Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2012, Article ID 427246, 10 pages
doi:10.1155/2012/427246

Research Article

PCA-Guided Routing Algorithm for Wireless Sensor Networks

Gong Chen, Liansheng Tan, Yanlin Gong, and Wei Zhang

Department of Computer Science, Central China Normal University, Wuhan 430079, China

Correspondence should be addressed to Liansheng Tan, liansheng tan688@gmail.com

Received 13 August 2012; Revised 24 October 2012; Accepted 11 November 2012

Academic Editor: Hongxiang Li

An important performance concern for wireless sensor networks (WSNs) is the total energy dissipated by all the nodes in the network over the course of network lifetime. In this paper, we propose a routing algorithm termed as PCA-guided routing algorithm (PCA-RA) by exploring the principal component analysis (PCA) approach. Our algorithm remarkably reduces energy consumption and prolongs network lifetime by realizing the objective of minimizing the sum of distances between the nodes and the cluster centers in a WSN network. It is demonstrated that the PCA-RA can be efficiently implemented in WSNs by forming a nearly optimal K-means-like clustering structure. In addition, it can decrease the network load while maintaining the accuracy of the sensor measurements during data aggregating process. We validate the efficacy and efficiency of the proposed algorithm by simulations. Both theoretical analyses and simulation results demonstrate that this algorithm can perform significantly with less energy consumption and thus prolong the system lifetime for the networks.

1. Introduction

Wireless sensor networks (WSNs) [1] consist of battery-powered nodes which inherit sensing, computation, and wireless communication capabilities. Although there have been significant improvements in processor design and computing issues, limitations in battery provision still exist, bringing energy resource considerations as the fundamental challenge in WSNs. Consequently, there have been active research efforts devoted to lifting the performance limitations of WSNs. These performance limitations include network throughput, energy consumption and, network lifetime. Network throughput typically refers to the maximum amount of packets that can be successfully collected by the cluster heads (CHs) in the network, energy consumption refers to the minimize energy dissipation that nodes in the network consume, and network lifetime refers to the maximum time limit that nodes in the network remain alive until one or more nodes drain up their energy.

The routing algorithms have been specifically designed for WSNs because the energy optimization is an essential design issue. A good routing scheme is helpful in improving these performance limits such as reducing the energy consumption, prolonging the network lifetime, and increasing the network throughput. Network researchers have studied a great variety of routing protocols in WSNs differing based on the application and network architecture. As demonstrated in [2, 3], it can be classified into four categories: flit, hierarchical clustering, location-based routing, and QoS-based routing. The current routing protocols have their own design trade-offs between energy and communication overhead savings, as well as the advantages and disadvantages of each routing technique.

As per the representative hierarchical clustering protocol, low energy adaptive clustering hierarchy (LEACH) [4] has simplicity, flexibility, and scalability because its manipulations rely on randomized rotation of the cluster heads (CHs), but its features of unregulated distribution, unbalanced clustering structure, uniform initial energy, and so on, hinder its performance. Based on LEACH, there are many variants, such as [5–8]. LEACH-E [5] more likely selects the nodes with higher energy as the CHs. LEACH-C [5] analytically determines the optimum number of CHs by taking into account the energy spent by all clusters. PEGASIS [6], TEEN [7], and ATEEN [8]

improve the energy consumption by optimizing the data transmission pattern. HEED [9] is a complete distributed routing protocol which has different clustering formations and cluster-heads selecting measures. These protocols have many restrictive assumptions and applicable limitations, so it has great improvement space and extensibility. The rapid development of wireless communications technology, and the miniaturization and low cost of sensing devices, have accelerated the development of wireless sensor networks (WSNs) [10, 11]. As in [12], Zytoune et al. proposed a uniform balancing energy routing protocol (UBERP). The BP K-means [13] and BS K-means [14] can improve the structure of clusters and perform better load-balance and less energy consumptions. HMP-RA [15] proposes a solution to address this issue through a hybrid approach that combines two routing strategies, flat multihop routing and hierarchical multihop routing. ESCFR and DCFR can map small changes in nodal remaining energy to large changes in the function value and consider the end-to-end energy consumption and nodal remaining energy [16]. Biologically inspired intelligent algorithms build a hierarchical structure on the network for different kinds of traffic, thus maximizing network utilization, while improving its performance [17]. In the case where sensor nodes are mobile, as in [18, 19], the nodes can adjust their position to help balance energy consumption in areas that have high transmission load and/or mitigate network partition.

In this paper, we consider an overarching algorithm that encompasses both performance metrics. It desires to minimize the sum of distances in the clusters. We show that the principal component analysis (PCA) [20], a useful statistical technique that has found application in fields such as face recognition and image compression, and a common technique for finding patterns in data of high dimension, can form a near-optimal K-means-like clustering structure, in which the distance between the non-CH nodes and CHs is near minimized.

Moreover, the data aggregating issue associated with the measurements accuracy calls for a careful consideration in scheme about data collecting and fusing. In this paper, we investigate the PCA technology in a high relative measurements context for WSNs. Our objective is to obtain a good approximation to sensor measurements by relying on a few principal components while decreasing the network load.

The remainder of this paper is organized as follows. In Section 2, we describe the network and energy model. Section 3 presents the PCA-guided routing algorithm (PCA-RA) model and gives numerical results to demonstrate the working mechanism of PCA-RA. Section 4 discusses the PCA-RA algorithm solution strategies. In Section 5, we simulate the PCA-RA and compare it with LEACH and LEACH-E. Finally, Section 6 concludes this paper.

2. Syestem Model

2.1. The Network Model. Let us consider a two-tier architecture for WSNs. Figure 1 shows the physical network topology for such a network. There are three types of nodes in the networks, namely, a base station (BS), the cluster-head nodes (CHNs), and the ordinary sensor nodes (OSNs).

For each cluster of sensor nodes, there is one a CHN, which is different from an OSN in terms of functions. The primary functions of a CHN are (1) data aggregation for data measurements from the local clusters of OSNs and (2) relaying the aggregated information to the BS. For data fusion, a CHN analyzes the content of each measurement it receives and exploits the correlation among the data measurements. An CHN has a limited lifetime, so we need consider rotating the CHNs to balance to the energy consumption.

The third component is the BS. We assume that there is sufficient energy resource available at the BS and thus there is no energy constraint at the BS.

2.2. The Energy Consumption Model. We compute the energy consumption using the first-order radio model [5]. The equations, which are used to calculate transmission costs and receiving costs for an L-bit message to cross a distance d, are shown below,

$$E_{\mathrm{TX}}(L,d) = \begin{cases} L(E_{\mathrm{elec}} + \varepsilon_{\mathrm{fs}} d^2), & d < d_0 \\ L\left(E_{\mathrm{elec}} + \varepsilon_{\mathrm{mp}} d^4\right), & d \geq d_0, \end{cases} \quad (1)$$

$$E_{\mathrm{RX}}(L) = LE_{\mathrm{elec}}.$$

In (1), the electronics energy, E_{elec}, depends on factors such as the digital coding, modulation, filtering, and spreading of the signal, whereas the amplifier energy, $\varepsilon_{\mathrm{fs}} d^2$ or $\varepsilon_{\mathrm{mp}} d^4$, depends on the distance to the receiver and the acceptable bit-error rate. d_0 is the distance threshold.

3. PCA-Guided Routing Algorithm Model

PCA is a classic technique in statistical data analysis, data compression, and image processing. PCA transforms a number of correlated variables into a number of uncorrelated variables called principal components. The objective of PCA is to reduce the dimensionality of the dataset, but not only retain most of the original variability in the data. The first principal component accounts for as much of the variability in the data as possible. Mathematically, how to pick up the dimensions with the largest variances is equivalent to finding the best low-rank approximation of the data via the singular value decomposition (SVD) [21].

The design of routing algorithm is important in wireless sensor networks. Although plenty of interests are drawn on it, there is still a challenge to face on the aspect of efficiency and energy consumption. In this section, we will describe the notations about PCA-guided routing algorithm model firstly and then propose the PCA-guided clustering model, and finally, we will present the PCA-guided data aggregating model.

3.1. Notations. Let $X = \{x_1, x_2, \ldots, x_n\}$ represents the location coordinate matrix of a set of n sensors, $Y = \{y_1, y_2, \ldots, y_n\}$ represents the centered data matrix, where

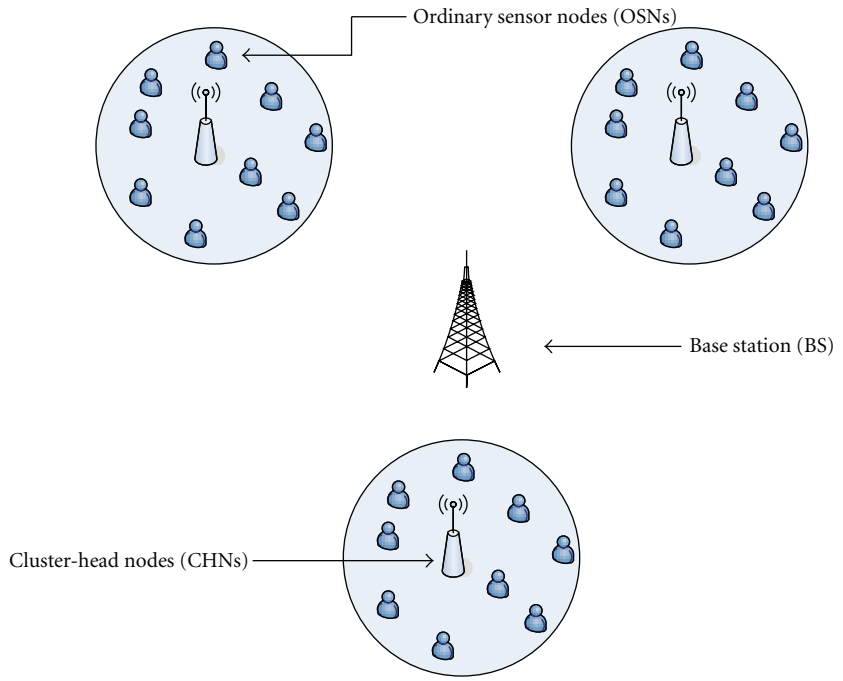

Figure 1: Physical topology for two-tier wireless sensor networks.

$y_i = x_i - \overline{x}$, which defines the centered distance vector column wise, and $\overline{x} = \sum_i x_i/n$ is the mean vector column wise of X matrix.

Let $M = \{m_1, m_2 \cdots m_p\}$ be a group of measurements collecting from the sampling period. Each sensor generates a stream of data. Let $D_{N\times P}$ be a matrix with elements D_{ij}, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, being the measurement taken by sensor i at point j. Let $Q_{N\times P}$ be a centered matrix with elements $q_{ij} = d_{ij} - d_{i.}/n$, $d_{i.} = \sum_j d_{ij}$.

3.2. PCA-Guided Clustering Model. We define the equation for the SVD of matrix Y [22] as follows:

$$Y = \sum_k \lambda_k u_k v_k. \tag{2}$$

The covariance matrix (ignoring the factor $1/n$) is

$$\sum_i (x_i - \overline{x})^T (x_i - \overline{x}) = Y^T Y. \tag{3}$$

The principal components v_k are eigenvectors satisfying

$$Y^T Y v_k = \lambda_k^2 v_k, \qquad v_k = \frac{Y^T u_k}{\lambda_k^2}. \tag{4}$$

3.2.1. K-Means Clustering Model. According to [23, 24], we find the PCA dimension reduction automatically by performing data clustering according to the K-means objective function [25, 26]. Using K-means algorithm, it can form a better cluster structure by minimizing the sum of squared errors. We define the squared distance between sensor nodes and cluster centers as

$$J_K = \sum_{k=1}^{K} \sum_{i\in C_k} (x_i - m_k)^2, \tag{5}$$

where $m_k = \sum_{x_i\in C_k} x_i/n_k$ is the center of cluster C_k and n_k is the number of sensor nodes in C_k. Given the fact that by minimizing the distance between sensor nodes and cluster centers, the energy consumption can be effectively reduced. Our clustering algorithm is thus designed to be capable of minimizing the above metric J_K.

For the sake of convenience, let us start with the case of $K = 2$. To obtain the explicit expression for J_k, letting

$$d\left(C_p, C_l\right) = \sum_{i\in C_p} \sum_{j\in C_l} \left(x_i - x_j\right)^2 \tag{6}$$

be the sum of squared distances between two clusters C_p and C_l, after some algebra one obtains the following:

$$J_2 = \frac{d(C_1, C_1)}{2n_1} + \frac{d(C_2, C_2)}{2n_2}, \tag{7}$$

where n_1 and n_2 are the numbers of sensor nodes in C_1 and C_2, n is the total number of sensor nodes; therefore, we get $n = n_1 + n_2$.

If denoting

$$\overline{y^2} = \frac{\sum_i y_i^T y_i}{n} = \frac{d(C_1, C_1)}{2n^2} + \frac{d(C_2, C_2)}{2n^2} + \frac{d(C_1, C_2)}{n^2}, \tag{8}$$

$$J_D = \frac{n_1 n_2}{n}\left[2\frac{d(C_1, C_2)}{n_1 n_2} - \frac{d(C_1, C_1)}{n_1^2} - \frac{d(C_2, C_2)}{n_2^2}\right], \tag{9}$$

we thus have

$$\begin{aligned} n\overline{y^2} - \frac{1}{2}J_D &= n\left(\frac{d(C_1,C_1)}{2n^2} + \frac{d(C_2,C_2)}{2n^2} + \frac{d(C_1,C_2)}{n^2}\right) \\ &\quad - \frac{1}{2}\left(\frac{n_1n_2}{n}\left[2\frac{d(C_1,C_2)}{n_1n_2} - \frac{d(C_1,C_1)}{n_1^2}\right.\right. \\ &\quad \left.\left. - \frac{d(C_2,C_2)}{n_2^2}\right]\right) \\ &= \frac{d(C_1,C_1)}{2n} + \frac{d(C_2,C_2)}{2n} + \frac{d(C_1,C_2)}{n} \\ &\quad - \frac{2n_1n_2d(C_1,C_2)}{2nn_1n_2} + \frac{n_1n_2d(C_1,C_1)}{2nn_1^2} \\ &\quad + \frac{n_1n_2d(C_2,C_2)}{2nn_2^2} \\ &= \left(\frac{d(C_1,C_1)}{2n} + \frac{n_1n_2d(C_1,C_1)}{2nn_1^2}\right) \\ &\quad + \left(\frac{d(C_2,C_2)}{2n} + \frac{n_1n_2d(C_2,C_2)}{2nn_2^2}\right) \\ &\quad + \left(\frac{d(C_1,C_2)}{n} - \frac{2n_1n_2d(C_1,C_2)}{2nn_1n_2}\right) \\ &= \left(\frac{n_1d(C_1,C_1)}{2nn_1} + \frac{n_2d(C_1,C_1)}{2nn_1}\right) \\ &\quad + \left(\frac{n_2d(C_2,C_2)}{2nn_2} + \frac{n_1d(C_2,C_2)}{2nn_2}\right) \\ &= \frac{(n_1+n_2)d(C_1,C_1)}{2nn_1} + \frac{(n_2+n_1)d(C_2,C_2)}{2nn_2} \\ &= \frac{d(C_1,C_1)}{2n_1} + \frac{d(C_2,C_2)}{2n_2} = J_2. \end{aligned} \tag{10}$$

That is

$$J_2 = n\overline{y^2} - \frac{1}{2}J_D, \tag{11}$$

where $\overline{y^2}$ is a constant and it denotes the distance between the sensor nodes and the center for all nodes; thus min (J_K) is equivalent to max (J_D), and because the two resulting clusters are as separated and compact as possible. Because the averaged intracluster distance is greater than the sum of the averaged intercluster distances, that is

$$\frac{d(C_1,C_2)}{n_1n_2} - \frac{d(C_1,C_1)}{n_1^2} - \frac{d(C_2,C_2)}{n_2^2} > 0. \tag{12}$$

From (9), it is seen that J_D is always positive. This is to say, evidenced from (11), for $K = 2$ minimization of cluster objective function J_K is equivalent to maximization of the distance objective J_D, which is always positive.

When $K > 2$, we can do a hierarchical divisive clustering, where each step using the $K = 2$ is a clustering procedure. This procedure can get an approximated K-means clustering structure.

3.2.2. PCA-Guided Relaxation Model. In [24], it proves that the relaxation solution of J_D can get via the principal component. It sets the cluster indicator vector to be

$$q(i) = \begin{cases} \sqrt{\dfrac{n_2}{nn_1}}, & \text{if } i \in C_1 \\ -\sqrt{\dfrac{n_1}{nn_2}}, & \text{if } i \in C_2. \end{cases} \tag{13}$$

The indicator vector satisfies the sum-to-zero and normalization conditions. Consider the squared distance matrix $H = (h_{ij})$, where $h_{ij} = \|x_i - x_j\|^2$. $q^THq = -J_D$ is easily observed.

(1) *The First Relaxation Solution.* Let q take any value in $[-1, 1]$; the solution of minimization of $J(q) = q^THq/q^Tq$ is given by the eigenvector corresponding to the lowest eigenvalue of the equation $Hz = \lambda z$.

(2) *The Second Relaxation Solution.* Let $\hat{H} = (\hat{h}_{ij})$, where the element is given by

$$\hat{h}_{ij} = h_{ij} - \frac{h_{i.}}{n} - \frac{h_{.j}}{n} + \frac{h_{..}}{n^2}, \tag{14}$$

in which, $h_{i.} = \sum_j h_{ij}$, $h_{.j} = \sum_i h_{ij}$, $h_{..} = \sum_{ij} h_{ij}$.

After computing, we have $q^T\hat{H}q = q^THq = -J_D$, then relaxing the restriction q, the desired cluster indicator vector is the eigenvector corresponding to the lowest eigenvalue of $\hat{H}z = \lambda z$.

(3) *The Third Relaxation Solution.* With some algebra, we can obtain $\hat{H} = -2Y^TY$. Therefore, the continuous solution for cluster indicator vector is the eigenvector corresponding to the largest eigenvalue of the covariance matrix Y^TY which by definition, is precisely the principal component v_1.

3.2.3. PCA-Guided Clustering Model. According to the mentioned above, for K-means clustering where $K = 2$, the continuous solution of the cluster indicator vector is the principal component v_1, that is, clusters C_1 and C_2 are given by

$$C_1 = \{i \mid v_1(i) \le 0\}, \qquad C_2 = \{i \mid v_1(i) > 0\}. \tag{15}$$

We can consider using PCA technology to clustering sensor nodes for WSNs. It near minimizes the sum of the distances between the sensor nodes and cluster centers.

Example 1. Let us assume in a WSN there are 20 sensors distributed in the network. X represents the 2D coordinate matrix.

$$X = \begin{bmatrix} 29.471, & 4.9162, & 69.318, & 65.011, & 98.299, \\ 55.267, & 40.007, & 19.879, & 62.52, & 73.336, \\ 37.589, & 0.98765, & 41.986, & 75.367, & 79.387, \\ 91.996, & 84.472, & 36.775, & 62.08, & 73.128; \\ 19.389, & 90.481, & 56.921, & 63.179, & 23.441, \\ 54.878, & 93.158, & 33.52, & 65.553, & 39.19, \\ 62.731, & 69.908, & 39.718, & 41.363, & 65.521, \\ 83.759, & 37.161, & 42.525, & 59.466, & 56.574 \end{bmatrix}. \tag{16}$$

Compute the eigenvector of the matrix, $Y^T Y$, that is, the principal component v_1:

$$v_1^T = \begin{bmatrix} -0.11623, & -0.47282, & 0.10624, & 0.058301, \\ 0.40896, & 0.0013808, & -0.20539, & -0.22552, \\ 0.033439, & 0.17823, & -0.15446, & -0.45626, \\ -0.067391, & 0.18908, & 0.16501, & 0.22147, \\ 0.2697, & -0.11445, & 0.04397, & 0.13673 \end{bmatrix}. \tag{17}$$

If $v_1(i) \leqslant 0$, sensor i belongs to C_1, otherwise it belongs to C_2. We depicted the above clustering results into Figure 2.

When $K > 2$, we do a hierarchical divisive clustering, where each step uses the $K = 2$ clustering procedure.

In summary, a PCA-guided clustering model can be used to form a nearly optimal K-means-like clustering structure.

3.3. PCA-Guided Data Aggregating Model. As mentioned above, $D_{N\times P}$ represents the data measurement matrix collected from the sampling period by the cluster heads. Q is a centered matrix about D.

The vector $\{w_k\}_{1\leq k\leq N}$ is the principal components satisfying

$$QQ^T w_k = \lambda_k w_k. \tag{18}$$

Because in most cases there exist high correlations between sensor measurements, good approximations to sensor measurements can be obtained by relying on few principal components. The first principal component accounts for as much of the variability in the data as possible, so we can find the first principal component vector w_1, such that their projection data can effectively express the original measurements. Approximations $\hat{Q}$ to Q are obtained by

$$\hat{Q} = w_1 w_1^T Q = w_1 Z, \tag{19}$$

where

$$Z = w_1^T Q. \tag{20}$$

The CHs only send three packets about Z, w_k, and the mean vector column-wise G of D matrix to the base station:

$$G = \left(g_i = \frac{d_{i\cdot}}{p}\right), \qquad d_{i\cdot} = \sum_j d_{ij}. \tag{21}$$

Because the mean vector column-wise G is subtracted by the CHs prior to the aggregation of its value, the base station can add back after the computation of the approximation.

Example 2. If the CH collects the matrix D as follow:

$$D = \begin{bmatrix} 20.3 & 20.2 & 20.8 & 20.3 & 20.3 & 20.4 & 20.5 & 20.4 \\ 20.4 & 20.3 & 20.6 & 20.4 & 20.4 & 20.5 & 20.6 & 20.5 \\ 20.2 & 20.1 & 20.4 & 20.2 & 20.2 & 20.3 & 20.4 & 20.3 \\ 20.3 & 20.2 & 20.8 & 20.3 & 20.3 & 20.4 & 20.5 & 20.4 \\ 20.3 & 20.2 & 20.8 & 20.3 & 20.3 & 20.4 & 20.5 & 20.4 \\ 20.4 & 20.3 & 20.6 & 20.4 & 20.4 & 20.5 & 20.6 & 20.5 \\ 20.2 & 20.1 & 20.4 & 20.2 & 20.2 & 20.3 & 20.4 & 20.3 \\ 20.3 & 20.2 & 20.8 & 20.3 & 20.3 & 20.4 & 20.5 & 20.4 \\ 20.3 & 20.2 & 20.8 & 20.3 & 20.3 & 20.4 & 20.5 & 20.4 \\ 20.2 & 20.1 & 20.4 & 20.2 & 20.2 & 20.3 & 20.4 & 20.3 \end{bmatrix}, \tag{22}$$

Then the CH can compute the matrix Q and the principal components w_1,

$$\begin{aligned} Z = w_1^T Q &= \begin{bmatrix} -0.26306, & -0.56555, & 0.93394, \\ -0.26306 & -0.26306, & 0.039433, \\ 0.34193, & 0.039433 \end{bmatrix}, \\ w_1 &= \begin{bmatrix} 0.39468, & 0.21031, & 0.21031, & 0.39468, \\ 0.39468, & 0.21031, & 0.21031, & 0.39468, \\ 0.39468, & 0.21031 \end{bmatrix}, \\ G &= \begin{bmatrix} 20.4, & 20.462, & 20.262, & 20.4, & 20.4 \\ 20.462, & 20.262, & 20.4, & 20.4, & 20.262 \end{bmatrix}, \end{aligned} \tag{23}$$

then the packet about Z, w_1, and G are delivered to the base station. The base station will compute the approximation $\hat{Q}$.

After adding back the subtracted mean value G, we can obtain:

$$\widehat{D} = \begin{bmatrix} 20.296 & 20.177 & 20.769 & 20.296 & 20.296 & 20.414 & 20.535 & 20.416 \\ 20.407 & 20.344 & 20.659 & 20.407 & 20.407 & 20.471 & 20.534 & 20.471 \\ 20.207 & 20.144 & 20.459 & 20.207 & 20.207 & 20.271 & 20.334 & 20.271 \\ 20.296 & 20.177 & 20.769 & 20.296 & 20.296 & 20.414 & 20.535 & 20.416 \\ 20.296 & 20.177 & 20.769 & 20.296 & 20.296 & 20.414 & 20.535 & 20.416 \\ 20.407 & 20.344 & 20.659 & 20.407 & 20.407 & 20.471 & 20.534 & 20.471 \\ 20.207 & 20.144 & 20.459 & 20.207 & 20.207 & 20.271 & 20.334 & 20.271 \\ 20.296 & 20.177 & 20.769 & 20.296 & 20.296 & 20.416 & 20.535 & 20.416 \\ 20.296 & 20.177 & 20.769 & 20.296 & 20.296 & 20.416 & 20.535 & 20.416 \\ 20.207 & 20.144 & 20.459 & 20.207 & 20.207 & 20.271 & 20.334 & 20.271 \end{bmatrix}. \tag{24}$$

4. PCA-Guided Routing Algorithm Solution Strategies

In Section 3, we have actually proposed a PCA-guided clustering and data aggregating model for routing optimization problem in WSNs by theoretical analyses and numerical examples. The following steps provide an overview of the solution strategy.

4.1. Initialization Stage. In the first stage, we assume that a set of N location coordinates are gathered at the base station. The base station computes the first principal component v_1. The two clusters C_1 and C_2 are determined via v_1 according to (15) by the BS.

4.2. Clusters Splitting Stage. When the number of sensor nodes is huge, two clusters are not enough and can induce the energy consume rapidly, considering splitting these clusters whose memberships are more than the CH can support.

If there are K clusters, there are on average N/K nodes per cluster (one CH node and non-CH nodes). We define that

$$\text{Ave} = N/K. \tag{25}$$

We can estimate the average energy dissipated per round to get the most energy efficient number of the clusters as follows:

$$\begin{aligned} E_{\text{round}} &= k(E_{\text{CH}} + E_{n\text{-CH}}) \\ &= k\left[lE_{\text{elec}}\left(\frac{N}{k} - 1\right) + alE_{\text{elec}} + al\varepsilon_{\text{mp}} d_{\text{toBS}}^4 \right. \\ &\quad \left. + \left(\frac{N}{k} - 1\right)\left(lE_{\text{elec}} + l\varepsilon_{\text{fs}} d_{\text{toCH}}^2\right)\right]. \end{aligned} \tag{26}$$

The notation and definition of the parameters in (26) are described as Table 1.

We can get

$$E_{\text{round}} = l\left[(2N + k)E_{\text{elec}} + 3k\varepsilon_{\text{mp}} d_{\text{toBS}}^4 + \varepsilon_{\text{fs}}\frac{M^2}{2\pi}\frac{N}{k} - \varepsilon_{\text{fs}}\frac{M^2}{2\pi}\right]. \tag{27}$$

According to the average energy dissipated per round, the scope of the clusters' number can be estimated when it is the most energy efficient. In [4, 13, 14], the authors use the average energy dissipated per round to obtain the optimal cluster number. Based on this methodology, here in this paper, we study the appropriate upper limit of the cluster nodes to perform the clusters splitting and thus extend this methodology.

Example 3. If we assume the number of sensor nodes is 100, the average energy dissipated per round as the number of clusters is varied between 1 and 20. Figure 3 shows that it is most energy efficient when there are between 3 and 5 clusters in the 100-node network. We define that the most appropriate number of clusters is varied from 3 to 5 because

$$\begin{aligned} \frac{(E_{\text{round}}(3) - E_{\text{round}}(\min))}{E_{\text{round}}(\min)} &< 0.03, \\ \frac{(E_{\text{round}}(5) - E_{\text{round}}(\min))}{E_{\text{round}}(\min)} &< 0.03. \end{aligned} \tag{28}$$

We obtain that the appropriate upper limit of the cluster nodes is 100/3 = 34.

If the number of cluster nodes is more than 34, the PCA-guided clustering algorithm will implement to split it.

4.3. Cluster Balancing Stage. We use the BS running the PCA-guided clustering algorithm to divide sensor nodes based on the geographical information. We get K clusters from N nodes in the field rapidly and form better clusters by dispersing the CHs throughout the network.

Now let us introduce the basic idea of the cluster balancing stage. Above, we get the number of clusters K if there are N nodes. We define the average node number per cluster as Ave. The cluster-balanced step is added in each iteration process. If $|C_j| > \text{Ave}$, $1 \le j \le K$, the BS computes $\text{dist}(s_i, \overline{u_q})$, where $s_i \in C_j, j \ne q$ and $|C_q| < \text{Ave}$. This means to compute the distances between the nodes and each cluster center whose cluster's node number is less than Ave. The BS gets s_i if its value is minimum and adjusts the node into the corresponding cluster computed. After implementing the cluster-balanced step, the BS limit the node number in each

Table 1: Parameter description.

Parameters	Head description
k	The number of the clusters
N	The total number of the sensor nodes
E_{elec}	The energy consumption per bit when sending and receiving
l	The sending data bit
$\varepsilon_{mp}/\varepsilon_{fs}$	The energy consumption about the amplifier
a	The data aggregating rate. It is application specific, where we assume $a = 3$ as mentioned in Section 2.
d_{toBS}	The average distance between the CHs and the BS. We assume it = 75 m
d_{toCH}	The average distance between the CHs and the non-CH nodes. From [5], we obtain $d_{toCH} = \sqrt{(1/2\pi)(M^2/k)}$, in which, M denotes the area of this field.

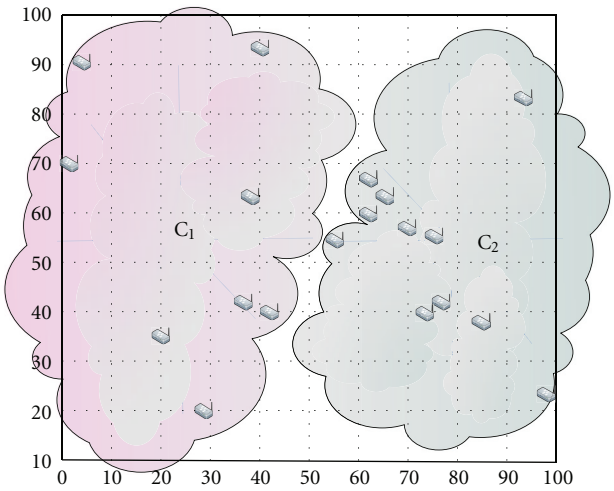

Figure 2: The clustering structure for 20 sensor nodes.

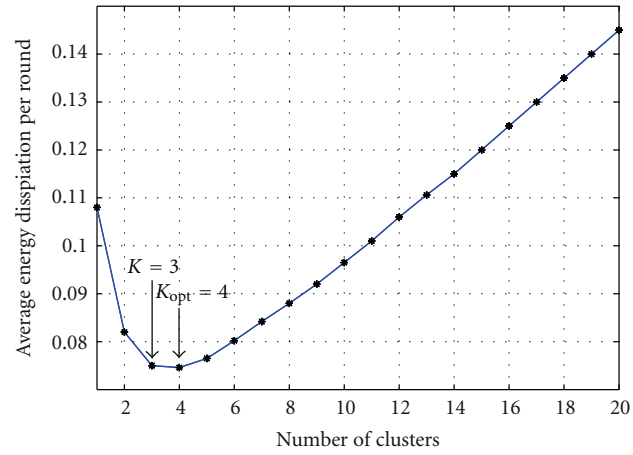

Figure 3: Average energy dissipation per round with varying the number of clusters.

cluster and change the clusters unequal distribution in the space of nodes originally.

Example 4. Figure 4 gives an example to illustrate how the cluster-balanced step works. In this example, if we only consider the geographical information of sensor nodes while using PCA-guided splitting algorithm, the sensor nodes s_1 to s_1 should belong to the C_1. However, the nodes in C_1 are more than Ave(=5). Thus, this scenario motivates the cluster-balanced step. To have a balance among all the clusters, in our method, we suggest that the sensor s_6 should be grouped into C_2 because its distance is nearest to C_2. We obtain the final balanced cluster structure by a serial standard PCA-guided splitting stage and the cluster balancing stage.

4.4. Cluster Heads Selecting Stage. After the base station divides the appropriate clusters using PCA technique, it needs to select the optimal cluster heads in these clusters. Assume that the initial energy is same, the base station can rank the matrix Y for each cluster and section the sensor nodes which is the nearest to the cluster centers.

4.5. Data Aggregating Stage. The sensor nodes begin to transfer the data to the cluster heads after finishing the cluster formation. The cluster heads collect the measurements from the sensor nodes and then compute the first principal component w_1 (as (18)), the mean vector G (as (21)), and Z (as (20)). They can be delivered to the base station with a constant packet size for each cluster head. At last, the base station will compute the approximate measurements by these packets.

4.6. The Description for PCA-Guided Routing Algorithm. The procedure taken by the base station is as follows.

Step 1: compute the first principal component v_1,

Step 2: according to (15), the two clusters C_1 and C_2 are determined via v_1.

Step 3: compute E_{round} and get the appropriate upper limit of the cluster nodes.

Step 4: while the cluster node number are more than the appropriate upper limit.

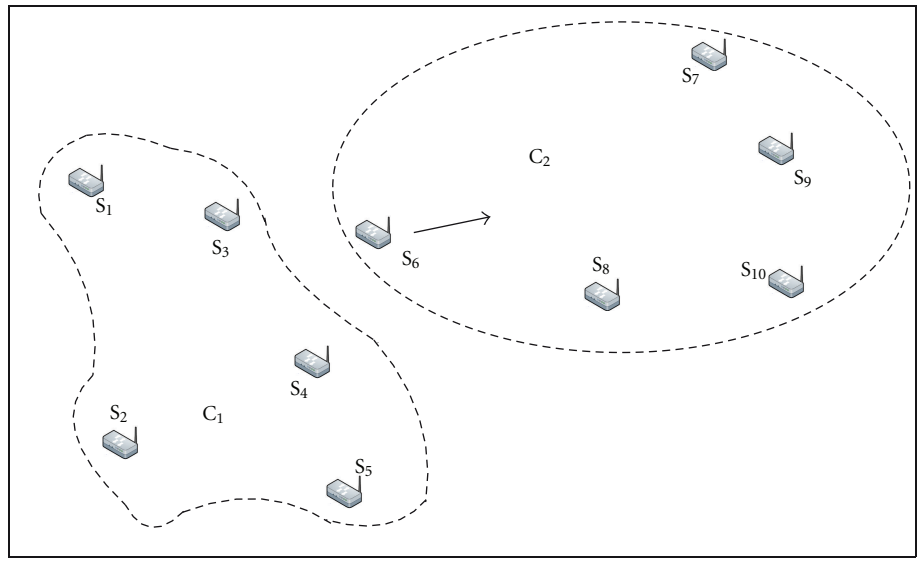

Figure 4: An example of the cluster balancing stage.

Step 5: compute a new v_1 for the cluster.

Step 6: repeat Step 2 and Step 3.

Step 7: end

Step 8: if needed, implement the cluster balancing stage.

Step 9: select the cluster heads.

Step 10: compute w_1 (according to (18)), G (according to (21)), and Z (according to (20)) for the approximate measurements.

5. Simulation Results

To evaluate the performance of PCA-RA, we simulate it, LEACH and LEACH-E, using a random 100-node network. The BS is located at $(50, 150)$ in a $100 \times 100\,\mathrm{m}^2$ field.

Figures 5 and 6 show the clustering structure for using LEACH and PCA-RA. Comparing Figures 5 and 6, one finds that each cluster is as compact as possible, and the cluster heads locate more closely to the cluster centers by using PCA-RA. This gives us an intuition that it is more efficient to balance the load of network and to even distribute the nodes among clusters by using PCA-RA.

The benefits of using PCA-RA are further demonstrated in Figures 7 and 8, where we compare the network performance of network lifetime and throughput under the PCA-RA with that under LEACH and LEACH-E. Figure 7 shows the total number of nodes that remain alive over the simulation time, while the first dead node remains alive for a more long time in PCA-RA, this is because PCA-RA takes into account the structure of clusters and the location of the cluster heads. Figure 8 shows that PCA-RA sends much more data in the simulation time than LEACH and LEACH-E.

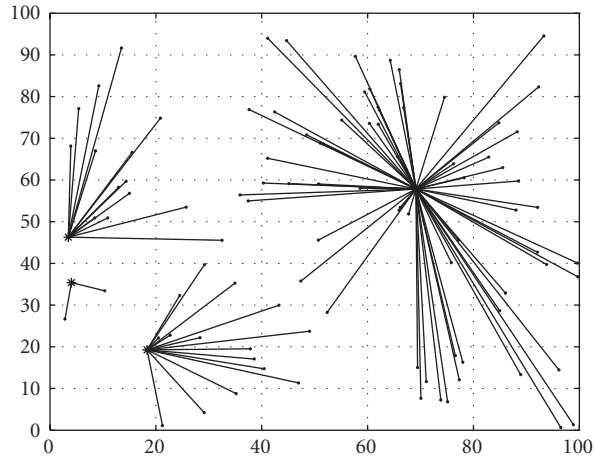

Figure 5: The clustering structure for using LEACH.

Note that the survey nodes in each round are NS_i; each node can send D_i data. Then, the maximum throughput can be expressed as follows;

$$\text{Throughput}_{\text{PCA-RA}} = \sum_{i=1}^{T_{\max}} NS_i \cdot D_i. \tag{29}$$

From Figure 7, the $T_{\max}$ for PCA-RA are more than the $T_{\max}$ for LEACH and LEACH-E. Then, $\text{Throughput}_{\text{PCA-RA}} > \text{Throughput}_{\text{LEACH}}$ and $\text{Throughput}_{\text{PCA-RA}} > \text{Throughput}_{\text{LEACH-E}}$.

For example, under the given test data, there are $60.456 * 10^3$ bits data sent in whole network lifetime with PCA-RA. And there are $33.394 * 10^3$ bits and $28.235 * 10^3$ bits by using LEACH-E and LEACH, respectively. The mathematics demonstrates that PCA-RA has 80.01% increase about

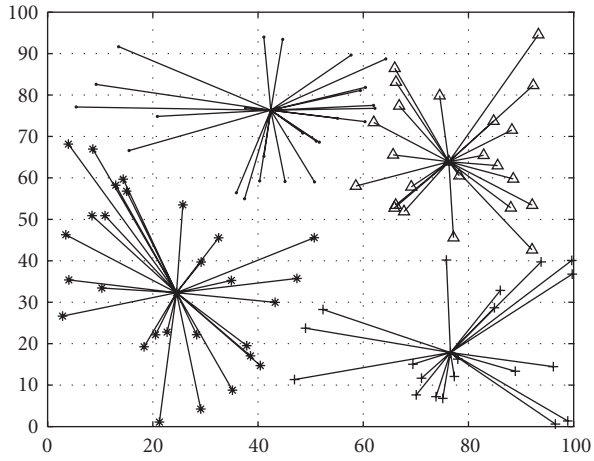

Figure 6: The clustering structure for using PCA-RA.

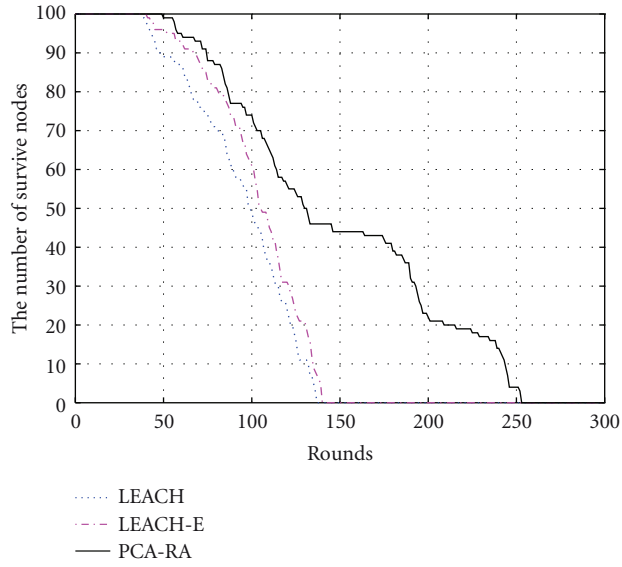

Figure 7: System lifetime using LEACH, LEACH-E and PCA-RA.

throughput compared with LEACH-E and 114.12% increase about throughput compared with LEACH.

PCA-RA is a centralized algorithm, and the complexity and communication cost of PCA-RA mostly happen in BS. About the balance structure stage, we think the effect on time complexity is small and consider that the time complexity is comparable to LEACH-C.

Table 2 displays the network lifetime (in terms of the time that the first node becoming dead) and the resulted square error function of the senor node structure under K-means algorithms and PCA-RA.

From Table 2, we can find that K-means algorithm can get a minimum square error. Because of the cluster-balanced step, PCA-RA can get a bigger square error, but the sensor nodes can survive a longer time. This implies that one can reach a certain tradeoff between the total spatial distance of sensor structure and the network lifetime. The suboptimal solution in PCA-RA can achieve such tradeoff.

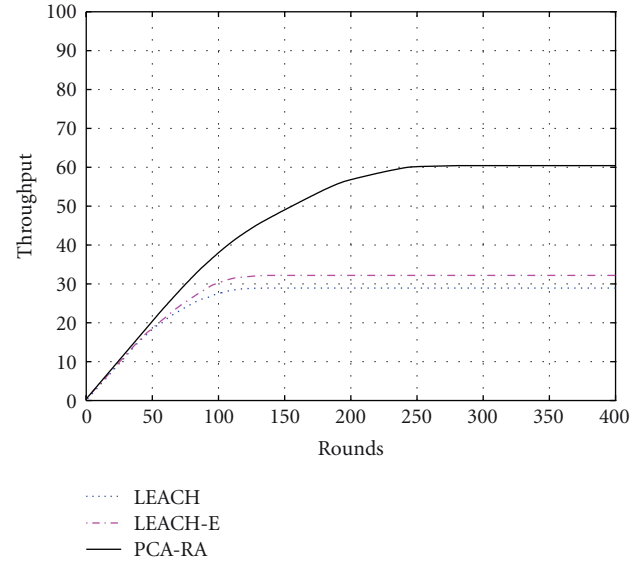

Figure 8: Throughput using LEACH, LEACH-E, and PCA-RA.

Table 2: A numerical example.

Parameters	K-means	PCA-RA
Square error	34392.980	38405.656
The first node dead time	88 rounds	99 rounds

In Figure 9, we simulate the sensor nodes collecting the measurements about temperature in some regions. Assume that the base station receive the packets from the cluster head in two minutes interval. Figure 9 demonstrates the approximations obtained by the base station about a certain sensor node in some intervals.

6. Conclusions

In this paper, we propose the PCA-guided routing algorithm for WSNs. By disclosing the connection between PCA and K-means, we design a clustering algorithm by utilizing PCA technique which efficiently develops a clustering structure in WSNs. Moreover, as a compression method, we demonstrate that the PCA technique can be used in data aggregation for WSNs as well. We establish the explicit procedure of PCA-guided routing algorithm for WSNs by incorporating PCA technique into both the data aggregating and routing process. The advantages of the proposed algorithm are demonstrated through both theoretical analyses and simulation results. The simulation results show that the PCA-guided routing algorithm significantly reduces the energy consumption, prolongs the lifetime of network, and improves network throughput when compared with LEACH and LEACH-E. Further, it keeps the accuracy about the measurements while reducing the network load.

Future research will focus on the distributed strategies of PCA-guided data aggregation and will investigate the performance of PCA-RA with different values of parameter K.

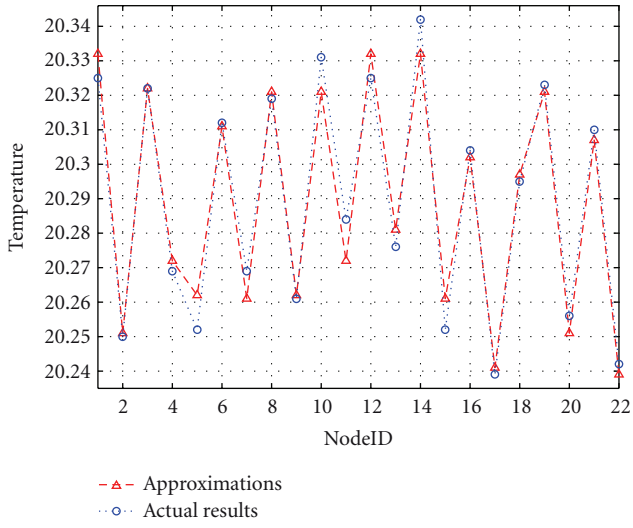

Figure 9: Approximation obtained for the sensor nodes.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China under Grant no. 61070197 and by self-determined research funds of CCNU from the colleges' basic research and operation of MOE under Grant CCNU11A01017.

References

[1] I. F. Akyildiz, W. L. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–28, 2004.

[3] Y. Tang, M. T. Zhou, and X. Zhang, "Overview of routing protocols in wireless sensor networks," *Journal of Software*, vol. 17, no. 3, pp. 410–421, 2006.

[4] W. Heinzelman, A. Chanrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, vol. 2, pp. 1–10, January 2000.

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

[6] S. Lindsey and C. Raghavendra, "PEGASIS: power-efficient gathering in sensor information systems," in *Proceedings of the IEEE Aerospace Conference*, vol. 3, pp. 1125–1130, 2002.

[7] A. Manjeshwar and D. P. Agarwal, "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks," in *Proceedings of the 15th International Conference on Parallel and Distributed Processing Symposium*, pp. 2009–2015, April 2002.

[8] A. Manjeshwar and D. P. Agarwal, "APTEEN: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks," in *Proceedings of the 15th International Conference on Parallel and Distributed Processing Symposium*, pp. 195–202, 2002.

[9] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[10] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[11] M. Tubaishat and S. Madria, "Sensor networks: an overview," *IEEE Potentials*, vol. 22, no. 2, pp. 20–23, 2003.

[12] O. Zytoune, M. El Aroussi, and D. Aboutajdine, "A uniform balancing energy routing protocol for wireless sensor networks," *Wireless Personal Communications*, vol. 55, no. 2, pp. 147–161, 2010.

[13] L. S. Tan, Y. L. Gong, and G. Chen, "A balanced parallel clustering protocol for wireless sensor networks using K-means techniques," in *Proceedings of the IEEE 2nd International Conference on Sensor Technologies and Applications*, pp. 300–305, Cap Esterel, France, August 2008.

[14] Y. L. Gong, G. Chen, and L. S. Tan, "A balanced serial K-means based clustering protocol for wireless sensor networks," in *Proceedings of the 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computing*, Dalian, China, October 2008.

[15] A. Abdulla, H. Nishiyama, and N. Kato, "Extending the lifetime of wireless sensor networks: a hybrid routing algorithm," *Computer Communications*, vol. 35, no. 9, pp. 1056–1063, 2012.

[16] A. F. Liu, J. Ren, X. Li, and Z. C. Xuemin, "Design principles and improvement of cost function based energy aware routing algorithms for wireless sensor networks," *Computer Networks*, vol. 56, no. 7, pp. 1951–1967, 2012.

[17] L. Cobo, A. Quintero, and S. Pierre, "Ant-based routing for wireless multimedia sensor networks using multiple QoS metrics," *Computer Networks*, vol. 54, no. 17, pp. 2991–3010, 2010.

[18] S. He, J. Chen, Y. Sun, D. K. Y. Yau, and N. K. Yip, "On optimal information capture by energy-constrained mobile sensors," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2472–2484, 2010.

[19] X. Cao, J. Chen, C. Gao, and Y. Sun, "An optimal control method for applications using wireless sensor/actuator networks," *Computers and Electrical Engineering*, vol. 35, no. 5, pp. 748–756, 2009.

[20] J. Shlens, "A tutorial on principal component analysis," December 2005, http://www.snl.salk.edu/~shlens/pca.pdf.

[21] D. P. Berrar, W. Dubitzky, and M. Granzow, in *A Practical Approach to Microarray Data Analysis*, pp. 91–109, Norwell, Mass, USA, 2003.

[22] G. Golub and C. Van Loan, *Matrix Computation*, Baltimore, Md, USA, 3rd edition, 1996.

[23] H. Zha, C. Ding, M. Gu, X. F. He, and H. Simon, "Spectral relaxation for K-means clustering," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '01)*, vol. 14, pp. 1057–1064, 2002.

[24] C. Ding and X. F. He, "K-means clustering via principal component analysis," in *Proceedings of the International Conference on Machine Learning (ICML'04)*, p. 29, Alberta, Canada, July 2004.

[25] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[26] J. Ham and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, 2nd edition, 2006.