Signal Processing 133 (2017) 156–168

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Data-driven sensors clustering and filtering for communication efficient field reconstruction☆

Jia Chen, Akshay Malhotra, Ioannis D. Schizas*

University of Texas at Arlington, Department of Electrical Engineering, 416 Yates Street, Arlington, TX 76010, USA

ARTICLE INFO

Keywords:
Field reconstruction
Norm-one regularized canonical correlation analysis
Clustering
Adaptive filtering
Communication efficiency

ABSTRACT

A novel communication efficient scheme for reconstructing a field sensed by spatially scattered sensors is proposed. The field is formed by multiple sources, while a fusion center gathers the sensor measurements. The goal is to reconstruct the field at the fusion center using only the measurements of a small number of sensors. The framework entails learning the correlation structure of the field by determining clusters of correlated sensors observing the same set of sources. Combining moving-average filtering along with principal component analysis applied in the sensor data covariance the number of sources can be determined, while norm-one regularized canonical correlations are utilized to determine the different correlated clusters. A novel iterative interplay of regularized canonical correlations with principal component analysis is designed that determines correctly the correlated clusters as the number of training data goes to infinity. From each cluster only a head sensor transmits data to the fusion center, while the measurements of the remaining sensors are reconstructed using proper linear filters that learn the correlation pattern within a cluster via normalized least mean squares. The novel approach substantially reduces the communication cost. Extensive numerical tests demonstrate the effectiveness of the proposed scheme in field recovery.

1. Introduction

The usage of large number of scattered sensors in environmental monitoring is gaining more and more popularity due to the low cost and simplicity of the sensing units. Sensors' main task is to monitor a field and continuously acquire spatio-temporal samples that contain information about the physical phenomena of interest. Such phenomena are induced by sources, such as thermal or chemical sources, causing temperature variations in a field [35], air pollution, or light intensity variations [11]. One popular architecture when using sensing units is the fusion-center (FC) based sensor network, where the sensing units transmit their measurements to the FC. Traditionally, all sensors transmit their data to the FC which usually causes large communication costs and results in resource depletion across all sensors [17,5]. The main focus of this work is to design a scheme where only a small number of sensors communicate their data to the FC, while the FC will reconstruct the field measurements at all other sensors using only the information acquired from properly selected sensing units.

The main task of this paper therefore is to derive an efficient centralized data fusion scheme to reconstruct the sensor measurements induced by multiple uncorrelated sources present in the field. Toward this end, a novel framework is proposed that carries out the following three tasks: (i) estimating the number of field sources; (ii) clustering the sensors based on which sources they are observing; and (iii) utilizing only the measurements of a few cluster head sensors to reconstruct *all* sensors' measurements at the FC. To achieve the first task, moving-average (MA) filtering [10] is combined with principal component analysis (PCA) [3] to eliminate sensing noise variance and extract the number of principal components (PCs) in the sensor data covariance corresponding to the uncorrelated sources. To group sensors in clusters according to their source information, the main idea is to realize that sensor measurements containing information about the same sources are statistically correlated. To exploit such spatial correlations, a norm-one regularized canonical correlation analysis technique [4] will be used to extract correlated sensor measurements and group them in clusters. PCA and adaptive filtering will also be utilized to result in an improved clustering approach that groups sensors according to their information content. Normalized least mean squares is utilized to enable the FC to reconstruct each cluster's sensor measurements using only data from a few cluster head sensors that communicate their observations to the FC. Different from existing source-based clustering techniques [4,6] the proposed frame-

☆ This work is supported by the National Science Foundation under Grants CCF-1218079 and ECCS-1509780.
* Corresponding author.
E-mail addresses: jia.chen81@mavs.uta.edu (J. Chen), akshay.malhotra@mavs.uta.edu (A. Malhotra), schizas@uta.edu (I.D. Schizas).

http://dx.doi.org/10.1016/j.sigpro.2016.10.024
Received 1 July 2016; Received in revised form 15 October 2016; Accepted 31 October 2016
Available online 10 November 2016

work is able to perform accurate clustering even when sensors observe multiple sources. The main idea is to separate the network into clusters of similar information content and use within each cluster the data of only one sensor to reconstruct the remaining sensor measurements at the FC. Such an approach will reduce significantly the number of scalars transmitted from the sensors to the FC potentially extending the sensors' lifespan.

The paper is structured as follows. Section 3 introduces the system setting and problem statement. Section 4 outlines the basic steps involved in the novel algorithmic framework. Specifically, MA filtering is combined with PCA to determine the number of sources (Section 4.1). Clustering of the sensors along with learning/reconstructing all sensor measurements is delineated in Sections 4.2, 4.2.2 and 4.4, respectively. In Section 4.3 it is demonstrated that for sufficiently large number of measurements, the proposed scheme achieves flawless sensor clustering, while a simple technique is put forth to improve clustering performance for a small number of training data. Section 5 provides communication and computational costs at the FC. Extensive numerical tests are carried out in Section 6, demonstrating the potential of the proposed framework.

2. Related work

Typically, sensors are limited in terms of communication and computational capabilities. A straightforward approach to collect the information across all sensors to the FC is to allow each sensor to forward its acquired measurements, possibly via multi-hop communications, to the FC. Such a process can place a heavy operational burden in all sensors due to the high communication cost. To tackle such a challenge, data aggregation techniques have become crucial to prolong the overall lifespan of a sensor network. There are four different strategies for data aggregation: (i) centralized approaches, (ii) in-network aggregation, (iii) tree-based approaches, and (iv) cluster-based aggregation [20]. The work in [1] considers a dynamic spanning-tree approach to minimize the energy consumption by taking into consideration the data traffic load. Support vector machines are used in [21] to reduce the redundant data and eliminate false data. An efficient cluster-based data aggregation scheme for heterogeneous sensor networks was developed in [15], where inter- (intra-) cluster data aggregation is performed to eliminate redundant data. The cluster-based approach in [25] uses a context-aware approach to validate data, while intra-cluster and inter-cluster redundancy is eliminated when sensors belong to the same cluster or neighboring clusters, respectively, for the validated data. The work in [28] focuses on the issues of accuracy, traffic load, redundancy elimination and delay when performing data aggregation, and proposed a model to address the aforementioned issues.

In the aforementioned cluster-based approaches, the geographical area is divided into multiple grid-based clusters, while the cluster heads are elected as those sensors with the highest energy and largest number of one-hop neighboring sensors, or as the sensors whose positions are closer to the centroid position of the cluster. Further, there is no basic principle in deciding the number of clusters needed. There are fundamental differences with the work proposed here. Specifically, the sensors will be clustered in groups according to their information content and the sources they sense (possibly multiple). Thus, the clusters here are formed based on the sensor information content and not according to an ad hoc splitting of the area monitored. Further, the number of clusters will be determined by the number of underlying information sources in the monitored field. Last but not least, the clustering proposed here is done to facilitate a form of reduction in the number of data transmitted and achieve accurate reconstruction at the FC.

Clustering sensors based on their source information has been considered in [26,4]. In [26], a distributed sparsity-aware matrix decomposition framework to estimate the support of the sparse covariance factors was proposed, which was further used in identifying source-informative sensors. A more generalized scheme for grouping sensors according to their source content was put forth in [4], which was able to deal with nonlinear data models. One challenge in the aforementioned approaches is that the number of sources is assumed to be known, while the source-to-sensor propagation channels used consider only flat fading. Another contribution of our work is the consideration of sensors acquiring information about multiple sources, which was not addressed effectively in [26,4].

There are various algorithms addressing the problem of clustering data into different groups in which the data share similar properties. The most popular technique is the K-means algorithm [13,7], which represents clusters by centroid points and allocates each data vector to the cluster that has the most similar centroid with respect to a distance metric. These clustering schemes are challenged by the fact that in our setting, sensors belonging to the same cluster do not necessarily exhibit similar magnitude in their observations, while sensors may be observing multiple sources resulting in overlapping clusters.

The field reconstruction problem has been previously addressed in the literature in [2] under the assumption that the monitored field is spatially governed by known partial differential diffusion equations. An interesting work can also be found in [23], where an optimal dimensionality-reduced approximation method was developed to recover thermal maps. Based on Bayesian estimation and Kalman filtering, the papers [18,31,33,30] consider statistical estimation methods for non-static fields. Algorithms to estimate a single source's parameters are studied in [16,9,12] to fully recover the monitored field. The work in [19] puts forth an algorithm that can successfully reconstruct sensor signals as long as the field bandwidth is sufficiently small, the field adheres to an one-dimensional model and the sensor locations are known to the FC. A distributed cluster-based signal reconstruction for non-bandlimited fields is proposed in [24], by locally adapting the field model. In [29], a distributed adaptive node-specific signal estimation (DANSE) algorithm is considered in wireless sensor networks to estimate a set of desired signals, however, the effects of nonideal propagation channels are not considered. However, the aforementioned schemes either assume that the fields are driven by assumed known spatiotemporal diffusion equations or statistical models, or only take a single-source into consideration. Different from the existing methods, our method is attractive since it does not require all sensor to communicate with the FC, it does not require knowledge of the sensors' positions, it can address settings where the field consists of multiple spatially scattered sources and the sources-to-sensors propagation channels may be multipath. There is no need to estimate the source signals, and our focus is on reconstructing the field only in points of interest where sensors are deployed rather than the entire field.

3. Problem formulation

3.1. System model

Consider p spatially scattered sensors. The sensors monitor a field which is generated by M underlying sources, while the number of sources M is not available. The sources are placed in different spatial locations, e.g., thermal sources, while the source signals are modeled as random uncorrelated processes, namely $s_m(t)$, where m is the source index and t denotes time. The source signals are assumed to be wide sense stationary, which implies that their ensemble average is time-invariant.

The field sources' signals are reaching the sensing units via multipath propagation channels. The channel coefficients from the mth source to sensor j is modeled as a finite impulse response filter with coefficients $\mathbf{h}_{j,m} = [h_{j,m}(0), \ldots, h_{j,m}(L-1)]$, where L corresponds to the maximum number of taps these filters can have. The channel coefficients are not known and modeled as random Gaussian variables. The