Received January 11, 2017, accepted January 26, 2017, date of publication January 27, 2017, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2660770

Joint Design of Energy-Efficient Clustering and Data Recovery for Wireless Sensor Networks

XUAN LIU[1], JUN LI[2,4], (Senior Member, IEEE), ZY DONG[1], (Fellow, IEEE), AND FEI XIONG[3]

[1]School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia
[2]School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[3]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[4]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Corresponding author: J. Li (jun.li@njust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501238, in part by the Jiangsu Provincial Science Foundation under Project BK20150786, in part by the Specially Appointed Professor Program in Jiangsu Province, 2015, in part by the Fundamental Research Funds for the Central Universities under Grant 30916011205, and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D04.

ABSTRACT The two-layer network structure has been widely adopted in wireless sensor networks (WSNs) for managing sensor nodes. In such a structure, the low layer nodes communicate with their cluster head, followed by the cluster-head nodes communicating with the base station operating in either a one-hop or a multi-hop manner. The main focus of node-clustering algorithms is minimizing energy consumption due to strictly limited resources in WSNs. Also, WSNs are data intensive networks with the capability of providing users with accurate data. Unfortunately, data missing is common in WSNs. In this paper, we propose a novel joint design of sensor nodes clustering and data recovery, where the WSNs is organized in a two-layer manner with our developed clustering algorithm, and then, the missing data are recovered based on this two-layer structure. Furthermore, in the proposed clustering algorithm, we take both the energy-efficiency and data forecasting accuracy into consideration and investigate the tradeoff between them. This is based on the key observation that the high energy-efficiency of the network can be achieved by reducing the distances among the nodes in a cluster, while the accuracy of the forecasting results can be improved by increasing the correlation of the data stream among the nodes in a cluster. Simulation results demonstrate that our joint design outperforms the existing algorithms in terms of energy consumption and forecasting accuracy.

INDEX TERMS Wireless sensor networks, energy efficiency, node clustering, data forecasting

I. INTRODUCTION

With the development of electronic and sensor technologies, wireless sensor network (WSN) becomes a popular network architecture for current and future wireless communications. Particularly in recent years, WSNs have been widely applied to various practical scenarios, such as intelligent transportation, health-care monitoring, industrial manufacture, robotics, and so on [1]. Furthermore, WSNs will be prevailing with the emergences of intelligent applications, e.g., Smart City [2], Wearable Computing Devices [3], Tactile Internet [4], etc. A major responsibility of the WSNs is accurately sensing and collecting useful data, for example, the measurements of air quality, humidity, biomedical and chemical information, and yielding sensed big data for further analysis [5]. At the same time, cloud-computing enabled technologies, e.g., Cloud-RAN [6] and Fog-RAN [7], provide the WSNs with the leverages of computation, communication and storage resources [8], as well as a promising method to manage, process and preserve the privacy of massive aggregated data [9].

A wireless sensor node consists of multiple modules, including battery, data process units, storage, transmitter/receiver pair, and one or several sensor devices. These sensor nodes collect the information about the surrounding environment and forward it to the base station through a one-hop or multi-hop manner. As such, WSNs serve as bridges between the physical world and human societies, resulting in a cyber-physical system [10]. However, due to limited resources, sensor nodes shall cooperate with each other to carry out complicated tasks [11]–[13]. For example, mobile crowd-sensing has proved to be an effective and efficient way to collect and process environmental data [14], as well as reconstruct the spatial field of a physical quantity (e.g., traffic condition) [15], [16].

Apart from data transmissions, WSNs need to efficiently eliminate the redundant data, query the necessary data, fuse the correlated data and recover the missing data [17]. As shown in Fig. 1, the base station sends a query message to a specific cluster head to request the readings of a node. After receiving the query message, the cluster head communicates

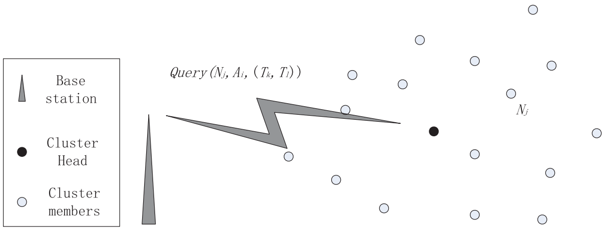


FIGURE 1. A typical scenario of query the useful readings.

with its members to obtain the readings, and then transmits them to the base station. Each node's readings may include several attributes, such as temperature, humidity, wind speed, and so on [18]. However, it is likely that the base station cannot successfully obtain the desired data due to the hostile environment of the communications. In this case, forecasting of missing data is needed. An intuitive method is collecting all correlated data to the base station and then forecasting the data in a centralized manner through forecasting algorithms.

At the same time, intensive data transmissions and processing will cause a large amount of energy consumption at sensor nodes [19]–[21]. As WSNs are usually battery-driven with limited power supply, battery lifetime is a vital factor for long-term operations of sensor nodes. Generally, there are two strategies to extend battery life. One is to charge the battery from other energy sources, such as energy harvesting and power transfer [22]. The other one is to develop protocols for efficiently managing energy consumptions, which has been widely studied as a hot topic in the academia society [23], [24]. Specifically in [23], an effective power allocation algorithm is proposed for interference alignment based cognitive radio networks, which can improve the energy efficiency of the secondary users significantly.

In particular, a hybrid two-layer structure has been proposed to deal with the energy efficiency issue, where the sensor nodes are divided into multiple clusters and each cluster is managed by its cluster head. This hybrid distributed method can achieve a good tradeoff between the fully centralized and distributed approaches. Furthermore, the hybrid approach consists of two modules, including a node clustering module and a missing data recovery module. In conventional wisdom, node-clustering algorithms only focus on decreasing energy consumption problem given that energy is very strictly limited in WSNs. Meanwhile, WSNs are data-centered networks aiming at providing users with accurate data. A well developed clustering algorithm should take both energy-efficiency and recovery-accuracy into account simultaneously.

In this paper, we propose a distributed data recovery scheme to address the above mentioned issue. The WSNs are assumed to be managed in a two-layer structure, i.e., the network is divided into multiple clusters and the cluster heads play a role as a bridge between the sensor nodes and the base station. To be specific, we first define both the spatial Euclidean distances between nodes [25] and the distance between the reading series generated by the nodes. Then we propose a clustering algorithm, where only the nodes having similar readings and small pairwise distances can be dispatched to the same cluster. Within each cluster, the forecasting process is conducted by each cluster head. We note that over-fitting problem imposes negative effects on the data forecasting accuracy. To address this problem, we further develop a lightweight missing data forecasting algorithm for the WSNs under the case of strictly limited resources.

In the simulations, we compare the proposed scheme with the existing approaches in terms of energy consumption and forecasting accuracy. The proposed scheme is the first of its kind that integrates node clustering and missing data forecasting into a unified framework. Simulation results show that the proposed scheme performs much better in terms of forecasting accuracy compared with the existing approaches, e.g., MUlti-SequenCe LEast Squares (MUSCLES) forecasting algorithm [26].

The major contributions are summarized as follows.

1) We identify the significance of achieving high energy efficiency and forecasting accuracy to design massive connected sensor devices networks.
2) A unified framework with the novel two-layer approach is proposed to improve the forecast accuracy and energy efficiency simultaneously. This is achieved by the sensor nodes clustering phase, followed by the missing data forecasting phase.
3) We present a clustering algorithm based on the similarities in the Euclidean distances and aggregated data. In particular, a specific approach is proposed to address the outliers in the process of clustering.
4) A lightweight missing data forecasting algorithm is developed to address the over-fitting issue, thereby significantly improving the forecasting accuracy.
5) Simulation results will demonstrate the effectiveness of the proposed algorithms to design energy efficient and missing data forecasting accurate WSNs, compared with the state-of-the-art algorithms.

The remainder of the paper is organized as follows. In Section II, we summarize the related work in two aspects, i.e., node clustering algorithms and data forecasting techniques. We then propose a novel node clustering algorithm and a data forecasting approach in Section III and IV, respectively. Extensive simulations are provided in Section V to evaluate the performance of the proposed method compared with several existing approaches. Finally, we conclude this paper and illustrate our future work in Section VI.

II. RELATED WORK

In this section, we review the related works in two aspects, i.e., clustering protocols and data forecasting algorithms.

A. NODES CLUSTERING APPROACHES

Nodes clustering problem in WSNs has been intensively investigated in the literatures and many classic approaches have been proposed. In particular, linked cluster algorithm (LCA) [20], [21] is one of the earliest clustering algorithms, which requires no central controller and is fully distributed. In LCA, the cluster heads form a backbone network and they connect with all the sensor nodes in its cluster directly. This structure is very flexible to implement a wide variety of routing strategies and can be used to avoid the problem of hidden terminals. The hierarchical control clustering algorithm proposed in [27] treats the network as a graph.

A cluster is defined as a subset of vertices whose included graph is connected. Finally, a multi-tier hierarchical cluster structure is formed and it satisfies several constrains simultaneously. In the process of clustering WSNs, the authors in [28] argue that it was very unwise to ignore the geographical information of the sensor nodes, especially for a large WSN. They then propose a novel clustering algorithm, which used geographical radius of cluster instead of logical radius.

Another classic clustering algorithm, namely, LEACH, is proposed in [29]. LEACH forms clusters based on the received signal strength and then the cluster heads serve as a bridge between the cluster members and the base station. Various applications of LEACH illustrate that it can always produce relatively good results in terms of energy efficiency and data transmission quality.

B. DATA FORECASTING ALGORITHMS

Data forecasting techniques have been widely used in WSNs to reduce data transmission and improve the energy-efficiency [30], [31]. The authors in [30] propose an on-mote filtering approach relying on a local multi-step assessment of sensor data with forecasting and assessing value of information. Simulation results showed that the proposed approach reduces the number of data transmissions and the energy consumption significantly. In [31], the authors discuss the implementation of an Artificial Neural Network (ANN) algorithm in a low cost system-on-chip and develop an autonomous intelligent wireless sensor network.

Additionally, there are several common missing data forecasting algorithms proposed in the field of time series mining. Specifically, the simplest approach to forecasting the missing data is Yesterday, in which, we replace the missing data with the nearest previous data. The major disadvantage of this method is that the forecasting error accumulates with the increasing of continuous missing data in which situation Yesterday becomes malfunctioning. Auto-regression based approaches are also very popular and they have been used in various scenarios. They forecast the missing data of a time series by first mining the pattern underneath the time series and then using the pattern to forecast the missing data. Similar to Yesterday, the forecasting accuracy decreases significantly with the increasing of the missing data.

Different from Yesterday and auto-regression based approaches, MUSCLES [26] makes full use of the high correlations between the co-evolving time series and construct a relation between them through linear mathematic tools. Simulation results illustrate that MUSCLES outperforms Yesterday and auto-regression based approaches in terms of forecasting accuracy. However, in MUSCLES, it is hard to define the correlations between the time series and the over-fitting problem is ignored.

Although the above discussed approaches in both the two fields are well developed, it is very difficult to integrate the two fields into a unified framework. In the process of designing nodes clustering algorithms, the goal is to reduce energy consumption and ignore the forecasting problem which is unwise for data-centered networks. On the other hand, most of the existing forecasting algorithms are not designed for the WSNs to recover the missing data and are not designed to be energy efficient. Besides, there are also some disadvantages for the forecasting approaches such as the over-fitting problem. As far as we know, there is no existing approaches that can solve the forecasting problem in WSNs.

III. NODE CLUSTERING ALGORITHM BASED ON LOCATION AND DATA CORRELATIONS

In this section, we will present a novel node clustering algorithm by considering both the locations of all the nodes and data correlations between each pair of data streams generated by the nodes. We first assume that all the nodes in the network are located in a plain area and each node has a standard radio radius r that can be adaptively changed by the nodes in some exceptional cases. Each node in the network has the capability of severing as a cluster head and the nodes take turns to be a cluster head considering that the cluster head consumes much more energy compared with the other nodes. The clusters of all the nodes need to be reconstructed when the lasting time of a round exceeds a threshold or some cluster heads cannot serve as a cluster head anymore because of limited resources.

The operation of the cluster algorithm is divided into multiple rounds. Similar to most of the existing approaches, each round of the proposed node-clustering algorithm consists of two phases: a phase of cluster head selection and clusters formation phase. In the cluster heads selection phase, the locations and residual energy of the nodes perform more important role to reduce the energy consumption of data transmission. However, data correlations take a more important role in the process of cluster formation to increase the correlations between the nodes in a cluster. The correlations among the time series have significant impact on the forecasting accuracy.

Overall, the proposed approach is operated in a half-distributed manner, i.e., the cluster heads selection phase is centralized and the cluster forming phase is distributed, and this is a tradeoff between the energy consumption and missing data forecasting accuracy. The base station chose the cluster heads in a centralized way to make the distribution of cluster heads reasonable. On the contrary, the clusters formation is totally distributed and each node has its own choice to improve the forecasting accuracy. In the following, we introduce cluster heads selection and clusters' formation in Section III-A and III-B, respectively.

A. CLUSTER HEADS SELECTION

Considering that most distributed approaches offer no guarantee about the number and distribution of the cluster heads,

Algorithm 1 Pseudocode of Cluster Heads Election

```
All the nodes in the network transmit their node IDs, locations and residual energy to the base station
Sort the residual energy of the nodes and select the top-half nodes with more energy as the candidates of cluster heads
Adopt the method in [32] to choose the $m$ final cluster heads
Broadcast the IDs and locations of the selected cluster heads in the network
```

in this paper, we designed a centralized clustering algorithm to choose the cluster heads. The entire process is shown in Algorithm 1. In the initial of each round, all the nodes first transmit their IDs, location information and the residual energy to the base station. After receiving all the information about the nodes, the base station first selects the top half nodes that have more residual energy as the candidates of the cluster heads to balance the energy load among all the nodes and prolong the life time of the network.

We assume that the communication energy scales exactly with squared distance and our goal is minimizing the amount of energy consumption for all the non-cluster nodes to communicate with their nearest cluster head. Note that the non-cluster head nodes may not select the nearest cluster head node as its cluster head and the clusters forming process will be discussed in Section III-B.

Finding the optimal result of the cluster heads is an NP-hard problem and it is impractical to use the brute force algorithms considering that the numbers of the nodes in WSNs are very large. Authors in [32] propose a heuristic method to solve the problem based on genetic algorithms and this algorithm. In this paper, the method in [32] is adopted by the base station to choose the m cluster heads.

Given the final selected cluster heads, the base station broadcasts the IDs of the nodes in the whole network. Then each node compares its own ID with the received IDs to realize whether it is a cluster head or not. Considering that transmitting all the readings of the nodes to the base station is impractical, we ignore the data correlations between the readings of the nodes when selecting the cluster heads and only the information of location and residual energy of the nodes are considered by the station.

Intuitively, a straightforward solution of node clustering is to transmit all the necessary information to the base station and both cluster heads selection and clusters formation are operating in a centralized way. Then the base station broadcasts the clustering result in the network, which is very similar to LEACH [29]. However, this pattern is impractical for our approach, since the data correlation is also taken into consideration and the readings of the nodes are of large amount, which cannot be transmitted to the base station totally. Therefore, we design a distributed approach to form the clusters, as presented in the following.

B. CLUSTERS FORMING

We first introduce the definition of trend closeness among the readings of nodes, which is a common measurement of the correlations between the time series.

Definition 1 (Spatial Closeness between Sensor Nodes): Under the assumption that all the nodes are located in a plain regime, the spatial distance of two nodes *dist* is defined as the Euclidean Distance. Sensor node n_p is spatial close to n_q if n_p is at worst d far from n_q. Parameter d is set by the users of the network and naturally not too much larger than the radio radius r, otherwise, the nodes in a cluster cannot communicate with each other very well. The spatial closeness between sensor nodes has important affection on the selection of cluster heads. ■

Definition 2 (Trend Closeness between Readings of Nodes: Each node in the network generates readings about the surrounding environment which can be treated as a time series and in most cases the readings are strongly correlated between neighboring nodes. For each node, we treat its readings as a time series and the time series is infinite. In this paper, we define the trend closeness between two time series as the Dynamic Time Wrapping (DTW) which is more robust than the Euclidean Distance. In this paper, for convenience, we compute the trend closeness based on the latest ten readings of the nodes rather than all the historical readings of the nodes. ■

Based on the above two definitions, we propose a novel clusters formation algorithm for WSNs considering both the spatial locations and data correlations. The pseudocode of clusters formation is presented in Algorithm 2 and will be discussed as follows.

Algorithm 2 Pseudocode of Clusters Forming

```
The base station broadcasts the IDs and locations of the cluster heads in the network
Each non-cluster head node $n_j$ computes the distances between itself and the cluster heads
Add the cluster heads that spatial close to $n_j$ to the candidate set $S_{n_j} = \{ch_1, ch_2, \cdots, ch_o\}$
$n_j$ communicates with the cluster heads in $S_{n_j}$ and get the most recent s readings from each node in $S_{n_j}$
$n_j$compute the trend closeness between its own readings and that of each node in $S_{n_j}$
Select the node $ch_{o'}$ with the least distance as the cluster head and send a message to $ch_{o'}$ to join its cluster
The cluster heads maintain all the information of its cluster members
```

After receiving the IDs and locations of the cluster heads, each non-cluster heads node computes the distances between itself and the cluster heads and then add the cluster heads in its communication range to the candidate set S_{n_j}. The node needs to choose a cluster head in S_{n_j} as its cluster head and join the corresponding cluster based on the similarities between the readings of itself and that of the cluster heads. In this paper, we employ DTW as the distances between the reading streams and each node select the cluster head with the most similar reading streams as its cluster head. For a cluster head, all the members share a similar reading stream with itself. As a result, all the nodes in the same cluster have similar reading streams.

In the process of forming the clusters, the similarities between the readings of the nodes and that of the cluster heads play more important roles compared with the locations of the nodes. This can improve the forecasting accuracy significantly and it will be discussed in the next section.

IV. MISSING DATA FORECASTING

In this section, we will develop the missing data forecasting algorithms for the given clusters, calculated by the algorithms presented in the previous section. Intuitively, the base station can collect all the time series from the network and then recover the missing values based on conventional data forecasting approaches, such as MUSCLES [26]. However, this method is impractical considering that energy is strictly limited and the jamming problem in the network. Therefore, we will design a distributed framework to forecast the missing values based on the node clustering approach proposed previously.

In Section IV-A, we first illustrate the over-fitting problem in time series forecasting through a complicated simulation. Simulation results show that over-fitting problem is an obvious drawback for most of the existing forecasting algorithms. Then, we propose a novel distributed forecasting algorithm for WSNs in Section IV-B. The forecasting approach executed by the cluster heads as well as the organizing of the clusters will be discussed in detail in the following.

A. THE OVER-FITTING PROBLEM IN TIME SERIES FORECASTING

In this section, we conduct a detailed experiment to present the over-fitting problem in time series forecasting field. This is the first clear illustration that too many uncorrelated time series will decrease the forecasting accuracy significantly. Therefore, we need to select several similar time series to forecast the missing data rather than employ all the time series. In the following, we present the experiment including collecting datasets, forecasting the missing data and analyzing of the simulation results.

We first collect 13 important stock price indexes from March 2^{nd}, 2015 to February 29^{th}, 2016 all over the world and these indexes are LNSZZS (CN), SZSE (CN), GEI (CN), HIS (HK), Nikkei (JPN), Kospi (KR), FTSE (UK), CAC40 (FR), DAX (GER), MIB (IT), SSMI (CH), DJIA (US), Nasdaq (US). The normalized indexes are presented in Fig. 2. For convenience, we employ letters $a \sim m$ to represent the time series. Intuitively, we can find that each time series may be very similar with some time series and also can be very different with some other time series. Consider a as an example, a is very similar with b and c. This is reasonable, because a, b and c are all belong to China and they influence each other significantly. However, a is very different with the other time series because China stock market is not very mature compared with that of developed countries.

Each time series is composed of 700 data points and we need to measure the distances between each pair of time series x_i and x_j quantitively by Euclidean distance, i.e.,

$$Dist(x_i, x_j) = \left(\sum_{k=1}^{N} (x_i(k) - x_j(k))^2 \right)^{1/2}. \qquad (1)$$

The distances between each pair of stock price index series are presented in Table 1.

Furthermore, the forecasting error is defined as the distance between the actual missing values and the forecasted results. For each time series, we sort all the other time series based on the distances in an ascending style. For a, the sorting result is $b, c, d, f, l, i, m, k, g, e, h, j$. We assume that the 301^{th} to 400^{th} data point of a time series is missing and we use the first n most similar time series to forecast the missing data of a time series.

As an example, the actual missing values and the forecasted results for time series i when $n = 6$ is presented in Fig. 3.

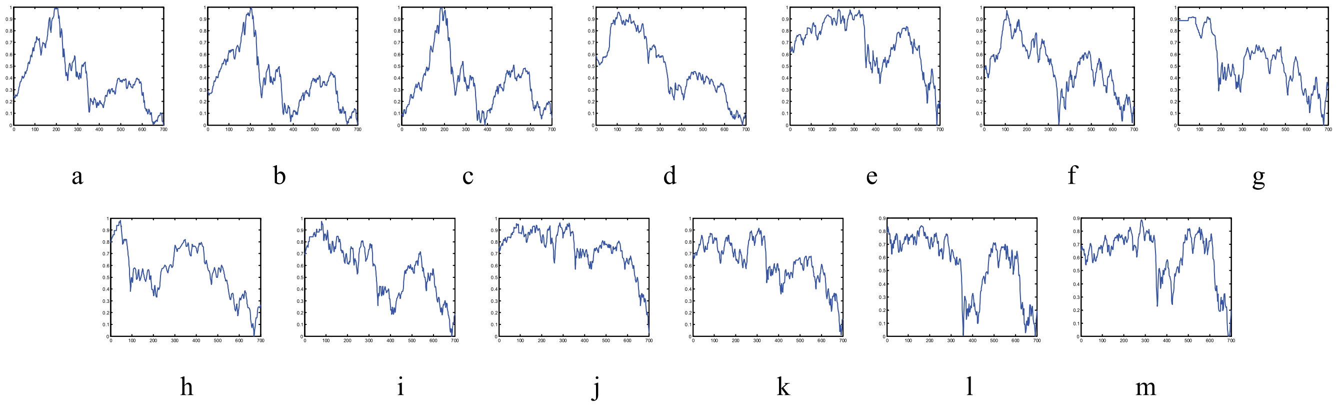


FIGURE 2. Fluctuations of stock price indexes.

TABLE 1. Distances Between Time Series

Distances	a	b	c	d	e	f	g	h	i	j	k	l	m
a	0	1.33	3.11	3.81	8.59	4.51	7.94	8.97	6.36	10.01	7.41	5.68	7.31
b	1.33	0	2.43	4.64	9.23	5.18	8.33	9.40	6.82	10.65	8.03	6.02	7.78
c	3.11	2.43	0	6.31	10.0	6.19	9.58	10.27	8.14	11.50	9.01	7.08	8.37
d	3.81	4.64	6.31	0	6.99	2.91	5.94	7.62	4.11	8.02	5.48	4.51	6.35
e	8.59	9.23	10.0	6.99	0	6.73	7.66	7.44	5.11	3.00	3.33	5.10	3.23
f	4.51	5.18	6.19	2.91	6.73	0	5.96	7.30	4.22	7.70	5.10	4.07	5.54
g	7.94	8.33	9.58	5.94	7.66	5.96	0	4.58	5.25	7.25	5.52	6.45	7.03
h	8.97	9.40	10.27	7.62	7.44	7.30	4.58	0	6.37	7.13	5.42	7.33	6.78
i	6.36	6.82	8.14	4.11	5.11	4.22	5.25	6.37	0	6.06	3.24	3.09	4.47
j	10.01	10.65	11.50	8.02	3.00	7.70	7.25	7.13	6.06	0	3.97	6.42	4.47
k	7.41	8.03	9.01	5.48	3.33	5.10	5.52	5.42	3.24	3.97	0	4.17	3.25
l	5.68	6.02	7.08	4.51	5.10	4.07	6.45	7.33	3.09	6.42	4.17	0	3.36
m	7.31	7.78	8.37	6.35	3.23	5.54	7.03	6.78	4.47	4.47	3.25	3.36	0

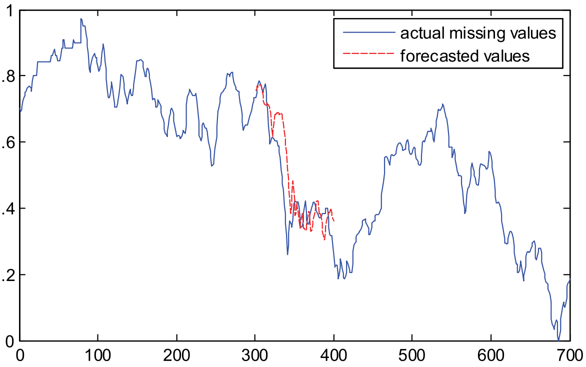


FIGURE 3. The actual missing values and the forecasted values.

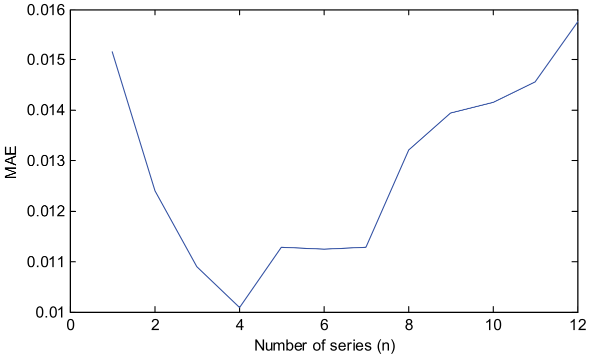


FIGURE 4. Average forecasting error with different number of series *n*.

We can find that the change trends between them are very similar which proves that the correlated time series can be used to forecast missing values. We employ mean absolute error (MAE) to measure the forecasting accuracy quantifiably and MAE is defined as

$$MAE = \sum_{t=1}^{N} \frac{|f_t - o_t|}{N}, \tag{2}$$

where f_t is the forecast value for the tth missing value, o_t is the real value of the tth missing value, and N is the total number of all the missing values.

For each time series from a to m, we forecast the missing values with n most similar time series and n range from 1 to 12. The simulation result is presented in Fig. 4. We can observe that the average forecasting error decreases significantly with the increasing of n when $n \leq 4$. This phenomenon can be explained that the information being used to forecast the missing values is not enough when n is too small. On the contrary, when $n \geq 4$, the average forecasting error increases with an increasing n. This strange phenomenon can be explained that somewhat uncorrelated time series may lead the forecasting process to the over-fitting problem which decreases the forecasting accuracy.

Another drawback for a large n is the time complexity, i.e., with the increasing of n, the average running time of the forecasting program increases significantly. As shown in Fig. 5, the running time of forecasting process increases significantly and the running time with $n = 12$ is about 6 times of that with $n = 1$.

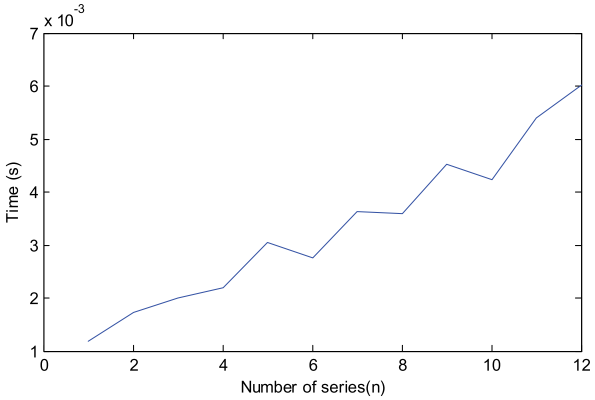


FIGURE 5. Running time of forecasting process with different number of series *n*.

In conclusion, too many uncorrelated time series cannot improve the forecasting accuracy through the interweaving process and even decrease the accuracy significantly. In addition, the redundant time series increase the time complexity of the forecasting process which makes it very difficult to forecast the future values in real time. Therefore, before forecasting the missing values based on interweaving module, we need to select a subset of all the giving time series to forecast the missing values of a time series. As presented in Section IV-B, we will select a proper subset time series based on clustering algorithms before forecasting the missing data.

B. A NOVEL DISTRIBUTED DATA FORECASTING ALGORITHM

We assume that each sensor node in a cluster update its reading periodically and then send the updated reading to the cluster head when it can communicate with the cluster head. The cluster head maintain the latest 10 reading for each node to measure the similarities between the time series generated by the nodes in the cluster. In this paper, to reduce energy consumption, the cluster heads select the corrected time series through a very simple method, i.e., selecting the most similar 5 time series to forecast the missing data as discussed in Section IV-A. For some special cases, the number of the members in a cluster may less than 5 and, in this case, we use all the time series in the cluster to forecast the missing data.

Algorithm 3 Pseudocode of Clusters Forming

```
Input: Node $n_j$, time period $(T_k, T_l)$
Output: The corresponding readings $R_s$ of $n_j$ in time period $T = (T_k, T_l)$
The base station sends a query message $Query\left(N_j, (T_k, T_l)\right)$ to a corresponding cluster head $ch_i$
$ch_i$ communicates with its cluster member $n_j$
if $n_j$ is normal then
    Query the readings $\mathbf{X}'$ of $N_q$ at time period of $(T_k, T_l)$
    Send $\mathbf{y}' = \mathbf{X}'$ to the base station
else
    Query the N latest conserved readings $\mathbf{y}$ of $n_j$
    Compute similarities between $\mathbf{y}$ and that of the other nodes in the same period by (1)
    Select the $q$ most related nodes $\mathbf{N_q}$ to $n_j$
    Query the readings $\mathbf{X}$ of all the nodes in $\mathbf{N_q}$
    Construct the relations between $\mathbf{X}$ and $\mathbf{y}$ by computing the coefficient matrix
        $\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}\right)$
    Query the readings $\mathbf{X}'$ of $\mathbf{N_q}$ at time period of $(T_k, T_l)$
    Computing the readings $\mathbf{y}'$ of $n_j$ at time period of $(T_k, T_l)$ by $\mathbf{y}' = \mathbf{X}'\mathbf{a}$
    Send $\mathbf{y}'$ to the base station
end if
```

The pseudocode of the data forecasting algorithm for WSNs is presented in Algorithm Algorithm 3. The total data query process presented in this algorithm is based on the query-response mode. In the initial, the base station sends a query message $Query\left(N_j, (T_k, T_l)\right)$ to a specified cluster head ch_i to get the node n_j's reading at time period of (T_k, T_l).

Note that the time period T should not be very far from the present time stamp, since the nodes cannot store all the historical data due to limited resources. Once a cluster head receives a query message, it communicates with all its cluster members to get the corresponding readings and responses a message containing $\mathbf{y}'$ to the base station. In this condition, the process is the same with the traditional approaches.

However, since most of the nodes in WSNs are not robust enough, it is likely that some of the nodes cannot provide the correct readings due to node failures or harsh transmission environment. In this case, we need to forecast the missing values by mining the strongly related time series. In this paper, we employ N readings of each node to forecast the missing values and N is a very important parameter that affect the forecasting accuracy. We assume that the set of nodes $\mathbf{N_q}$ is composed of the q most related nodes to n_j,

$$X = \begin{bmatrix} x_1[1] & x_2[1] & \cdots & x_q[1] \\ x_1[2] & x_2[2] & \cdots & x_q[2] \\ \vdots & \vdots & \ddots & \vdots \\ x_1[N] & x_2[N] & \cdots & x_q[N] \end{bmatrix} \tag{3}$$

is composed of the conserved readings of nodes in $\mathbf{N_q}$ and

$$y = \begin{bmatrix} y[1] \\ y[2] \\ \vdots \\ y[N] \end{bmatrix} \tag{4}$$

is the conserved readings of node n_j. We need to construct a relation between $\mathbf{y}$ and $\mathbf{X}$ by $\mathbf{y} = \mathbf{Xa}$, where we have

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{bmatrix}, \tag{5}$$

and it can be computed by minimizing

$$\mathsf{SE} = \sum_{i=1}^{N} \left\{ y[i] - \left(a_1 x_1 + a_2 x_2 + \cdots + a_q x_q\right) \right\}^2. \tag{6}$$

Furthermore, by setting $\frac{\partial \mathsf{SE}}{\partial a_1} = \frac{\partial \mathsf{SE}}{\partial a_2} = \cdots = \frac{\partial \mathsf{SE}}{\partial a_q} = 0$, we can get that $\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}\right)$. After obtaining $\mathbf{a}$, the cluster head can forecast the missing readings $\mathbf{y}'$ of n_j at time period of (T_k, T_l) by $\mathbf{y}' = \mathbf{X}'\mathbf{a}$ and then send $\mathbf{y}'$ to the base station.

Different from the traditional missing data forecasting approaches, a preprocessing step is added into the forecasting algorithm which can tackle the over-fitting problem discussed in the previous section. However, selecting the best parameter q to get an accurate forecasting result is very difficult. In this paper, we set q to 4 based on the observation in Section IV-A. Selecting a proper q can improve the forecasting accuracy and meanwhile decrease the energy consumption which will be illustrated in the simulations.

V. SIMULATIONS

In this section, we evaluate the performance of the proposed approach and compare it with MUSCLES [26], Yesterday and Auto-regression in terms of energy-efficiency and forecasting accuracy. We first illustrate the setup of the simulation including the network structure, generation of the nodes' readings, several measurements of the simulation results and some other default parameters. Then, the simulation results under the setup are detailed presented and discussed.

A. SIMULATION SETUP

To the best of knowledge, there is no generally accepted approach that integrates both node clustering and missing data forecasting into one framework considering both energy-efficiency and forecasting accuracy. MUSCLES, Yesterday and Auto-regression are three representative forecasting methods and it is impossible to introduce them into WSNs directly, because collecting all the data to the base station and forecasting the missing data in a centralized way is impossible in WSNs. Therefore, we manage the networks in a two-layer manner and a proper node clustering algorithm is need for all of three forecasting algorithms. In this paper, we employ LEACH as the clustering algorithm and divided the network into clusters, and then the cluster heads employ these forecasting algorithms to get the missing data.

In our simulation, we randomly scatter 200 (with 20 nodes that fails to generate the readings) nodes in a $100m \times 100m$ squared region and generate the readings of the nodes by the way discussed in Section IV-A. Each node generates a reading in a second and all the recent 100 readings are stored by the nodes. The base station randomly queries 10 nodes about their recent 5 readings respectively until at most 20 nodes are out of energy. To estimate the forecasting accuracy with different s which indicates the length of the readings when computing the similarities for time series in the process of forming clusters and different N which is the length of the readings that used in the process of forecasting missing values, we employ mean absolute error (MAE) which is defined as follows:

$$MAE = \sum_{t=1}^{N} \frac{|f_t - o_t|}{M}, \tag{7}$$

where f_t is the forecast value for the tth missing value, o_t is the real value of the tth missing value, and M is the total number of all the missing values . In addition to the measurement MAE, mean absolute percentage error (MAPE) is also used in this study and MAPE is defined as

$$MAPE = \frac{1}{M} \sum_{t=1}^{M} \left| \frac{f_t - o_t}{o_t} \right|. \tag{8}$$

Besides the forecasting accuracy, energy efficiency is also a big concern for WSNs. In our simulation, the ns-3 simulator implements a 1 Mb/s 802.11 MAC layer. As in [29] and [33], the energy consumption for sending a message is given by $len \times Eelec + len \times \varepsilon \times dist^2$, and for receiving a message, the energy consumption is $len \times Eelec$, where len is the length of a message, $dist$ is the distance of message transmission, $Eelec$ is set at 50 nJ/bit. As in [33], each sensor node begins with only $2J$ of energy. We employ the average energy consumption which measures the ratio of the total dissipated energy of the whole network per round (20s) to the number of the sensor nodes and the number of the nodes alive in the network to estimate the performances of the approaches in terms of energy efficiency. For different m (i.e., the number of clusters), the average energy consumption will be presented in Section V-B.

B. SIMULATION RESULTS AND DISCUSSION

In this section, we present the simulation results and discuss the advantages and disadvantages of the proposed approach.

Obviously, for the approaches of Yesterday, Auto-Regression and MUSCLES, parameter s has no affection on the performance, because all the three approaches employ LEACH-C as the node clustering method which do not contain the parameter s. As shown in Fig. 6, with the increasing of s, the MAE decreases for the propose approach and new approach always outperforms Yesterday and Auto-Regression in terms of forecasting accuracy. However, when $s < 4$, MUSCLES performs better than the new approach and, when $4 < s < 10$, the new approach performs better than MUSCLES. In addition, when $s > 7$, the MAE of the new approach performs stable and we set $s = 8$ in the following to decrease the energy consumption.

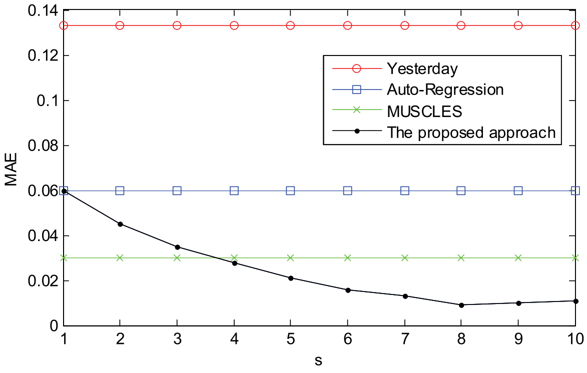


FIGURE 6. **MAE with different *s*.**

MAE is the most important measurement to evaluate the performance of the proposed approach and the simulation results are presented in Fig. 7. For Yesterday, there is no

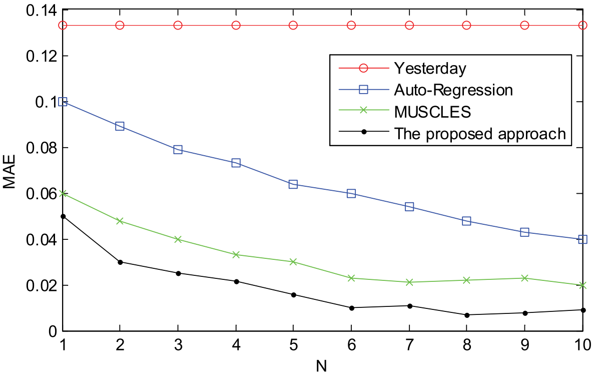


FIGURE 7. **MAE with different *N*.**

parameter N and the MAE is constant with different N. However, compared with the other three approaches, Yesterday is the easiest approach and has the worst performance. The MAE of all the other three approaches decrease with the increasing of N, because more and more valuable information are used in forecasting the missing data. However, MUSCLES and the proposed approach outperform Auto-Regression because the information contained in the co-evolving time series is also extracted. As discussed previously, the redundant information has negative affection on the forecasting accuracy and this is the reason why the proposed approach performs better compared with MUSCLES.

Another measurement that can illustrate the forecasting accuracy is MAPE and the simulation results are very similar with MAE, as shown in Fig. 8. Except for the missing data forecasting accuracy, energy efficiency is another important consideration in WSNs and we will use the average energy consumption and the number of nodes alive to evaluate these approaches.

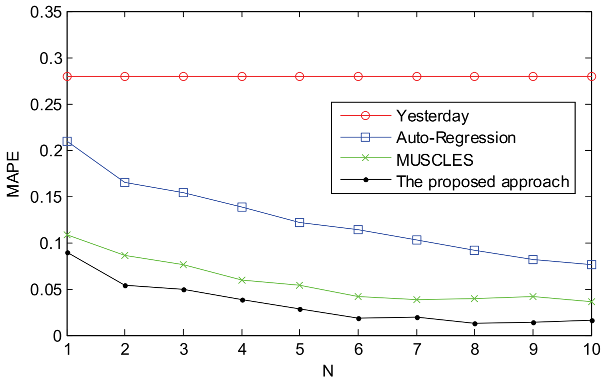


FIGURE 8. MAPE with different *N*.

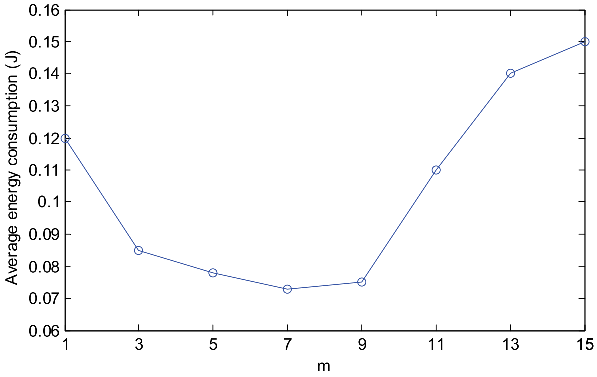


FIGURE 9. Average energy consumption with different *m*.

We present the average energy consumption per node per round with different number of clusters in Fig. 9. We can find that both too less and too many clusters consume much more energy compared with a proper number of clusters. In case of the assumed topology of the network, 7 to 9 clusters are good choices. For convenience, in the following experiment, m is set to 8.

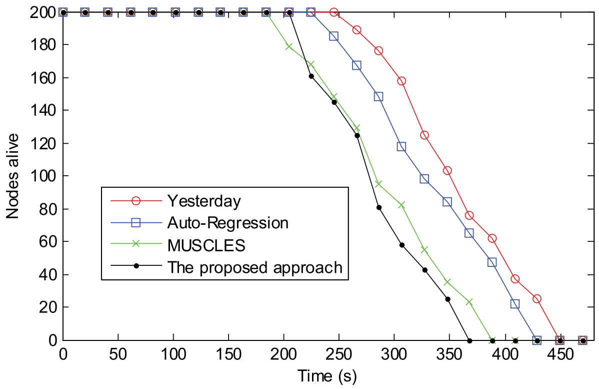


FIGURE 10. Number of nodes alive over time.

As shown in Fig. 10, both Yesterday and Auto-Regression performs very well in terms of prolonging the lifetime of the network. At the 300^{th} second, more than 80% nodes are still working which is much larger than that of MUSCLES and the proposed approach. This can be explained that more energy is consumed in forecasting the missing data because of the high computing complex. MUSCLES and the proposed approach have a similar performance in term of energy efficiency.

Overall, the proposed approach outperforms the existing forecasting algorithms in terms of forecasting accuracy significantly, because more valuable information is fully used. On the other hand, more energy is consumed by the proposed approach. This can be explained that some extra energy are used in the process of forming clusters and forecast the missing data. For the networks that used to monitor the environment with high requirements of accuracy, both MUSCLES and the proposed approach are good choices for the user. However, the proposed approach outperforms MUSCLES in terms of accuracy with very similar energy consumption.

VI. CONCLUSION AND FUTURE RESEARCH

In this paper, we propose a novel clustering algorithm for WSNs, which takes both the energy-efficiency problem and data correlations between the nodes into the unified consideration. The nodes can be assigned to a same cluster only when they have both close space distances and data correlations. Unfortunately, there are always some outliers that their readings are uncorrelated with most of their neighbors and it is impractical to generate some clusters for the outliers only. Therefore, after generating the clusters for most of the nodes, the outliers are assigned to the existing clusters based on the distances between the outliers and the cluster heads. On one hand, close distances between the nodes make it energy-efficient for the nodes in a cluster to communicate with each other. On the other hand, high data correlations make it accurate for the cluster head to forecast the missing data. In addition, we design a distributed missing values forecasting approach to decrease data transmission in the network. Different from the traditional forecasting approaches, a pre-processing stage is integrated into the framework. Only the high correlated data streams of the nodes are used to forecast the missing data with each other and the uncorrelated data streams are ignored.

For the future work, we will focus on reducing the energy consumption further to execute more in-network data processing and prolong the lifetime of the networks. Besides, it is interesting to design a pairwise-nodes correlation measurement monitoring system which is executed by the cluster head. In the system, the pairwise correlations need to be updated in real time which is particular important for real time forecasting.

REFERENCES

[1] M. R. Palattella *et al.*, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[3] S. Cirani and M. Picone, "Wearable computing for the Internet of Things," *IT Prof.*, vol. 17, no. 5, pp. 35–41, Sep. 2015.

[4] M. Simsek, A. Aijaz, M. Dohler, and J. Sachs, "5G-enabled tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[5] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[6] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[7] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[8] Y. Shi, J. Zhang, K. B. Letaief, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

[9] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[10] P. Derler, E. A. Lee, and A. S. Vincentelli, "Modeling cyber–physical systems," *Proc. IEEE*, vol. 100, no. 1, pp. 13–28, Jan. 2012.

[11] Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4729–4743, Sep. 2015.

[12] K. Yang, Y. Shi and Z. Ding, "Low-rank matrix completion for mobile edge caching in fog-RAN via riemannian optimization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[13] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental tradeoff between computation and communication in distributed computing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1814–1818.

[14] X. Zhang, Z. Yang, W. Sun, and Y. Liu, "Incentives for mobile crowd sensing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 54–67, 1st Quart., 2016.

[15] Y. Zhao and Q. Han, "Spatial crowdsourcing: Current state and future directions," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 102–107, Jul. 2016.

[16] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, "Sparse mobile crowdsensing: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 161–167, Jul. 2016.

[17] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Apr. 2009.

[18] R. Zhang, J. Shi, Y. Zhang, and C. Zhang, "Verifiable privacy-preserving aggregation in people-centric urban sensing systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 268–278, Sep. 2013.

[19] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, "Practical data prediction for real-world wireless sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2231–2244, Aug. 2015.

[20] D. Baker and A. Ephremides, "The architectural organization of a mobile radio network via a distributed algorithm," *IEEE Trans. Commun.*, vol. 29, no. 11, pp. 1694–1701, Nov. 1981.

[21] D. J. Baker, A. Ephremides, and J. Flynn, "The design and simulation of a mobile radio network with distributed control," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 1, pp. 226–237, Jan. 1984.

[22] N. Zhao, F. R. Yu, and V. C. M. Leung, "Opportunistic communications in interference alignment networks with wireless power transfer," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 88–95, Feb. 2015.

[23] N. Zhao, F. R. Yu, H. Sun, and M. Li, "Adaptive power allocation schemes for spectrum sharing in interference-alignment-based cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3700–3714, May 2016.

[24] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[25] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 12–30, Nov. 2015.

[26] B.-K. Yi, N. D. Sidiropoulos, T. Johnson, A. Biliris, H. V. Jagadish, and C. Faloutsos, "Online data mining for co-evolving time sequences," in *Proc. 16th Int. Conf. Data Eng.*, 2000, pp. 13–22.

[27] S. Banerjee and S. Khuller, "A clustering scheme for hierarchical control in multi-hop wireless networks," in *Proc. IEEE 12th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 2. Apr. 2001, pp. 1028–1037.

[28] H. Zhang and A. Arora, "GS^3: Scalable self-configuration and self-healing in wireless sensor networks," *Comput. Netw. Int. J. Comput. Telecommun. Netw.*, vol. 43, no. 4, pp. 459–480, 2003.

[29] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 660–670, Oct. 2002.

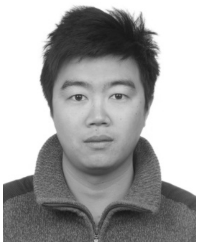
[30] S. Zoller, C. Vollmer, M. Wachtel, R. Steinmetz, and A. Reinhardt, "Data filtering for wireless sensor networks using forecasting and value of information," in *Proc. IEEE Conf. Local Comput. Netw.*, Oct. 2013, pp. 441–449.

[31] J. Pardo, F. Zamora-Martínez, and P. Botella-Rocamora, "Online learning algorithm for time series forecasting suitable for low cost wireless sensor networks nodes," *Sensors*, vol. 15, no. 4, pp. 9277–9304, 2015.

[32] T. Murata and H. Ishibuchi, "Performance evaluation of genetic algorithms for flowshop scheduling problems," in *Proc. 1st IEEE Conf. Evol. Comput., IEEE World Congr. Comput. Intell.*, vol. 2. Jun. 1994, pp. 812–817.

[33] M. Wu, J. Xu, X. Tang, and W. C. Lee, "Top-k monitoring in wireless sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 962–976, Jul. 2007.

XUAN LIU received the B.S. degree in communication engineering from Jilin University, Jilin, China, in 2010, and the M.S. degree from The Hong Kong University of Science and Technology, in 2012. She is currently pursuing the Ph.D. degree in wireless engineering with The University of Sydney, Sydney, Australia. Her research interests include wireless sensor networks and data forecasting.



JUN LI (M'09–SM'16) received the Ph. D degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. In 2009, he was with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell, as a Research Scientist. From 2009 to 2012, he was a Post-Doctoral Fellow with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. From 2012 to 2015, he was a Research Fellow with the School of Electrical Engineering, The University of Sydney, Australia. Since 2015, he has been a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, channel coding theory, wireless network coding, and resource allocations in cellular networks.

ZY DONG (M'99–SM'06–F'16) received the Ph.D. degree from The University of Sydney, Australia, in 1999. He previously held academic and industrial positions with the Hong Kong Poly technic University, the University of Queensland, Australia, and also with Transend Networks, Australia. He is currently a Professor and the Head of the School of Electrical and Information Engineering, The University of Sydney, Australia. He is immediate Ausgrid Chair Professor and the Director of the Center for Intelligent Electricity Networks , University of Newcastle, Australia. His research interest includes smart grid, power system planning, power system security, load modeling, renewable energy systems, electricity market, and computational intelligence and its application in power engineering. He is an Editor of the IEEE Transactions on Smart Grid, and the IEEE Power Engineering Letters.

FEI XIONG received the Ph.D. degree in communication and information system from Beijing Jiaotong University, Beijing, China, in 2013. He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. His current research interests include the areas of web mining, network analysis and complex systems.