Acta Technica *62* No. 2B/2017, 85–94

Design and implementation of improved distributed data mining algorithm

Ang Li[1]

Abstract. Faced with the rapid growth of data, the traditional data mining algorithm is not capable of dealing with this issue. Although the existing data mining algorithm has taken into account the distributed operation, it does not apply to wireless sensor networks. Therefore, a method of data mining algorithm for wireless sensor networks is proposed. The design and implementation of improved distributed data mining algorithm DDMAW is described, and its performance is analyzed. Based on the performance comparison of finding events algorithm based on distributed K-means with the improved distributed algorithm DDMAW. After that, the DDMAW algorithm itself is discussed in parallel. The comparison results show that DDMAW algorithm has the characteristics of low time complexity and high accuracy, and it is more suitable for use in wireless sensor networks.

Key words. Wireless sensor network, data mining, distributed.

1. Introduction

In recent years, the research on all aspects of wireless sensor network has increased, and the research on distributed clustering algorithm and anomaly detection algorithm is growing to be more. But applying the traditional, centralized data mining algorithm to the wireless sensor network environment is not feasible [1]. The traditional data mining algorithm processes data in a unified and centralized way, but not the use of distributed processing method. If a large amount of data is processed in a centralized data analysis algorithm, it is possible to cause significant traffic and computational complexity, and cost a lot of time, which rapidly consumes the energy of the sensor and increase the overhead, which leads to the reduction of the sensor life. The wireless sensor network technology and distributed data mining technology should be dynamically integrated [2].

[1]Nanjing University of science & Technology Zijin College, Nanjing, 210023, China

2. Design and implementation of distributed clustering algorithm

2.1. The algorithm overview

As the basic algorithm of data mining in wireless sensor networks, it is accordingly improved to be centralized K-means algorithm. The improvement can make it can apply to the distributed environment [3]. The K-Dmeans algorithm solves the parallel problem, but there are still problems that will delete part of the point sets. For wireless sensor networks, the above method may lose some data points, resulting in the loss of observed events [4]. However, it is feasible to divide the detection area into grid.

2.2. The improved DMAW algorithm

In the wireless sensor network, the normal working sensor nodes are able to receive data. But only a part of the sensor nodes can detect the occurrence of the event. Therefore, in the process of data mining, there is only need to do the clustering analysis of selected point set including events rather than each data [5]. And the selected data, not only contains the observed data, but also should be attached with the data information of the geographical coordinates. After selecting the point sets containing events, the geographical position range of the point set is delineated to determine the area of the event. Within this range, the event and the location of occurrence can be obtained through the data mining algorithm [6]. The whole algorithm is divided into two parts. The first is the data processing, and the other part is the determination of the K value, finding clustering centroids, returning the observed value and the corresponding position coordinates.

Before data mining, the data is pre-processed so that the observed data is in an effective detection environment. Then SSE is calculated starting from $K = 1$ iteration. If the threshold is exceeded, iteration will continue. Otherwise stop, output K value, and event center point coordinates.

The time complexity of the grid-based algorithm is mainly the time complexity of the assignment point set to the grid cell, and the time complexity of deleting the sparse grid and induction [7]. Therefore, the time complexity based on the grid algorithm is $O(N)$. The time complexity of the improved algorithm—DMAW algorithm is approximately $O(N)$. And the time complexity based on the grid algorithm is also $O(N)$. That is to say, the time complexity of the DMAW algorithm is basically the same as the time complexity of the grid algorithm. The communication complexity of DMAW algorithm is also analyzed. The communication complexity of the DMAW algorithm is denoted as: T_{M}. According to the description of the DMAW algorithm, it can be assumed that the traffic is generated in the following parts.

1. Scan data points and finds data that exceeds the normal values
2. Determine the effective detection area
3. Iterate starting from $K = 1$ for the preprocessed data

Scan the data points and find out the data exceeds the normal value, and operate this on N points respectively. The communication complexity of this part is

$$T_{\mathrm{M1}} = N \times (T_{\mathrm{D}} + T_{\mathrm{C}}) . \tag{1}$$

Next, the process of determining the effective detection area is to compare the values of x and y, respectively. The communication complexity required for this process is:

$$T_{\mathrm{M2}} = N_{\mathrm{a}} \times (\log N_{\mathrm{a}}) . \tag{2}$$

Then, iterate the processed data.

The first traversal communication complexity is

$$T_{\mathrm{M31}} = T_{\mathrm{D}} . \tag{3}$$

The second traversal communication complexity is

$$T_{\mathrm{M32}} = 2 \times T_{\mathrm{D}} . \tag{4}$$

The Kth traversal communication complexity is

$$T_{\mathrm{M3K}} = K \times T_{\mathrm{D}} . \tag{5}$$

Therefore, the communication complexity of the iteration process is

$$T_{\mathrm{M3}} = (1 + 2 + \cdots K) \times T_{\mathrm{D}} + T_{\mathrm{C}} . \tag{6}$$

Based on the analysis above, the communication complexity of DMAW algorithm is

$$T_{\mathrm{M}} = T_{\mathrm{M1}} + T_{\mathrm{M2}} + T_{\mathrm{M3}} . \tag{7}$$

Namely,

$$T_{\mathrm{M}} = (N + N_{\mathrm{a}} \log N_a + K(K+1)/2) T_{\mathrm{D}} . \tag{8}$$

Since the connection setup time is much smaller than the actual communication time, the communication complexity of the DMAW algorithm can be approximated by formula (8).

The data set points in the whole detection area are 1000, 2000, 4000, 8000, 16000, 32000, respectively. The number of events in each detection area is 2, SNR is 10 dB. The analysis and discussion is undergone through the computation time and data mining effect of DMAW algorithm. The experimental results show that the computation time of DMAW algorithm is proportional to the number of nodes in the case of different nodes. The corresponding DMAW algorithm computing time at 16000 points is 146.767 milliseconds. And the corresponding DMAW algorithm computing time is 350.784 milliseconds at 32000 points. That the time complexity of the DMAW algorithm is $O(N)$, it can be obtained in the case of different numbers of nodes.

2.3. The improved DDMAW algorithm

The centralized algorithm can only be applied to local parts as for distributed environments [7]. As for the distributed, not only the local processing of data is necessary, but also it needs to consider how to improve in the overall situation [8]. In view of this problem, this paper proposes DDMAW algorithm. The core idea of the DDMAW algorithm is the parallel implementation of DMAW algorithm in each detection area, and then fuses data when the local model is aggregated to the global level.

The idea of data fusion is listed as follows:

1. To determine whether there exists a value v falling on the detection area boundary or the coordinates of the two events are close to each other.
2. For both centroids, find the effective detection area before the calculation.
3. Integrate the effective detection areas corresponding to the two centroids.
4. Implement the data mining in the new area.
5. Return the new centroids of events.

The DDMAW algorithm is based on the improvement of DMAW algorithm. The DMAW algorithm is used for local models. Multiple detection areas simultaneously use the DMAW algorithm to work in parallel to form a distributed underlying architecture. Each detection area reports its own event centroid after computation, and then on the basis of this to implement the data fusion. After the integration, a new event center is created [9].

As for DDMAW algorithm, it is analyzed by means of time complexity and communication complexity. It is assumed that the time complexity of the DDMAW algorithm is T_{D}. The communication complexity of DDMAW algorithm is T_{C}. The number of data points per detection area is N. Suppose the number of detection areas is M. The computation time is mainly composed of the parallel data mining time and the time of data fusion to the data reintegration. Thus, the time complexity of the DDMAW algorithm should be $O(N)$.

The communication complexity of DDMAW algorithm is discussed hereby. Assuming that T_{D} is the data communication time, and T_{C} is the connection setup time, the communication complexity of the DMAW algorithm is given by the formula (8) mentioned above.

Since $\ll N$, and $N_a \log N_a \ll N$, the communication complexity of the DMAW algorithm can be approximated as: $T_M = N \times T_{\mathrm{D}}$. Considering that there are four detection areas, then $T_{\mathrm{MM}} = N \times M \times T_{\mathrm{D}}$.

In the process of transmitting $K_1 + K_2 + \cdots + K_M$ event centroids to the upper level, the needed communication complexity is $(K_1 + K_2 + \cdots + K_M) \times T_{\mathrm{C}}$ which is negligible.

Extract the two effective detection areas for the event. Since the extracted area is given in the parallel execution of the DMAW algorithm, and the comparison is for

the values of the x and y values of the region. Thus, the communication complexity of this part is approximately $4T_{\mathrm{D}}$. Calculate the communication complexity for the new effective detection area event computation center. The DMAW algorithm is still adopted here. So, the communication complexity of this part of is $T_{\mathrm{M}} = N \times T_{\mathrm{D}}$. Based on what is discussed above, the communication complexity of DDMAW algorithm is $T_{\mathrm{M}} = (N \times M + 5) \times T_{\mathrm{D}}$. Since $N \times M \gg 5$, the communication complexity of the DDMAW algorithm can be approximated as $T_{\mathrm{M}} = N \times M \times T_{\mathrm{D}}$.

The data set points of all the detection area are 1000, 2000, 4000, 8000, 16000, 32000. The number of events of each detection area is 6, SNR is 10 dB. The analysis and discussion is undergone through the computation time and data mining effect of DMAW algorithm. Experiments show that in the detection area, you can accurately find its centroids as for the occurrence of specific events, and the event falls on the edge of the detection area. The DDMAW algorithm fuses the data for this event and other events may occur in the adjacent area according to its principle.

1. The number of nodes in the detection area is different, the number of events is the same and the noise condition is the same. The DDMAW algorithm can accurately find out each event and its centroid point in the case of the variation of the number of nodes in the detection area. Therefore, the DDMAW algorithm can also be used if the detection area node is small. The running time of DDMAW algorithm is basically linear with the number of nodes in the detection area. And the computation time of the algorithm increases with the increase of the number of nodes in the detection area.

2. The number of nodes in the detection area is different, the number of events is the same and the noise condition is the same. The DDMAW algorithm can accurately determine the number of event, its time and position in the case of the variation of the number of events in the detection area. Therefore, the DDMAW algorithm can apply to the case of single event occurrence and multi-events occurrence. The running time of DDMAW algorithm is basically linear with the number of events in the detection area. And the computation time of the algorithm increases with the increase of the number of nodes in the detection area.

3. The number of nodes in the detection area is the same, the number of events is the same and the noise condition is different. The negative value of SNR is regarded that the noise power exceeds the signal power. The DDMAW algorithm set a threshold for excluding most of the noises in data processing. The result of finding the centroids of event points may drift to some degree due to the noise disturbance, but the disturbance is little under the condition of a certain SNR. The event could be accurately found out when the SNR value is negative.

3. Comparison of the DMAW algorithm and other algorithms

3.1. Performance analysis and comparison

The time complexity of the iteration is $O(KTN)$ through analyzing the time complexity of K-means. Since $K, T \ll N$, so the time complexity of this part is approximately $O(N)$. The spatial complexity of K-means algorithm is $O((K+N)m)$ through analyzing the spatial complexity of K-means algorithm. Analyze the communication complexity of K-means algorithm, and the communication complexity of K-means algorithm is

$$T_{\mathrm{K}} = T \times N \times K \times T_{\mathrm{D}} \,. \tag{9}$$

The time complexity of DMAW algorithm is $O(N)$, and the communication complexity of DMAW algorithm is given by formula (9).

3.2. Experimental analysis and comparison

The method to find the event based on the K-means is the analysis and computation of, all the data points, including the points of the normal circumstances. Therefore, there is a certain error in the determination of the event and the event centroids. The DMAW algorithm is more accurate in finding events, since it only processes the event area and excludes the redundant data in the sensor network.

This paper also explores the time required to find an event based on the K-means. The time required to find the event based on K-means is analyzed through MATLAB simulation. The DMAW algorithm runs faster because it only processes the event occurrence area. And the number of events in the DMAW algorithm is determined by iteration, which greatly improves the accuracy of determining the number of events. In addition, the data mining of the effective detection area of the event can accurately determine the event position.

Analyze the cases that the number of nodes of each detection area is 1000, 2000, 4000, 8000, 16000, and 32000, respectively. Assume that there is one event in each detection area, and the SNR is 10 dB. The required time comparison is shown in Fig. 1.

4. The comparison of the DMAW algorithm and other algorithms

4.1. Performance analysis and comparison

The time complexity and communication complexity of K-Dmeans are analyzed, and compared with DDMAW algorithm.

The time complexity of K-Dmeans algorithm is $O(N^2)$. Then the communication complexity of K-Dmeans algorithm is analyzed, and the result is $O(N^2)$.

The time complexity of DDMAW algorithm is $O(N)$. The time complexity of the distributed algorithm based on K-means is much larger than that of the DDMAW

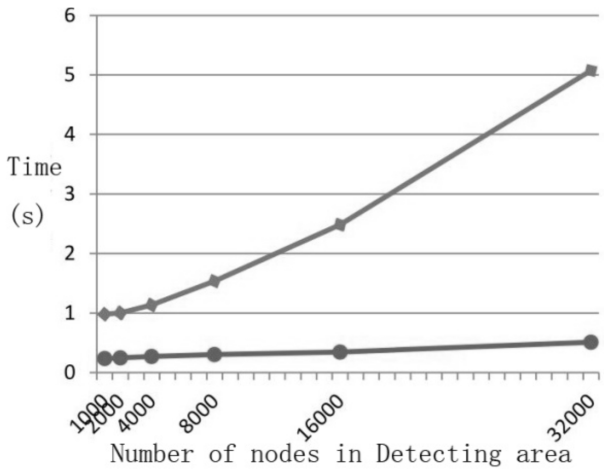

Fig. 1. Required time comparison of the two algorithms

algorithm. Therefore, the DDMAW algorithm performs better than the distributed K-means algorithm theoretically. The communication complexity of the K-Dmeans algorithm and the DDMAW algorithm are $O(N^2)$ and $O(N)$ respectively. Therefore, the DDMAW algorithm is superior to the distributed K-means algorithm in communication complexity.

4.2. Experimental analysis and comparison

The event centroids based on the K-means event finding algorithm is not consistent with the pre-assumed event in the detection area. In the same situation, DDMAW algorithm can accurately find all six events. Therefore, the accuracy of the DDMAW algorithm is better than that of the K-means in terms of the algorithm effect. For the computation time, the time required for the event finding algorithm based on K-means and the DDMAW algorithm both increases with the number of nodes in the four parallel areas. The time required for the event finding algorithm based on K-means is much higher than that of the DDMAW algorithm. The time required to find an event based on K-means is significantly higher than that of the DDMAW algorithm. The computation time for the event finding algorithm based on K-means and the DDMAW algorithm both increases with the number of nodes in the four parallel areas in the case of 8 parallel areas. Moreover, the time required of event finding algorithm based on K-means is higher than that of the DDMAW algorithm. Therefore, the event finding algorithm based on K-means has the drawbacks in determining the number of events, and the DDMAW algorithm can find out the number and the centroids of the event accurately.

4.3. The comparison of self- performance of DDMAW algorithm

The DDMAW algorithm is improved based on the DMAW algorithm. The DMAW algorithm and the DDMAW algorithm are compared, and the experiment adopts the man-made data sets. The number of data set points in all detection areas are 100, 1000, 4000, 8000, 16000, and 32000, respectively. The computation time and data mining results of the DDMAW algorithm parallel in 4 areas and the DDMAW algorithm parallel in 8 areas are analyzed and discussed. Assuming that there are 1000 nodes in the local area, and two events in the area, namely (4, 5), (16, 15), and the SNR is 10 dB. In the case of four areas in parallel, the DDMAW algorithm can still accurately find the location of the event.

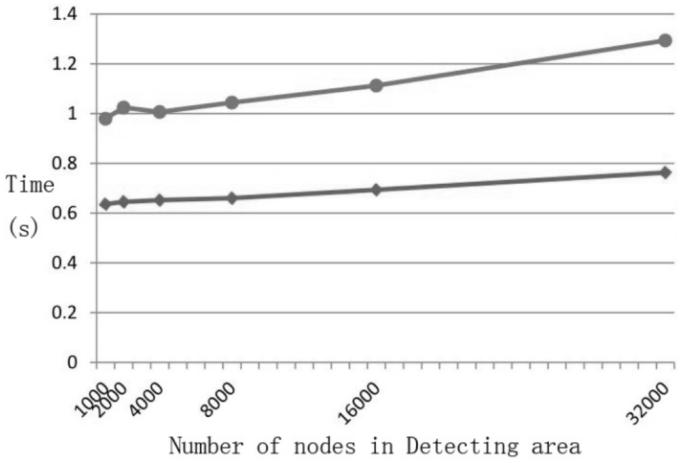

Fig. 2. Average required time of each area in different cases of parallel area number

Figure 2 compares the average time for each area of the four parallel areas and eight parallel areas respectively. It can be seen from the figure that the average time spent on each detection area in the four parallel areas (dots) is slightly higher than that of 8 parallel areas (square points). Among them, in the case of four parallel areas, there is a higher value when the data set point number is 2000. To explore the reasons, it is possible that in the four 2000-points parallel areas, the number of event occurrence is relatively more, therefore the computation time is relatively long. In the case of more parallel areas, the average computing time of DDMAW algorithm in each detection area is less adversely. This shows that the DDMAW algorithm efficiency is higher in the cases of relative more parallel areas.

5. Conclusion

In a wireless sensor network environment, the collected data is distributed across many nodes. If the data of these nodes are processed uniformly, the amount of data will be extremely enormous. And the large amount of data will lead to a

long computation time, which lacks timeliness in data application. Therefore, the distributed data mining technology is applied to wireless sensor networks, which can improve the computation efficiency and effectively find useful information. As the fact that it is not practical to apply the distributed data algorithm directly to the wireless sensor network for actual situation, the relevant data mining algorithm is improved, making it suitable for wireless sensor network structure.

Firstly, a new distributed model for wireless sensor networks is proposed according to the principle of distributed data mining algorithm. The model is divided into multiple parallel detection areas. There are several nodes in each detection area. The data of each node in the work is collected to one of the nodes. In the parallel process, each node aggregates the data to the cluster node. Then the cluster nodes fuse the data of the information. Then, an improved DMAW algorithm is proposed. The algorithm firstly determines the area of events. Secondly, it determines the number of events in the area and find the centroid of the event. At the same time, a distributed DDMAW algorithm is proposed based on the DMAW algorithm. The results show that the comprehensive performance of the DMAW algorithm and the DDMAW algorithm is better.

As for the distributed data mining technology in the sensor network in this paper, the algorithm design idea of the algorithm determines that the algorithm can accurately determine the number of events and find out the centroids of the events. Moreover, the algorithm time is superior to other data mining technology. However, considering the abnormal points in the experimental analysis, it is obvious that there may exist other factors impacting the running time of algorithm except for the number of data set point. The next work will specifically explore other factors that affect the running efficiency of the DMAW algorithm and the DDMAW algorithm.

References

[1] H. S. Shi, Y. X. Li, S. P. Zhang: *An energy-efficient MAC protocol for wireless sensor network*. Proc. IEEE International Conference on Advanced Computer Theory and Engineering(ICACTE), 20–22 August 2010, Chengdu, China, IEEE Conference Publications *4* (2010), V4-619–V4-623.

[2] K. Sohrabi, J. Gao, V. Ailawadhi, G. J. Pottie: *Protocols for self-organization of a wireless sensor network*. IEEE Personal Communications *7* (2000), No. 5, 16–27.

[3] J. Paek, K. Chintalapudi, R. Govindan, J. Caffrey, S. Masri: *A wireless sensor network for structural health monitoring: Performance and experience*. IEEE Workshop on Embedded Networked Sensors(EmNetS-II), 30–31 May 2005, Sydney, Queensland, Australia, IEEE Conference Publications (2005), 1–10.

[4] G. Werner-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees, M. Welsh: *Deploying a wireless sensor network on an active volcano*. IEEE Internet Computing *10* (2006), No. 2, 18–25.

[5] A. Arora, P. Dutta, S. Bapat, V. Kulathumani, H. Zhang, V. Naik, V. Mittal, H. Cao, M. Demirbas, M. Gouda, Y. Choi, T. Herman, S. Kulkarni, U. Arumugam, M. Nesterenko, A. Vora, M. Miyashita: *A line in the sand: A wireless sensor network for target detection, classification, and tracking*. Computer Networks *46* (2004), No. 5, 605–634.

[6] G. Werner-Allen, J. Johnson, M. Ruiz, J. Lees, M. Welsh: *Monitoring volcanic eruptions with a wireless sensor network*. IEEE Proceeedings of the Second European

Workshop on Wireless Sensor Networks, 31 Januar–2 Februar 2005, Istanbul, Turkey, IEEE Conference Publications (2005), 108–120.

[7] I. H. Witten, E. Frank: *Data mining: Pratical machine learning tools and techniques, second edition.* Morgan Kaufmann, San Francisco, CA, USA, (2005).

[8] T. Hastie, R. Tibshirani, J. H. Friedman: *The elements of statistical learning: Data mining, inference, and prediction, second edition.* Springer Series in Statistics, New York, USA (2016).

[9] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, A. M. Chinnaiyan: *ONCOMINE: A cancer microarray database and integrated data-mining platform.* Neoplasia *6* (2004), No. 1, 1–6.

Received August 7, 2017