-based approximation algorithm and deterministic error interval approximation algorithm. The networking style of wireless sensor networks is shown in the following figure (Fig. 1):

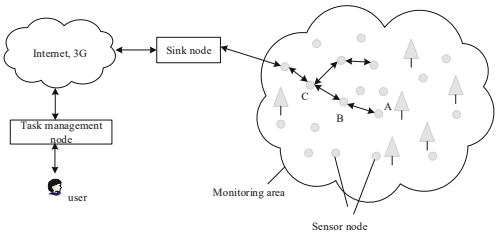

Fig. 1. Composition of wireless sensor networks

Based on the above structure, the sensor node of wireless sensor network is a micro-embedded system. In different application background, the composition of the sensor node is different (Fig. 2).

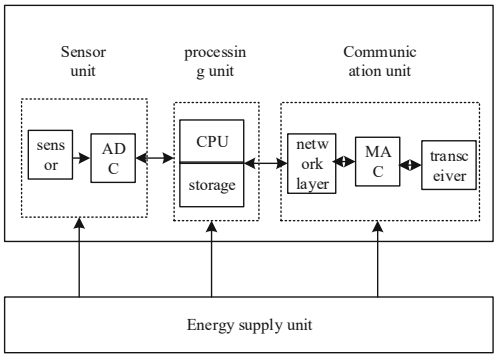

Fig. 2. Sensor node data feature processing module

The sensor unit is responsible for data acquisition, and the processing unit is responsible for data processing and controlling the whole node. In the frequent pattern mining of stream data, we can use the timeliness of stream data and the drift of stream center to combine the two models of landmark window and time attenuation. The technology of frequent pattern mining is mainly based on a dynamic system to form the overall pattern support number, and then calculate the frequency of patterns in the landmark window according to the time attenuation model. The algorithm has high mining precision, low memory cost, and can meet the requirements of high speed stream data processing, and can adapt to different number of transactions, different services and different average length of potential frequent pattern stream data mining.

2.2 Stream Data Mining Association Rule Algorithm Optimization

Stream data mining is to mine the arriving data stream according to a certain sequence, which is different from the mining of static data association rules in that the stream data is high-speed, continuous and without boundary. The unique characteristics of stream data bring a series of problems to the data mining and analysis, so as to identify the effective patterns in the mining cycle more quickly and efficiently. In order to improve the computational efficiency, 1 frequent itemset is proposed to approximate the difference between two datasets. The following formulas are given.

$$\text{error}(D, S) = \frac{|L_1(S) - L_1(D)| + |L_1(D) - L_1(S)|}{|L_1(S)| + |L_1(D)|} \tag{1}$$

In the above algorithm, L data feature set, S data difference, D data interference coefficient. The algorithm for further calculating the degree of difference caused by missing frequent itemsets is:

$$\begin{aligned} S = &(L(S) - L(D)| + |L(D) - L(S)D \\ &+ \ (MMS) - M\,(D)| + |M\,(D) - M\,(S)| \\ &- \ (0L(D) \cap M\,(S)|) + (0M\,(D) \cap L(S)D) \end{aligned} \tag{2}$$

Further optimize the process for calculating the differences in available data mining as follows:

$$\text{error}(D, S) = \frac{\sum_{i=1}^{k} |L_i(S) - L_i(D)| + |L_i(D) - L_i(S)|}{\sum_{i=1}^{k} |L_i(S)| + |L_i(D)|} \tag{3}$$

Each node can collect M kinds of attribute data in N times. This node with multiple sensing elements is called a multimode node. The data monitored by Atr, the i attribute of the sensor node, is a N-long time sequence, s = (sx), where s represents the data collected by the i attribute at j time. The raw data on the sensor is thus abstracted into a matrix:

$$A^0 = \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{M-1} \end{bmatrix} = \begin{bmatrix} s_{0,0} & s_{0,1} & \cdots & s_{0,N-1} \\ s_{1,0} & s_{1,1} & \cdots & s_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{M-1,0} & s_{M-1,1} & \cdots & s_{M-1,N-1} \end{bmatrix} \left[t_0, t_1, \cdots t_{N-1}\right] \tag{4}$$

Based on the above algorithm, the mining algorithm is re-run to extract association rules, so as to improve the mining efficiency and reduce the mining cost as far as possible under the limited system resources, and effectively scan the original data set once, and then make an incremental update with the saved results of the previous scan in the next periodic scan, and obtain the frequent itemsets near the support of the adjacent original data set to participate in the estimation of the variation degree of the two data sets, so as to determine whether it is necessary to run the mining algorithm to extract data patterns.

2.3 Implementation of Streaming Data Mining in Wireless Sensor Networks

Wireless sensor network is a kind of data network focusing on the transmission of a large amount of monitoring information. Users need data information rather than hardware devices or sensors themselves. Because of the characteristics of sensor nodes, the research concept of wireless sensor network is quite different from that of traditional information network devices only focusing on data transmission in basic design. Therefore, effective integrated management and operation of the data in the network become the core research technology for optimizing and improving the performance of wireless sensor networks. Wireless sensor networks usually focus on the monitoring data of interest, and take sensor nodes collecting data as the original stream or source of sampling data. Because all nodes build the whole network as the data transmission space or a large amount of data storage base, it can be said that the whole wireless sensor network is the bulk collation and integrated management of the sampled data. The main task and function of wireless sensor nodes are responsible for the sensing, internal storage, interrogation and data mining of the monitoring data, and separate the logical sensing map of the collected data in the monitoring environment from the physical reality of the network, and give the logical transmission structure of the query to the users. Of course, the effective management and processing of the network data from the beginning to the end of the integrated processing of the network through the whole network, the following steps are needed to improve the efficiency of the overall network management of the sensor network. The following steps are needed to consider the overall data (Fig. 3).

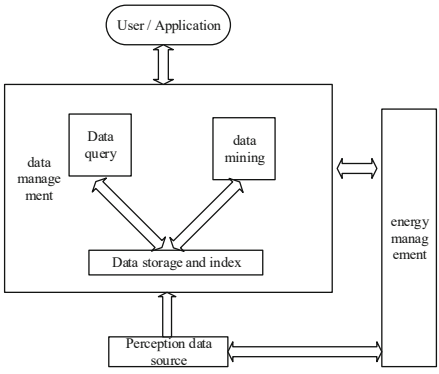

Fig. 3. Data management architecture

In the process of data management, we should consider the transmission path, redundant data and query optimization. Generally, the omni-directional management of data should include obtaining effective information, storing effective data, querying a large amount of stored information, mining deep data and the whole system management technology. The data management of wireless sensor networks is mainly located in the network layer and application layer in the longitudinal network architecture. The main task of the network layer is to provide sampling data and process the original data, and send the final data results to the application layer. Data mining technology is the key part

of all data processing. The intuitive definition is to extract or mine the useful "knowledge" from the redundant and repetitive sampled data, and use intelligent method to extract data model. Because the energy of sensor nodes and the whole sensor network is limited, we need to process a large number of data in the process of data transfer. Compression of the data at the sensor nodes and transmission of the compressed results can reduce the communication volume of the sensor network and prolong the lifetime of the network (Fig. 4).

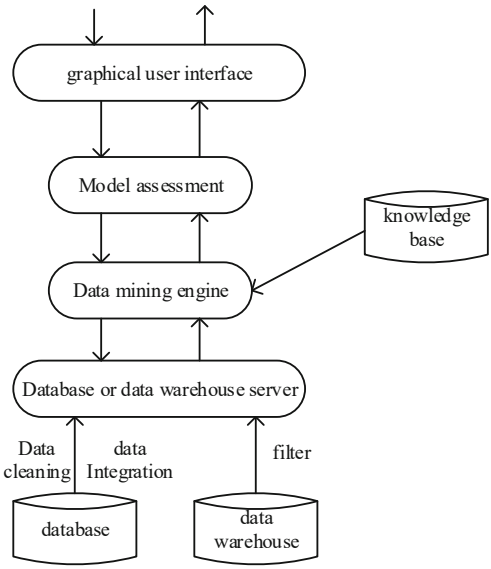

Fig. 4. Optimization of flow data mining steps in wireless sensor networks

We can define perceptual data as a relational database because of the temporal or spatial correlation of perceptual data collected in wireless sensor networks. Then we can mine the data in this relational database, and data mining is to mine the meaningful data contained in many information. Because there is a certain degree of correlation between the data, such as sampled data in a similar or identical point in time, the correlation can be eliminated to the greatest extent by some transformations; but after the adoption of some transformations, the loss of the original data and the error between the predicted data and the original data may sometimes occur, and such compression is called lossy compression. At present, there are many methods to compress the data, but the essence of which is reversible lossless compression and irreversible lossy compression. Data compression is at the cost of certain quality loss, and the quality loss is within the range of error allowed by the condition (Fig. 5).

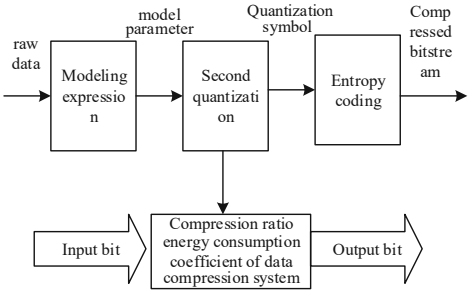

Fig. 5. Data compression processing step optimization

For the correlation of observation data, the approximate data can be obtained by constructing a suitable mathematical model of time series, so that the data amount is less than that of the original time series. Set the sampling data as:

$$\xi = ((t_1, d_1), (t_2, d_2), \mathrm{L}, (t_u, d_w)) \tag{5}$$

Secondly, the recursive curve of data feature fitting is given. The function takes t and d in the sampling data sequence as independent variable and dependent variable respectively.

$$d = \alpha + \beta t + \xi, \xi \sim \left(0, \delta^2\right) \tag{6}$$

The least square method is used to fit the above data features linearly, and the data stream features are estimated as follows:

$$\begin{cases} \alpha' = 1/n \sum_{i=1}^{n} d_i - \left(1/n \sum_{i=1}^{n} t_i\right)\beta' \\ \beta' = \left[\sum_{i=1}^{n} t_i d_i - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)\left(\sum_{i=1}^{n} d_i\right)\right] / \left[\sum_{i=1}^{n} t_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)^2\right] \end{cases} \tag{7}$$

Further, the characteristic regression equations of data are obtained.

$$\hat{d} = \alpha' + \beta' t \tag{8}$$

In order to solve the problem of energy consumption in data transmission, an effective data stream management framework is proposed, which aims to process data from different types of systems, aggregate any different and abnormal data streams, and further mine the anomaly feature data of data nodes. According to the initial cluster population value M collected from the data model, the individual characteristics of the optimal population are judged. If the number of cluster features collected is j, the dynamic adjustment range numerical algorithm for abnormal data is:

$$\zeta = \hat{d}\lambda \bigcap j - 1/(U - M) \tag{9}$$

Based on the above algorithm, the difference node numeric operators of anomaly data are globally optimized and clustered. In order to speed up the convergence rate of data

mining, it is necessary to carry out iterative processing of the above algorithms. Because the automatic data node anomaly mining method is relatively complex and a complete process, the process of data mining is usually relatively cumbersome, time-consuming and prone to bias. Therefore, the process of data mining of anomaly characteristics is optimized as follows (Fig. 6):

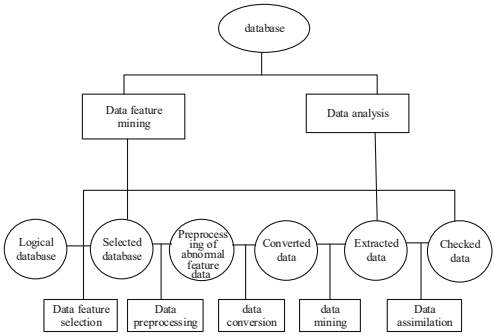

Fig. 6. Abnormal feature data mining procedure

In order to reduce the complexity of data mining to the greatest extent, the steps of anomaly feature data mining are further improved as follows:

1. Randomly collect the characteristic values of data nodes of the security resource pool.
2. Mining the internal and external characteristic data information of the collected data.
3. Further analyzing and transforming the characteristic data, establishing corresponding analysis models, and establishing corresponding SDN data centers.
4. Combine the feature data and SDN data center detection values to achieve automatic mining, obtain and output the mined data feature values.
5. The error Min value is given, then the data meeting the Min value is approximately compressed, and finally the compressed time series is decompressed to obtain the approximate prediction data, thus prolonging the network life cycle and realizing the data feature mining goal.

3 Analysis of Experimental Results

Using Matlab 7.0 as the simulation platform for simulation analysis. In order to verify the practicability of real-time stream data mining in wireless sensor networks based on edge computing, the following comparative experiments are designed. Two computers with the same configuration were used as the experimental object, in which the experimental group was loaded with deep mining method and the control group was loaded with SDN marking principle. A network device with high running stability is selected as the monitoring subject, and the changes of the influence parameters of the experimental group and the control group are recorded respectively.

In order to ensure the rationality of the experimental results, the experimental environment and parameters are set.

The hardware running environment for the experiment was Pentium- Core R980 @ 2.3 GHz dual core CPU, 64 GB memory hard disk.

Experimental software environment: Windows - XP VC+ system, C++ language.

The data of wireless sensor network is collected randomly, and the object is 4200. In order to ensure the efficiency of the experiment, it is always in the range of 18.24–20.04. In this case, the exception data is categorized and analyzed as follows (Table 2):

Table 2. Characteristics of streaming data in wireless sensor networks

Attribute category	Name
SOL	Back, land, neptune, smurf
STN	ftp_write, guess_passwd, imap, multi
R2L	Lpsweep, namp, portsweehop, phf
ROTP	overflow, loadmodule, perl, spy

In the above experimental environment, the experimental parameters are further standardized. Randomly selected feature mining data shall be set up, and the local density characteristic values, sparse characteristic values, support intrusion data flow and other relevant standard parameters of abnormal data shall be regulated, as shown in the following table (Table 3):

Table 3. Data stream standard parameters

Hierarchy	Characteristic value of local density	Support	Pattern (%)	Numerical simulation of sparsity characteristics
A	0.6	5	{d}	18
B	0.9	1	{e, f}	12
C	0.4	2	{a, c}	35
A	0.7	8	{l, e, k}	19
B	0.5	4	{d, l}	25
C	0.8	3	{d, g, t}	34

In the above experimental parameters and environment compared with the traditional method and the actual effect of this method, and record the detection results for subsequent analysis and research. By human intervention, the method of large data mining is changed, and the detailed picture of the change of experimental parameters is drawn according to the form of the specified parameters in the control host. Take 60 min as monitoring time, record the change of the total amount of data mining in big data node after using the method of experiment group and control group respectively (Fig. 7).

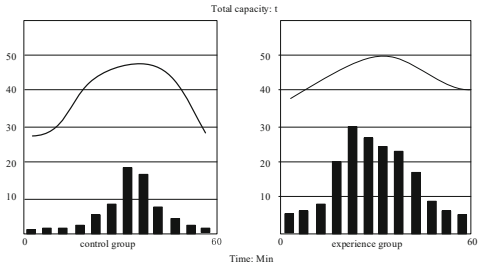

Fig. 7. Data mining load aggregate comparison chart

In the above figure, the actual value level of the column segment is used for data mining, while the curve segment only reflects the basic trend of the physical quantity. The total amount of data mining in the experimental group and the control group increased first and then decreased, but the extreme value of the experimental group was close to 30T, far higher than the extreme value of 19T. Therefore, the data mining effect of this method is better, and more data information can be mined in the same time. In summary, with the application of deep mining method, the total amount of node organization data mining does appear to be on the rise. Furthermore, it reflects the changes of the transmission rate of big data matching paths in the experimental group and the control group during the monitoring time of 60 min (Table 4).

Table 4. Experimental group data mining rates

Monitoring time /(min)	Transmission speed /(T/s)	Average value /(T/s)	Changing trend
5	9.6	10.7	
10	9.9		Rise
15	10.3		Rise
20	10.7		Rise
25	11.2		Rise
30	11.8		Rise
35	11.8		Stable
40	11.5		Decline
45	11.0		Decline
50	10.6		Decline
55	10.2		Decline
60	9.8		Decline

Analysis table shows that with the increase of monitoring time, the data mining rate of the experimental group keeps increasing, stable and decreasing trend, the maximum level reaches 11.8 T/s, and can keep steady state for 5 min (Table 5).

Table 5. Data mining rate of control group

Monitoring time/(min)	Transmission speed/(T/s)	Average value/(T/s)	Changing trend
5	4.2	4.9	
10	5.6		Rise
15	4.2		Decline
20	5.7		Rise
25	4.3		Decline
30	5.5		Rise
35	4.4		Decline
40	5.6		Rise
45	4.1		Decline
50	5.4		Rise
55	4.3		Decline
60	5.2		Rise

The global maximum value is only 5.7 T/s, which is much lower than the extreme value level of the experimental group, and this value can not maintain a stable state for a long time. This shows that under the application of this method, the data mining rate is higher, and the high efficiency can be maintained for a certain period of time, which has better performance than the traditional method. In summary, the edge computation-based wireless sensor network real-time stream data mining method has high practical value, data mining effect is significantly better than the traditional method.

4 Closing Remarks

In wireless sensor networks, sensor nodes will produce a large number of data, which have the characteristics of fast arrival, real-time update and so on, is a typical stream data. The application of stream data mining technology in many fields is more and more extensive, which makes communication and computer function more powerful and can provide better services for users. This paper analyzes and compares the main data mining technologies in wireless sensor networks, and proposes a workflow framework for streaming data mining in wireless sensor networks. The processing flow of WSN stream data mining is explained clearly, and the global pattern can be obtained by the deeper mining of local pattern. With the maturity of wireless sensor network technology, sensor sensing data will be increasingly rich.

References

1. Gombé, B.O., et al.: A SAW wireless sensor network platform for industrial predictive maintenance. J. Intell. Manuf. **30**(4), 1617–1628 (2017). https://doi.org/10.1007/s10845-017-1344-0
2. Lorenz, P., Schott, R., et al.: New path centrality based on operator calculus approach for wireless sensor network deployment. IEEE Trans. Emerg. Top. Comput. **7**(1), 162–173 (2019)
3. Babu, R.G., Karthika, P., Manikandan, G.: Polynomial equation based localization and recognition intelligent vehicles axis using wireless sensor in MANET. Procedia Comput. Sci. **167**(10), 1281–1290 (2020)
4. Gupta, G.P., Jha, S.: Biogeography-based optimization scheme for solving the coverage and connected node placement problem for wireless sensor networks. Wirel. Netw. **25**(6), 3167–3177 (2019)
5. Dugaev, D., Peng, Z., Luo, Y., et al.: Reinforcement-learning based dynamic transmission range adjustment in medium access control for underwater wireless sensor networks. Electronics **9**(10), 1727 (2020)
6. Wang, H., Tan, X., Huang, Z., et al.: Mining incomplete clinical data for the early assessment of Kawasaki disease based on feature clustering and convolutional neural networks. Artif. Intell. Med. **105**(4), 101859 (2020)
7. Jeihouni, M., Alavipanah, S.K., Toomanian, A., et al.: Digital mapping of soil moisture retention properties using solely satellite-based data and data mining techniques. J. Hydrol. **585**(5), 124786–124788 (2020)
8. Maind, A., Raut, S.: Mining conditions specific hub genes from RNA-Seq gene-expression data via biclustering and their application to drug discovery. IET Syst. Biol. **13**(4), 194–203 (2019)
9. Jimenez-Carvelo, A.M., Gonzalez-Casado, A., Gracia Bagur-Gonzalez, M., et al.: Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – a review. Food Res. Int. **122**(AUG), 25–39 (2019)
10. Ma, X., Ji, Y., et al.: Multidimensional visualization of Bikeshare travel patterns using a visual data mining technique: data cubes. J. Beijing Inst. Technol. **28**(2), 79–91 (2019)
11. Liu, S., Liu, D., Srivastava, G., Połap, D., Woźniak, M.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. **7**(4), 1895–1917 (2020). https://doi.org/10.1007/s40747-020-00161-4
12. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
13. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)