Entropy based Correlation Clustering for Wireless Sensor Network in Multi-Correlated Regional Environment

Nguyen Thi Thanh Nga, Nguyen Kim Khanh, Son Ngo Hong
School of Information and Communication Technology,
Hanoi University of Science and Technology
Hanoi, Vietnam
ngantt, khanhnk, sonnh@soict.hust.edu.vn

***Abstract*— The existence of correlation characteristics bring significant potential advantages for the development of efficient routing protocols in wireless sensor network. However, there has not an effective approach to divide sensor nodes into correlation group in multiple correlation-areas. This research proposed a new simple method on evaluating joint entropy of multiple sensed data to identify multi-correlation areas. The definition of correlation region based on entropy theory is also proposed. Following, a correlation clustering scheme with less computation is developed using proposed method and definition. The results are validated with real data set.**

Keywords*— *Entropy; Correlation Clustering; Entropy Correlation Coefficient

I. Introduction

Wireless Sensor Networks (WSNs) are simple low-cost approaches that can be used in distributed environment. In recent years, a considerable number of published researches on WSNs have deal with the issue of energy conservation [1]. Among these works, clustering is a well-established technique for reducing data collection costs in WSNs [2]. This class of WSNs is potentially viewed as the most energy-efficient and long-lived class of sensor networks [3]. However, these routing protocols do not yet consider the characteristics of environmental attributes.

Typical WSNs applications require spatially dense sensor deployment in order to achieve satisfactory coverage [1]. As a result, multiple sensors record information about a single event in the sensor field, i.e. these sensed data has correlation with each other. The existence of correlation characteristic brings significant potential advantages for the development of efficient communication protocols well-suited for the WSNs paradigm. There have been some research efforts to study the correlation in WSNs. In [4, 5], a theoretical framework to model the spatial and temporal correlations in sensor networks was developed. Following, some approaches to exploit spatial and temporal correlation for efficient medium access and reliable event transport in WSNs were proposed. In those papers, the correlation model was based on the assumption that all nodes in a sensor field observed the same physical phenomenon but with noises. The correlation coefficient was chosen to be a function of distance between nodes. However, this correlation coefficient measured only the linear correlation and for scalar data. In [6], the correlation characteristics for visual information was studied and entropy correlation coefficient was used. This correlation coefficient is more general than the previous correlation coefficient and directly related to amount of information transferred in the network. However, the entropy correlation coefficient is also assumed to be a function of sensing position and direction. Additionally, both researches considered only one correlation region, i.e. all nodes observed the same phenomenon. Therefore, the problem of correlation grouping/clustering has not been considered yet. Grouping/clustering of correlated nodes has been considered in the field of machine learning/data mining [7], and the results were also used in WSNs [8]. However, the meaning of correlation here simply is the similarity between two sets of data. Entropy concept is applied only to verify the effectiveness of the proposed correlation clustering.

Entropy/joint entropy concept is more general to describe the correlation among sets of data, especially in some cases, the correlation is not distance/location dependence [9, 10]. Therefore, in some researches, entropy theory was used for correlation clustering in WSNs [9, 11]. In these researches, correlation region was defined based on the increased amount of joint entropy when the number of calculated nodes increases. This definition has some problems such as it requires vast amount of computation time, because there are an enormous number of combinations of picking sensor nodes to calculate the joint entropy. Additionally, the correlation level has not been clarified yet.

In order to overcome these above difficulties, this research presents a correlation clustering scheme with less computation, based on entropy concept. At first, the joint entropy of group of nodes is evaluated based on the entropy of single nodes and entropy correlation coefficient of a pair of nodes. Then, the definition of correlation region is proposed continuously. Based on the proposed definition of correlation region, the correlation clustering scheme is proposed.

II. ENTROPY CORRELATION COEFFICIENT

In order to measure the correlation among sets of data, the concept of entropy and mutual information [12] is considered.

The entropy *H(X)* of the random variable X with probability distribution *P(X)* is:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \quad (1)$$

If random variables *X* and *Y* are jointly distributed according to *P(x, y)*, then joint entropy *H(X, Y)* is:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 P(x,y) \quad (2)$$

with equality if *X* and *Y* are independent.

Mutual information is a quantity that measures a relationship between two random variables which are sampled simultaneously and is given by:

$$I(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \quad (3)$$

The relation between mutual information and entropy is given by:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (4)$$

However, it is difficult to compare the correlation level between two pairs of random variables using mutual information or joint entropy, because their values depend on entropy of each individual data in the pair. To overcome this problem, we use normalized measures of mutual information called entropy correlation coefficient [13] that is given as follows:

$$\rho(X,Y) = 2\frac{I(X,Y)}{H(X)+H(Y)} = 2 - 2\frac{H(X,Y)}{H(X)+H(Y)} \quad (5)$$

The entropy correlation coefficient ρ varies from 0 to 1, depending on the correlation between two nodes. The larger the value of ρ, the higher the correlation is.

III. JOINT ENTROPY ESTIMATION

In order to determine whether a group of nodes is correlated or not, it is necessary to know the entropy of each node and joint entropy of all nodes in the group. However, the direct calculation of joint entropy of group of more than two nodes are waste of time and computation resources. In this section, we try to estimate joint entropy of a group of nodes from entropies of single nodes in the group and entropy correlation coefficients of all pairs in the group.

Suppose that there is a set of *N* data $\{X_1, X_2, ..., X_N\}$ with entropy of each data, $H(X_i)$, and entropy correlation coefficient, $\rho_{ij} = \rho(X_i, X_j)$, with any $1 \leq i \neq j \leq N$ satisfies the following conditions:

$$H_{min} \leq H(X_i) \leq H_{max} \quad (6)$$

$$\rho_{min} \leq \rho_{ij} \leq \rho_{max} \quad (7)$$

The joint entropy is estimated based on the idea of hierarchical clustering [14] as follows.

A. Determination of upper bound of joint entropy

With a group that has only one node, we have entropy of one node is limited by equation (6):

$$H_1 = H(X_i) \leq k_1 H_{max} \quad \text{where } k_1 = 1 \quad (8)$$

With a group of two nodes X_i and X_j, from the definition of entropy correlation coefficient in equation (5) we have:

$$H_2 = H(X_i, X_j) = \frac{2-\rho(X_i,X_j)}{2}\big(H(X_i) + H(X_j)\big) \quad (9)$$

In addition,

$H(X_i), H(X_j) \leq H_{max}$, and $\rho(X_i, X_j) = \rho_{ij} \geq \rho_{min}$

Then

$$H_2 \leq \frac{2-\rho_{min}}{2}(2H_{max}) = (2-\rho_{min})H_{max}$$

or $H_2 \leq k_2 H_{max} = bH_{max}$ where $k_2 = b = 2 - \rho_{min}$ (10)

With a group of three nodes X_i, X_j and X_k, at first, two nodes X_i and X_j are merged to create a new cluster represented by node X_{ij} with $H(X_{ij}) = H(X_i, X_j) \leq k_2 H_{max}$. According to hierarchical clustering [14], [6], the correlation coefficient between one cluster and another cluster can be obtained by the greatest/shortest/average correlation coefficient from any member of one cluster to any member of the other cluster. Therefore, $\rho(X_{ij}, X_k) = \min\{\rho(X_i, X_k), \rho(X_j, X_k)\}\} \geq \rho_{min}$

Then,

$$H_3 = H(X_i, X_j, X_k) = H(X_{ij}, X_k) = \frac{2-\rho(X_{ij},X_k)}{2}\big(H(X_{ij}) + H(X_k)\big)$$

$$\leq \frac{2-\rho_{min}}{2}(k_2 H_{max} + H_{max}) = \frac{b}{2}(k_2+1)H_{max} = k_3 H_{max} \quad (11)$$

where $k_3 = \frac{b}{2}(k_2 + 1)$

Similarly, joint entropy H_m of a group with *m* nodes can be considered to be joint entropy of a sub-cluster with *m-1* nodes and the remaining node. The entropy of the sub-cluster is joint entropy of *m-1* nodes and the entropy correlation coefficient between the sub-cluster and the main node is greatest/shortest/average correlation coefficient from any member of the sub-cluster to the remain node. Thus,

$$H_m \leq \frac{2-\rho_{min}}{2}(k_{m-1}H_{max} + H_{max}) = \frac{b}{2}(k_{m-1}+1)H_{max}$$

$$= k_m H_{max} \quad \text{where } k_m = \frac{b}{2}(k_{m-1}+1) \quad (12)$$

From recurrence relation of k_m, the general formula to calculate k_m can be obtained as follows ($m \geq 3$):

$$k_m = 2\left(\frac{b}{2}\right)^{m-1} + \left(\frac{b}{2}\right)^{m-2} + \cdots + \left(\frac{b}{2}\right)^2 + \frac{b}{2} \quad (13)$$

or in more compact way (in case b≠2):

$$k_m = \frac{\left(\frac{b}{2}\right)^m - 1}{\frac{b}{2}-1} + \left(\frac{b}{2}\right)^{m-1} - 1 \qquad (14)$$

B. Determination of lower bound of joint entropy

Lower bound of joint entropy of a group with m node could be determined in the similar way to upper bound. The results are as follows:

$$H_m \geq l_m H_{min} \qquad (15)$$

where $l_1 = 1$, $l_2 = c = 2 - \rho_{max}$;

in case $m \geq 3$: $$l_m = \frac{\left(\frac{c}{2}\right)^m - 1}{\frac{c}{2}-1} + \left(\frac{c}{2}\right)^{m-1} - 1 \qquad (16)$$

The above proposed joint entropy estimation can be easily validated with two special cases: all nodes are completely depend on each other and all nodes are completely independent with each other. Moreover, in order to verify the above estimation of joint entropy in a practical case, we have used sample temperature data supplied by Intel Berkeley Research Lab [15]. It is found that the joint entropy of one group is always between lower bound and upper bound. These above examples show the validity of the proposed estimation.

IV. CORRELATION REGION DEFINITION AND CORRELATION CLUSTERING ALGORITHM

A. Correlation Region Definition

As mentioned in [4], sensor nodes in the same correlation region record information of a single event in the sensor field, i.e. these sensed data has correlation with each other. Because the sensed data is taken from the same event, the number of bits to represent sensed data should be not so different, i.e. the entropy of sensed data is similar. On the other hand, the entropy correlation coefficient of all pairs in this region is also similar. Therefore, we can define a correlation region as follows:

Definition 1: A correlation region is a region where the sensed data of all nodes has similar entropy value and entropy correlation coefficient between all pairs of nodes are also similar. In practice, it is difficult to obtain similarity between two entropies or entropy correlation coefficient. Then, the correlation region could be defined by a more practical way as follows.

Definition 2: A group of m nodes $\{X_1, X_2, \dots, X_m\}$ is in a correlation region if:

- $H_0 \leq H(X_1), H(X_2), \dots, H(X_m) \leq H_0 + \Delta H$
- $\rho_0 \leq \rho_{ij} = H(X_i, X_j)$, $\forall i \neq j$

where ΔH is small enough.

H_0 is called "base entropy" and ρ_0 is called "correlation level" of the region. The higher the correlation level, the more correlation of the region is.

With the first definition, the upper bound and lower bound of joint entropy is almost the same, then the joint entropy can be estimated using upper bound or lower bound function. With the second definition, the upper bound of coefficient correlation ρ_{ij} has not been limited. Therefore, the joint entropy of the m nodes $\{X_1, X_2, \dots, X_m\}$ could be estimated by the following equation of upper bound function:

$$H(X_1, X_2, \dots, X_m) = k_m H_0 \qquad (17)$$

k_m is calculated by using equation (13) or (14) with $b = 2 - \rho_0$.

B. Correlation Clustering Algorithm

Using the definition of correlation region, a sensor field could be divided into correlation regions with specified base entropy and correlation level. The clustering process is described as in Fig. 1. In the step (*) of the algorithm, the base entropy and correlation level is chosen such as they can cover all possible values of entropy and entropy correlation coefficient in the network. The value of entropy correlation coefficient should be chosen from high to low. In the step (**) of the algorithm, if there are more than one node satisfy the condition $0 < C(X_i) = \max\{C(X_j), X_j \in G\}$, the node that has maximum entropy value will be removed.

```
BEGIN
REPEAT
    Choose $H_0$, $\rho_0$, $\Delta H$; (*)
    Initialize new group $G = \varnothing$;
    FOR each node $X_i$ in the network and not belong to any group
        IF $H_0 \leq H(X_i) \leq H_0 + \Delta H$
            Add $X_i$ into $G$
        ENDIF
    ENDFOR
    REPEAT
        FOR each node $X_i$ in $G$
            Calculate $C(X_i)$= number of node $X_j$ that $H(X_i, X_j) < \rho_0$
        ENDFOR
        FOR each node in G
            IF $0 < C(X_i) = max\{C(X_j), X_j \in G\}$
                    Remove $X_i$ from G (**)
            ENDIF
        ENDFOR
    UNTIL $max\{C(X_j), X_j \in G\}=0$
UNTIL all nodes are grouped
END
```

Fig.1 Correlation-based clustering algorithm

In order to verify the definition and clustering algorithm, we re-consider the data in [15]. The entropy is in a range from *2.70* to *2.90*, i.e. $H_0 = 2.70$; $\Delta H = 0.2$. The entropy correlation coefficients are all larger or equal to *0.60*, i.e. $\rho_0 = 0.60$. Fig. 2 shows the data of 11 nodes in the group. In Fig. 2, it is found that all nodes, except node 5, are quite similar, i.e. they are correlation. Data in node 5 looks different with the others, however, its negative is similar to the others, i.e. it is correlated with the others. Fig. 3 shows the estimated and the actual joint entropy according to number of nodes in the group. It is found that the estimated joint entropy and actual joint entropy is quite similar. The difference between them is because the actual entropy correlation coefficients are larger than those in estimated function.

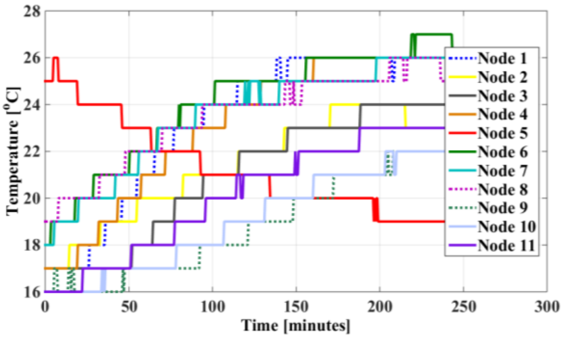

Fig. 2. Temperature data measured at 11 nodes

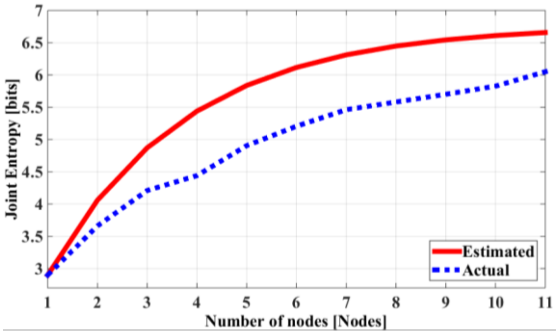

Fig. 3. Estimated and actual joint entropy with a group of 11 nodes.

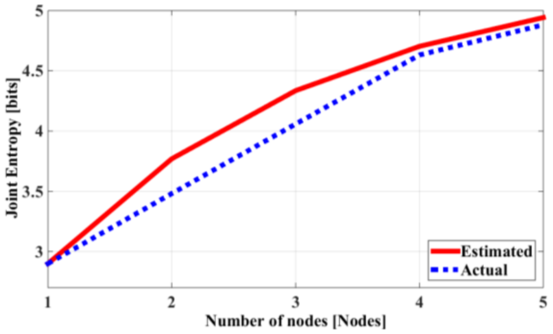

Fig. 4. Estimated and actual joint entropy with a group of 5 nodes.

Now, if entropy correlation coefficient is chosen to be $\rho_0=0.7$, a new group including nodes 1, 4, 5, 8 and 11 from 11 above nodes which correlation level is 0.7 is formed by using proposed clustering algorithm. The remaining nodes are grouped into another group with correlation level is 0.6. Fig. 4 shows the estimated and the actual joint entropy according to number of nodes in the group of node 1, 4, 5, 8, 11. It could be seen that in this case, the estimation is better because the entropy correlation coefficients are quite similar (in the range from 0.70 to 0.79). These examples show that our definition and algorithm are reasonable.

V. CONCLUSIONS AND FUTURE WORKS

The paper proposed a definition of correlation region in WSNs. Using this definition, a correlation clustering algorithm is proposed with less computation than previous algorithm. In addition, joint entropy of correlation region could be estimated to evaluate the effectiveness of the correlation based clustering in WSNs. In the future, by utilizing the advantages of correlation characteristics, it is expected that in each correlation region, the sensed data could be compressed in a higher rate to reduce the aggregated messages to save more energy for the transmission phase. On the other hand, the relation between the correlation level ρ and compression rate should be evaluated.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey", Computer Networks (Elsevier) Journal, vol. 38, no. 4, pp.393-422, March 2002.

[2] A. Abbasi and M. Younis, "A Survey on Clustering Algorithms for Wireless Sensor Networks," Computer Communications, vol. 30, no. 14-15, pp. 2826–2841, 2007.

[3] N. Vlajic and D. Xia, "Wireless Sensor Networks: To Cluster or Not To Cluster?" in Proc. International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), June 2006.

[4] Akyildiz, Ian F., Mehmet C. Vuran, and Ozgür B. Akan. "On exploiting spatial and temporal correlation in wireless sensor networks." Proceedings of WiOpt. Vol. 4. 2004.

[5] Shakya, Rajeev K., Yatindra N. Singh, and Nishchal K. Verma. "Generic correlation model for wireless sensor network applications." IET Wireless Sensor Systems 3.4 (2013): 266-276.

[6] Rui Dai, Ian F. Akyildiz, A Spatial Correlation Model for Visual Information in Wireless Multimedia Sensor Networks, IEEE transaction on multimedia, vol.11, No.6, 10. 2009

[7] Becker, Hila. "A survey of correlation clustering." Advanced Topics in Computational Learning Theory (2005): 1-10.

[8] Liu, Chong, Kui Wu, and Jian Pei. "A dynamic clustering and scheduling approach to energy saving in data collection from wireless sensor networks." SECON. Vol. 5. 2005.

[9] D. Maeda, H. Uehara, and M. Yokoyama, Efficient Clustering Scheme Considering Non-uniform Correlation Distribution for Ubiquitous Sensor Networks, IEICE Trans. on Fund. of Electronics, Comm. and Computer Sciences 2007,pp. 1344-1352.

[10] N. T. T. Nga, H. Uehara, T. Ohira, "Attribute change adaptation routing protocol for energy efficiency of wireless sensor networks," ICITA 2009.

[11] Taka, H., Uehara, H. and Ohira, T., Intermittent Transmission Method based on Aggregation Model for Clustering Scheme, ICUFN, 2011, pp.107-111

[12] Thomas M. Cover, Joy A. Thomas, "Elements of Information Theory," Copyright@1991 John Wiley & Sons.

[13] Cahill, Nathan D. "Normalized measures of mutual information with general definitions of entropy for multimodal image registration." Biomed. Image Reg. Springer 2010. 258-268.

[14] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol.31, No.3, 9.1999

[15] Intel Bekerley Research Lab
http://db.csail.mit.edu/labdata/labdata.html