IEIE Transactions on Smart Processing and Computing, vol. 5, no. 2, April 2016
http://dx.doi.org/10.5573/IEIESPC.2016.5.2.085

Entropy-based Correlation Clustering for Wireless Sensor Networks in Multi-Correlated Regional Environments

Nguyen Thi Thanh Nga, Nguyen Kim Khanh, and Son Ngo Hong

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
{ngantt, khanhnk, sonnh}@soict.hust.edu.vn

* Corresponding Author: Nguyen Thi Thanh Nga

Received February 19, 2016; Revised March 10, 2016; Accepted March 11, 2016; Published April 30, 2016

** Extended from a Conference: Preliminary results of this paper were presented at the ICEIC 2016. This present paper has been accepted by the editorial board through the regular reviewing process that confirms the original contribution.*

***Abstract*:** The existence of correlation characteristics brings significant potential advantages to the development of efficient routing protocols in wireless sensor networks. This research proposes a new simple method of clustering sensor nodes into correlation groups in multiple-correlation areas. At first, the evaluation of joint entropy for multiple-sensed data is considered. Based on the evaluation, the definition of correlation region, based on entropy theory, is proposed. Following that, a correlation clustering scheme with less computation is developed. The results are validated with a real data set.

***Keywords*:** Entropy, Correlation clustering, Entropy correlation coefficient, Multi-correlation regions

1. Introduction

Wireless Sensor Networks (WSNs) are simple low-cost approaches that can be used in a distributed environment. WSNs, consisting of a large number of small, inexpensive, battery-powered communications devices densely deployed throughout a physical space [1-3], can fulfill monitoring requests for surrounding environment characteristics, such as temperature, humidity, light, etc. In WSNs, energy conservation is commonly recognized as the key challenge to designing and operating the network [4, 5] because individual sensor nodes are expected to be low-cost, small in size, and powered by a non-replaceable battery.

In recent years, a considerable number of published studies on WSNs have dealt with the issue of energy conservation. Among these works, clustering is a well-established technique for reducing data collection costs in WSNs [6]. With this technique, sensor nodes are grouped into disjoint sets called clusters, and a cluster head (CH) is selected from among the sensor nodes to manage the cluster. Cluster members send their sensed data to their CH. CHs suppress local redundancies and send compressed data to the base station to avoid repeating redundant data. Organizing sensor nodes into such cluster-based topologies is widely accepted for energy conservation. Additionally, for local data compression, this approach coordinates the activities of cluster members, and addresses scalability issues (e.g., routing and communications costs) in large WSNs. Thus, this class of WSN is potentially viewed as the most energy-efficient and long-lived class of sensor network [7].

Numerous clustering algorithms for WSNs have been proposed in the literature. These algorithms vary in their objectives, which may include load balancing, fault-tolerance, increased connectivity, reduced delays, and maximal network longevity, which were discussed in detail by Abbasi and Younis [6]. However, these routing protocols do not yet consider the characteristics of environmental attributes.

Typical WSN 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 have correlations with each other. The existence of correlation characteristics brings significant potential advantages to the development of efficient communications protocols well-suited to the WSN paradigm. For example, due to the degree of correlation, data in a correlated group could be compressed at a high ratio; thus, the amount of data sent is reduced to save energy. Even with high-enough correlation, it may not be necessary for every sensor node in a correlation group to transmit its data to the sink; instead, a smaller number of sensor measurements might be adequate to communicate

the event features to the sink within a certain reliability/ fidelity level.

There have been some research efforts into studying this correlation in WSNs. For example, a theoretical framework to model the spatial and temporal correlations in sensor networks was developed [8, 9]. Following that, some approaches to exploit spatial and temporal correlations for efficient medium access and reliable event transport in WSNs were proposed. In those papers, the correlation models were based on the assumption that all nodes in a sensor field observe the same physical phenomenon, but with noise. The observation noise in each sensor node was modeled as a Gaussian random variable of zero mean and known variance. The samples from the event signal at each point of the event area were modeled as joint Gaussian random variables. The correlation coefficient chosen was a function of distance between nodes. However, this correlation coefficient measured only the linear correlation and was for scalar data. Dai and Akyildiz studied the correlation characteristics for visual information [10], and an entropy correlation coefficient was used. This correlation coefficient is more general than the previous correlation coefficient and directly related to the amount of information transferred in the network. However, the entropy correlation coefficient was also assumed to be a function of sensing position and direction. Additionally, both studies 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 was considered in the field of machine learning/data mining [11], and the results were also used in WSNs [12]. However, the meaning of correlation here is simply the similarity between two sets of data. The entropy concept is applied only to verify the effectiveness of the proposed correlation clustering.

An entropy/joint-entropy concept is more general to describe the correlation among sets of data, especially in some cases where the correlation is not distance/location dependent [13, 14]. Therefore, in some research, entropy theory was used for correlation clustering in WSNs [13, 15]. In that research, correlation region was defined based on the increased amount of joint entropy when the number of calculated nodes increases. This definition presents some problems, such as requiring vast amounts of computation time, because of the enormous number of combinations when picking sensor nodes in order to calculate joint entropy. In addition, the correlation level was not yet clarified.

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

The remainder of this paper is organized as follows. Section 2 presents the entropy concept and defines entropy correlation coefficient. In Section 3, joint entropy of a group of nodes is evaluated based on entropy of each node and the entropy correlation coefficient between two nodes. Validating this estimation is also discussed in this section. Section 4 proposes a definition of correlation region, and then, the correlation clustering scheme is described. Finally, the conclusion and future works are presented in Section 5.

2. Entropy Theory and Entropy Correlation Coefficient

2.1 Entropy Concept

In order to measure the correlation among sets of data, we first consider the concept of entropy and mutual information [16].

The entropy of a random variable is a function that attempts to characterize the "unpredictability" of a random variable. If random variable X takes on values in a set $X = \{x_1, x_2, ..., x_n\}$, and is defined by probability distribution $P(X)$, then entropy $H(X)$ of random variable X is written as:

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

Joint entropy is the entropy of a joint probability distribution, or a multi-valued random variable. If X and Y are jointly distributed according to $P(x, y)$, then joint entropy $H(X, Y)$ is:

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

The relation between entropy and joint entropy is:

$$H(X,Y) \le H(X) + H(Y) \tag{3}$$

with equality, if X and Y are independent.

Mutual information is a quantity that measures the relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, by one random variable about another. The formal definition of the mutual information of two random variables X and Y, where joint distribution is defined by $P(X, Y)$, is given by:

$$I(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{4}$$

The relation between mutual information and entropy is given by:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{5}$$

2.2 Entropy Correlation Coefficient

It was found that mutual information could be used to measure the correlation between two sets of data; the larger the value of $I(X, Y)$, the more correlation between X and Y. 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 the entropy of each individual data in the pair. To overcome this problem, we use normalized measures of mutual information, called entropy correlation coefficient [17], given as follows:

$$\rho(X,Y) = 2\frac{I(X,Y)}{H(X)+H(Y)} = 2-2\frac{H(X,Y)}{H(X)+H(Y)} \quad (6)$$

An entropy correlation coefficient presents the correlation level of a pair of data, which is independent of the value of individual entropy, and therefore, it can be used to compare the correlation level of two pairs of data.

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. If $\rho = 1$, (for $H(X) = H(Y) = H(X, Y)$), two sets of data totally depend on each other. If $\rho = 0$ (for $H(X, Y) = H(X)+H(Y)$), they are independent.

3. Joint Entropy Estimation

If nodes are correlated with each other, they can be collected into a correlated group. On the other hand, because they share so much information with each other, their joint entropy is much smaller than the total of all nodes' entropy. In order to determine whether a group of nodes is correlated or not, it is necessary to know the entropy of each node and the joint entropy of all nodes in the group. However, the calculation of joint entropy for a group of more than two nodes is a waste of time and computation resources. As a result, it is necessary to find a simple method to estimate joint entropy. In this section, we try to estimate the joint entropy of a group of nodes from the entropies of single nodes in the group and the entropy correlation coefficients of all pairs in the group.

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

$$H_{min} \le H(X_i) \le H_{max} \quad (7)$$

$$\rho_{min} \le \rho_{ij} \le \rho_{max} \quad (8)$$

The joint entropy is estimated based on the idea of hierarchical clustering [18] as described in the following sections.

3.1 Determining the Upper Bound of Joint Entropy

With a group that has only one node, we have the entropy of one node limited by equation (7):

$$H_1 = H(X_i) \le k_1 H_{max} \quad (9)$$

where $k_1 = 1$.

With a pair of nodes, X_i and X_j, from the definition of entropy correlation coefficient in equation (6), we have:

$$H_2 = H(X_i, X_j) = \frac{2-\rho(X_i, X_j)}{2}\left(H(X_i)+H(X_j)\right)$$

In addition,

$$H(X_i), H(X_j) \le H_{\max}, \text{ and } \rho(X_i, X_j) = \rho_{ij} \ge \rho_{min}$$

Then,

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

$$\text{or } H_2 \le k_2 H_{max} = bH_{max} \quad (10)$$

where $k_2 = b = 2-\rho_{min}$.

With a group of three nodes, X_i, X_j, and X_k, at first, the two nodes X_i and X_j are merged to create a new cluster represented by node X_{ij} with $H(X_{ij}) = H(X_i, X_j) \le k_2 H_{max}$. According to hierarchical clustering [18, 10], the correlation coefficient between one cluster and another can be obtained by the greatest/shortest/average correlation coefficient from any member of one cluster to any member of the other cluster. Therefore:

$$\rho(X_{ij}, X_k) = \min\{\rho(X_i, X_k), \rho(X_j, X_k)\} \ge \rho_{min}$$

Then,

$$\begin{aligned} H_3 &= H(X_i, X_j, X_k) = H(X_{ij}, X_k) \\ &= \frac{2-\rho(X_{ij}, X_k)}{2}\left(H(X_{ij})+H(X_k)\right) \\ &\le \frac{2-\rho_{min}}{2}\left(k_2 H_{max} + H_{max}\right) \\ &= \frac{b}{2}(k_2+1)H_{max} = k_3 H_{max} \end{aligned} \quad (11)$$

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

Similarly, joint entropy H_m of a group of 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 the joint entropy of $m-1$ nodes, and the entropy correlation coefficient between the sub-cluster and the main node is the greatest/shortest/average correlation

coefficient from any member of the sub-cluster to the remaining node. Thus:

$$H_m \le \frac{2-\rho_{min}}{2}\left(k_{m-1}H_{max}+H_{max}\right) = \frac{b}{2}\left(k_{m-1}+1\right)H_{max} = k_m H_{max} \quad (12)$$

where $k_m = \frac{b}{2}\left(k_{m-1}+1\right)$.

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

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

Or, in a more compact way (when $b \ne 2$):

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

3.2 Determining the Lower Bound of Joint Entropy

The lower bound of joint entropy of a group with m nodes could be determined in a way similar to determining the upper bound. The calculation is as follows.

With a group that has only one node, we have:

$$H_1 = H(X_i) \ge l_1 H_{min} \quad (15)$$

where $l_1 = 1$.

With a group of two nodes, we have:

$$H_2 \ge l_2 H_{min} = cH_{min} \quad (16)$$

where $l_2 = c = 2-\rho_{max}$.

With a group of m nodes ($m \ge 3$):

$$H_m \ge l_m H_{min} \quad (17)$$

where $l_m = \frac{c}{2}\left(l_{m-1}+1\right)$.

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

$$l_m = 2\left(\frac{C}{2}\right)^{m-1} + \left(\frac{C}{2}\right)^{m-2} + \ldots + \left(\frac{C}{2}\right)^{2} + \frac{C}{2} \quad (18)$$

Or, in a more compact way, (when $c \ne 2$):

$$l_m = \frac{\left(\frac{C}{2}\right)^{m}-1}{\frac{c}{2}-1} + \left(\frac{C}{2}\right)^{m-1} - 1 \quad (19)$$

3.3 Validating Entropy Estimation

In order to validate the proposed joint entropy estimation, at first, we consider two special cases. In the first case, all nodes completely depend on each other, i.e. all nodes measure the same information. In this case, $H(X_1) = H(X_2) = \ldots = H(X_m) = H$, and $\rho_{ij} = 1, \forall i, j = 1, 2, \ldots, m; i \ne j$.

Thus, $H_{min} = H_{max} = H$ and $\rho_{min} = \rho_{max} = 1$. Then, $k_m = l_m = 1$. Using (12) and (17), finally we have $H_m = H\left(X_1, X_2, ..., X_m\right) = H$.

These results show that, in this case, the estimated joint entropy is exactly equal to the actual joint entropy.

In the second case, all nodes are completely independent of each other. In this case, $\rho_{ij} = 0$, $\forall i, j = 1, 2, \ldots, m; i \ne j$, and thus, $\rho_{min} = \rho_{max} = 0$. Then, $k_m = l_m = m$. Using (12) and (17), finally we have

$$mH_{min} \le H_m = H\left(X_1, X_2, ..., X_m\right) \le mH_{max}$$

This inequality is true, because in this case:

$$H_m = H(X_1, X_2, ..., X_m) = H(X_1) + H(X_2) + \cdots + H(X_m).$$

Moreover, in order to verify the above estimation of joint entropy in a practical case, we will calculate joint entropy of some groups collected from sample data supplied by the Intel Berkeley Research Lab [19]. The sample data were collected from 54 sensors deployed in the Intel Berkeley Research Lab between February 28 and April 5, 2004. Mica2Dot sensors with weather boards collected time-stamped topology information, along with humidity, temperature, light and voltage values, once every 30 seconds. Data were collected using the TinyDB in-network query processing system, built on the TinyOS platform.

In this paper, only temperature data are considered. Let us choose a group of 11 nodes from among 54 nodes. The entropy of each node and the entropy correlation coefficient between each pair of nodes are shown in Tables 1 and 2, respectively. From these entropies and the entropy correlation coefficient, the lower bound and upper bound of some groups from among 11 nodes will be calculated.

The results are shown in Tables 1, 2, and 3, and Fig. 1.

Table 1. Entropy of each node in the group.

Nodes	1	2	3	4	5	6	7	8	9	10	11
Entropy	2.70	2.83	2.76	2.90	2.78	2.90	2.79	2.70	2.89	2.76	2.86

Table 2. Entropy correlation coefficients between pairs of nodes in the group.

Node	1	2	3	4	5	6	7	8	9	10	11
1	1	0.75	0.79	0.75	0.71	0.71	0.69	0.71	0.62	0.68	0.76
2	0.75	1	0.69	0.73	0.64	0.71	0.65	0.67	0.66	0.70	0.69
3	0.79	0.69	1	0.70	0.75	0.70	0.69	0.69	0.72	0.87	0.74
4	0.75	0.73	0.70	1	0.73	0.75	0.71	0.72	0.65	0.78	0.74
5	0.71	0.64	0.75	0.73	1	0.69	0.75	0.72	0.72	0.72	0.74
6	0.71	0.71	0.70	0.75	0.69	1	0.71	0.74	0.64	0.71	0.74
7	0.69	0.65	0.69	0.71	0.75	0.71	1	0.73	0.64	0.61	0.70
8	0.71	0.67	0.69	0.72	0.72	0.74	0.73	1	0.65	0.68	0.71
9	0.62	0.66	0.72	0.65	0.72	0.64	0.64	0.65	1	0.69	0.72
10	0.68	0.70	0.87	0.78	0.72	0.71	0.61	0.68	0.69	1	0.74
11	0.76	0.69	0.74	0.74	0.74	0.74	0.70	0.71	0.72	0.74	1

Table 3. Joint entropy, upper bound, lower bound, and estimated average joint entropy of some node groups.

Group	H_{min}	H_{max}	ρ_{min}	ρ_{max}	Joint entropy	Upper bound	Lower bound
1-3	2.70	2.83	0.69	0.79	3.96	4.28	3.61
1-4	2.70	2.83	0.69	0.79	4.44	4.66	3.82
1-5	2.70	2.90	0.64	0.79	4.72	5.46	3.94
1-6	2.70	2.90	0.64	0.79	5.10	5.69	4.02
1-7	2.70	2.90	0.64	0.79	5.37	5.84	4.07
1-8	2.70	2.90	0.64	0.79	5.62	5.94	4.10
1-9	2.70	2.90	0.62	0.79	5.84	6.27	4.11
1-10	2.70	2.90	0.61	0.87	5.94	6.33	3.50
1-11	2.70	2.90	0.61	0.87	6.05	6.38	3.50

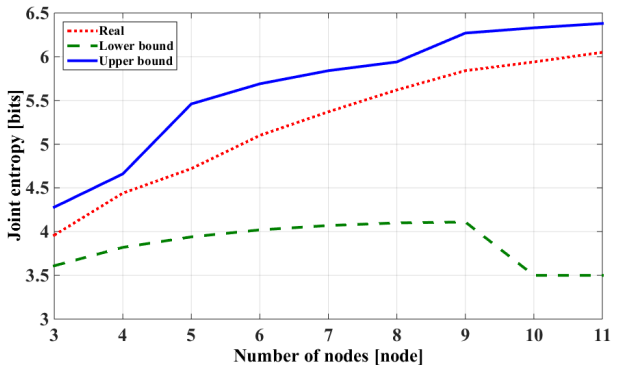

Fig. 1. Joint entropy, estimated lower bound and upper bound.

It is found that the joint entropy of one group is always between the lower bound and the upper bound. The above examples show the validity of the proposed estimation.

3.4. Joint Entropy and Correlation

Nodes in a correlation group will share a lot of information among themselves [13, 15]. Therefore, the joint entropy will not increase much when the number of nodes in the group increases. In other words, the joint entropy will go to "saturation" when the number of nodes increases. The faster the approach of the saturation state, the more correlation among the nodes.

Moreover, from inequalities (12) and (17), it is found that the upper bound and lower bound functions of joint entropy are the same. Only the arguments are different. For the upper bound function, the two arguments are the upper bound of entropy for a single node in the group, and the lower bound of the entropy correlation coefficient between two nodes in the group. For the lower bound function, two arguments are the lower bound of entropy for a single node in the group, and the upper bound of the entropy correlation coefficient between two nodes in the group. If the differences between the upper bound and the lower bound of entropy and of the entropy correlation coefficient are small enough, this function could be chosen to estimate joint entropy of the group.

Between these two arguments, the entropy correlation coefficient has a strong effect on the shape of the function. Fig. 2 shows the estimated joint entropy with different values for the entropy correlation coefficient.

From Fig. 2, it is found that with a high enough value of the entropy correlation coefficient, the estimated joint entropy has the same characteristics as the calculated joint entropy of the correlation group. This means joint entropy will attain "saturation" when the number of nodes increases. The faster the approach of the saturation state, the more correlation among the nodes. From these results, we can conclude the following.

- It is acceptable to use the lower/upper bound function to estimate joint entropy of a group, because both functions have similar characteristics when the

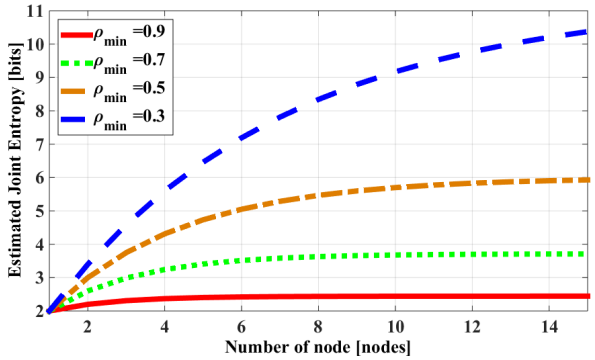

Fig. 2. Estimated joint entropy with different values of entropy correlation coefficients using the upper bound function (H_{max} = 2 [bits]).

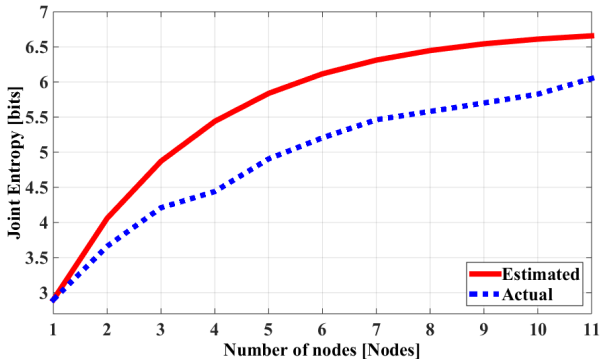

Fig. 3. Estimated and actual joint entropy with a group of 11 nodes.

number of nodes in the group increases.

- The entropy correlation coefficient of all pairs in the group can represent correlation of the group.

When correlated nodes are grouped, it is expected that the joint entropy is as small as possible. The worst case with joint entropy is its upper bound. Therefore, in order to evaluate the correlation of a group, the upper bound function should be used. Then, the actual joint entropy will always satisfy, if its upper bound is already satisfied.

Fig. 3 shows the estimated and actual joint entropy, according to the number of nodes in the group from the data in Tables 1 and 2; H_{max} = 2.90 and ρ_{min} = 0.61. It is found that the estimated joint entropy and the actual joint entropy are quite similar. The difference between them is because the actual entropy correlation coefficients are larger than those in the estimated function.

4. Correlation Region Definition and Correlation Clustering Algorithm

4.1 Correlation Region Definition

As mentioned by Akyildiz et al. [8], sensor nodes in the same correlation region record information on a single event in the sensor field, i.e. these sensed data are correlated with each other. Because the sensed data are taken from the same event, the number of bits to represent the sensed data should not be so very 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.

Moreover, as shown in the last section, with a group that has the two above properties (similar entropy and a similar entropy correlation coefficient), the group is a correlated group. 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 a similar entropy value, and the entropy correlation coefficient between all pairs of nodes is also similar.

In practice, it is difficult to obtain similarity between two entropies or entropy correlation coefficients. Then, the correlation region could be defined in a more practical way as follows.

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

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

where ΔH is small enough.

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

With this definition, we can estimate the joint entropy of m nodes $\{X_1, X_2, ..., X_m\}$ with the following equation:

$$H(X_1, X_2, ..., X_m) = k_m H_0 \quad (19)$$

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

4.2 Correlation Clustering Algorithm

Using the definition of correlation region, a sensor field could be divided into correlation regions with a specified base entropy and correlation level. The clustering process is described in Fig. 4:

In the first step of the algorithm (*), the base entropy and correlation level are chosen, such that they can cover all possible values for entropy and the entropy correlation coefficient in the network. The value of the entropy correlation coefficient should be chosen from high to low.

In the step marked ** in the algorithm, if more than one node satisfies the condition $0 < C(X_i) = \max\{C(X_j), X_j \in G\}$, the node that has the maximum entropy value will be removed.

```
BEGIN
REPEAT
        Choose H0, ρ0, ΔH; (*)
        Initialize new group G = ∅;
    FOR each node Xi in the network not
    belonging to any group
        IF H0 ≤ H(Xi) ≤ H0 + ΔH
                Add Xi into G
        ENDIF
    ENDFOR
    REPEAT
        FOR each node Xi in G
                Calculate C(Xi)= number of node
    Xj where H(Xi,Xj) < ρ0
        ENDFOR
        FOR each node in G
                IF 0 < C(Xj) = max{C(Xj), Xj∈G}
                        Remove Xi from G (**)
                ENDIF
        ENDFOR
    UNTIL max{C(Xj), Xj∈G}=0
UNTIL all nodes are grouped
END
```

Fig. 4. Correlation-based clustering algorithm.

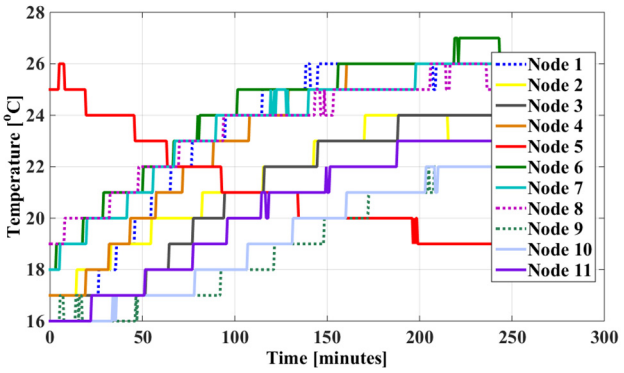

Fig. 5. Temperature data measured at 11 nodes.

4.3 Evaluations

In order to verify the definition and clustering algorithm, we reconsider the data in Tables 1 and 2. The entropy is in a range from 2.70 to 2.90, i.e. H_0 = 2.70; ΔH = 0.2. The entropy correlation coefficients are all larger than, or equal to, 0.60, i.e. ρ_0 = 0.60. Fig. 5 shows the data of 11 nodes in the group.

It is found that all nodes, except node 5, are quite similar, i.e. they are in correlation. Data in node 5 look different from the others, but its negative is similar to the others. Thus, it is correlated with the others. This example shows that our definition is reasonable.

Now, if the entropy correlation coefficient is chosen as ρ_0=0.7, then using the proposed clustering algorithm, we will have a new group from the above 11 nodes that have a correlation level of 0.7, including nodes 1, 4, 5, 8 and 11. The remaining nodes are placed into another group with a correlation level of 0.6. Fig. 6 shows the estimated and actual joint entropy according to the number of nodes in the group of nodes 1, 4, 5, 8, and 11. We can see that, in this case, the estimation is better because the entropy correlation coefficients are quite similar (in the range 0.70 to 0.79).

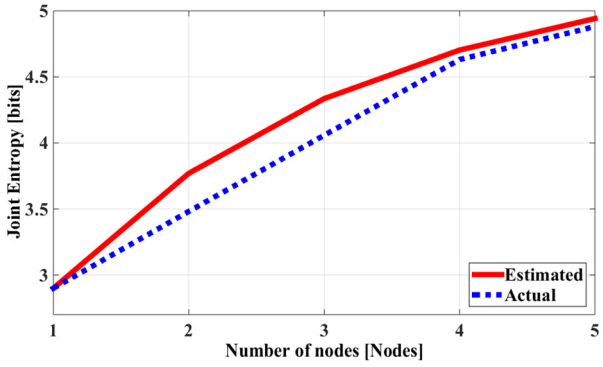

Fig. 6. Estimated and actual joint entropy with a group of 5 nodes.

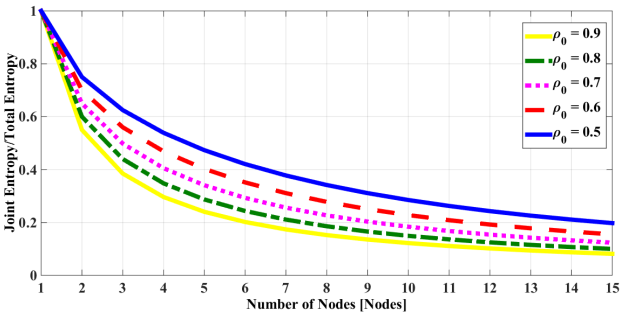

Fig. 7. Ratio between joint entropy over total entropy.

In addition, it is found that when the correlation level is higher, the number of nodes in a cluster/group is reduced. If the number of nodes in a cluster/group is small, the effectiveness of correlation-based clustering may not be clear. In order to evaluate the effectiveness of clustering, let us consider the ratio between joint entropy and total entropy of a group.

According to the first condition in the definition of correlation region, total entropy of a correlation group of m nodes can be approximated as mH_0. Using estimated joint entropy (18), the ratio between joint entropy and total entropy can be calculated as:

$$\frac{Joint\ entropy}{Total\ Entropy} = \frac{k_m H_0}{mH_0} = \frac{k_m}{m}$$

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

$$= \frac{}{m} \qquad (19)$$

Fig. 7 shows the relation between this ratio and the number of nodes with a difference value of ρ_0.

It is found that the larger the number of nodes, the smaller the ratio between joint entropy and total entropy (with the same correlation level). The higher the correlation level, the smaller the ratio between joint

entropy and total entropy, i.e. the more effective the correlation-based clustering. Therefore, it is necessary to consider both correlation level and the number of nodes in a correlation region in order to obtain the effectiveness of correlation clustering.

5. Conclusions and Further Studies

The paper has shown that joint entropy of a group of data can be estimated by using information on the entropy of a single data and the entropy correlation coefficient of a couple of data. Therefore, the correlation region is defined based on entropy and the entropy correlation coefficient, which makes the correlation clustering algorithm become simple with less computation than in previous algorithms.

In the future, by utilizing the advantages of correlation characteristics, it is expected that in each correlation region, the sensed data can be compressed at a higher rate in order to reduce the aggregated messages to save more energy in transmission. Moreover, the relation between correlation level ρ and the compression rate should be evaluated. On the other hand, the effects of the entropy correlation coefficient on joint entropy of a data group mentioned in this paper need further evaluation to achieve the optimal value for each application. The correlation characteristic could also be used to turn on/off sensor nodes to save energy when the correlation level is high enough with acceptable distortion.

Acknowledgement

This research was supported by the 911 project of the Ministry of Education and Training of Vietnam.

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. Article (CrossRef Link)

[2] D. Culler, D. E. M. Srivastava, “Overview of Sensor Network”, IEEE Computer Magazine, vol. 37, no. 8, pp. 41-49, August 2004. Article (CrossRef Link)

[3] C. Siva Ram Murthy and B. Manoj, “Ad Hoc Wireless Networks: Architectures and Protocols”, Prentice Hall, 2004.

[4] J. N. Al-Karaki and A. E. Kamal, “Routing techniques in wireless sensor networks: a survey”, IEEE Wireless Comm., vol. 11, pp. 6-28, 2004. Article (CrossRef Link)

[5] Ignacio Solis and Katia Obraczka, “Isolines: Energy efficient mapping in Sensor Networks”, Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC'05), Cartagena, Spain, June 2005. Article (CrossRef Link)

[6] 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. Article (CrossRef Link)

[7] 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. Article (CrossRef Link)

[8] 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.

[9] 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. Article (CrossRef Link)

[10] 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. Article (CrossRef Link)

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

[12] 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. Article (CrossRef Link)

[13] D. Maeda, H. Uehara, and M. Yokoyama, Efficient Clustering Scheme Considering Non-uniform Correlation Distribution for Ubiquitous Sensor Networks, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences 2007 E90-A(7):1344-1352. Article (CrossRef Link)

[14] N. T. T. Nga, H. Uehara, T. Ohira, “Attribute change adaptation routing protocol for energy efficiency of wireless sensor networks”, ICITA 2009. Article (CrossRef Link)

[15] Taka, H., Uehara, H. and Ohira, T., Intermittent Transmission Method based on Aggregation Model for Clustering Scheme, Third International Conference on Ubiquitous and Future Networks (ICUFN), 2011, pp.107-111, Print ISBN 978-1-4577-1176-3, 15-17 June 2011.Article (CrossRef Link)

[16] Thomas M. Cover, Joy A. Thomas, “Elements of Information Theory”, Copyright@1991 John Wiley & Sons, Inc. Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1 Chapter2 pp.13-49.

[17] Cahill, Nathan D. “Normalized measures of mutual information with general definitions of entropy for multimodal image registration”. Biomedical Image Registration. Springer Berlin Heidelberg, 2010. 258-268. Article (CrossRef Link)

[18] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, 9.1999. Article (CrossRef Link)

[19] Intel Berkeley Research Lab

[20] Article (CrossRef Link)

Nguyen Thi Thanh Nga is a doctoral candidate at the Hanoi University of Science and Technology (HUST), Hanoi, Vietnam. She received her BSc from Faculty of Electrical and Telecommunication, HUST in 2002, and her MSc in Electronics and Information Engineering from Toyohashi University of Technology (TUT), Japan, in 2009. Her research interests include wireless sensor networks. She is a student member of the IEEE.

Nguyen Kim Khanh received BSc from the Faculty of Electronics and Telecommunication, Hanoi University of Science and Technology in 1985, and a PhD in IT from the Belarusian State University in 1991. Currently, he is Head of the Department of Computer Engineering, School of Information and Communication Technology, Hanoi University of Science and Technology. His main research interests include advanced computer architecture, embedded systems, wireless sensor networks, and speech signal processing.

Son Hong Ngo received a BSc from the Faculty of Information Technology, Hanoi University of Science and Technology (HUST) and a PhD from the Japan Advanced Institute of Science and Technology (JAIST) in 2000 and 2006, respectively. From 2006 to 2007, he stayed in GSIS, Tohoku University, Japan, for postdoctoral research on routing in WDM optical networks. He is now an associate professor at the NCT Lab, SoICT. His current research interests include Future Internet, Wireless Sensor Networks, and Distributed Systems.