Reconstruction of Big Sensor Data

Yongshuai Shao, Zhe Chen, Fangfang Li, Chong Fu
School of Computer Science and Engineering
Northeastern University
Shenyang, China
Shao115_shuai@163.com
chenzhe@mail.neu.edu.cn

Abstract—With ubiquitous sensors continuously monitoring and collecting large amounts of information, there is no doubt that this is an era of big data. One of the important sources for scientific big data are the datasets collected by Internet of things (IoT). For an IoT application to analyze big sensor data, it is necessary that the data are clean and lossless. However, data loss in IoT is common due to unreliable wireless link or hardware failure in the nodes. To reconstruct the big sensor data, this paper first presents an algorithm based on matrix completion method. Then for multi-attributes sensor data, a tensor-based method is provided to estimate missing values. Moreover, an effective solution is proposed using the alternating direction method of multipliers. Finally, the experiments with real-word sensor data show the effectiveness of the proposed methods.

Keywords—big sensor data; data loss; matrix completion; tensor

I. INTRODUCTION

With ubiquitous sensors continuously monitoring and collecting large amounts of information, there is no doubt this is an era of big data. One of the important sources for scientific big data are the datasets collected by Internet of things (IoT) [1]. Learning from these large amounts of data is expected to bring significant science advances and improvements in quality of life. However, due to unreliable wireless link or hardware failure in the nodes, big data from sensors are often subject to corruption and losses. For an IoT application to analyze big sensor data, it is necessary that the data received are clean and lossless.

To address this problem, some related work have been done. A general method for reconstruction of missing data is suggested in [2], which exploits both temporal and spatial redundancy to characterize the phenomenon being monitored and distributed system. Recently, the intrinsic low-rank property of the high-dimensional data matrix has been considered. In paper [3-6], matrix completion theory is used to recover the missing data in sink node for large-scale wireless sensor networks. However, their main work focuses on energy saving from a sample of entries that are selected uniformly and randomly. In paper [7-9], compressive sensing is applied to the reconstruction of sensor data. A novel approach based on compressive sensing to reconstruct massive missing sensor data is proposed in [7]. By analyzing the real sensor data, the features of spatial correlation, temporal stability, and low-rank structure are exhibited. A multiple attributes-based recovery algorithm is proposed in paper [8]. The algorithm combines the benefits of compressive sensing and the correlation of attributes. In paper [9], an algorithm combining the benefits of compressive sensing, environmental space-time, and multi-attribute correlation features is proposed.

Based on the existing related works introduced above, this paper investigates the methods to reconstruct the big sensor data. Firstly, matrix completion technique is used to recover the missing sensor data. To solve the matrix completion optimal problem, an Alternating Direction method of multipliers (ADMM) based Matrix Completion (ADMC) algorithm is provided. Secondly, considering that sensors in IoT often monitor and measure multiple attributes data, a tensor-based method is proposed to reconstruct the multi-attributes sensor data. Thirdly, the proposed algorithms are simulated using real data.

Our contributions are summarized as follows:

- We use matrix completion technique to recover missing sensor data and provide an algorithm, named ADMC, which is based on ADMM model.
- To the best of our knowledge, this is the first work to apply tensor-based method to sensor data reconstruction problem.
- We design a tensor-based algorithm, named ADMC, to reconstruct multi-attributes sensor data.

The rest parts of the paper are organised as follows. In section II, we formulate the big sensor data reconstruction problem. Section III provides the algorithm for matrix sensor data reconstruction. Section IV proposes the algorithms for multi-attributes sensor data reconstruction. The performance is evaluated in Section V. Section VI concludes the paper and discusses future work.

II. PROBLEM FORMULATION

Suppose $n$ nodes are deployed in an area, each of which equips $k$ sensors to monitor attributes. Each sensor measures information during $t$ time slots.

Let $\mathbf{M}_i$ denote matrices of different attributes sensor data, $i=1,...,k$. Each $\mathbf{M}_i$ is a $n \times t$ matrix. Due to data loss in IoT, $\mathbf{M}_i$ is usually an incomplete matrix. The available information about $\mathbf{M}_i$ is a set of entries $(m_{ij})_{p,q}$, $(p,q)$ $\in \Omega_i$, where $\Omega_i$ is the set of sampled entries in $\mathbf{M}_i$. This
process is represented by using a sampling operator \( P_\Omega(\cdot) \), which is defined by:

\[
[P_\Omega(X)]_{ij} = \begin{cases} 
 x_{i,j}, & \text{if (i, j) } \in \Omega \\
 0, & \text{otherwise} 
\end{cases}, \quad (1)
\]

Therefore, the matrices we obtain are \( P_\Omega(M_i), i=1,\ldots,k \).

Our problem is to recover a series of raw data \( M \) from sampled incomplete matrices \( P_\Omega(M_i) \) as precisely as possible.

When \( k=1 \), only consider one attribute to reconstruct the missing data in sampled matrix. The problem is defined as follows:

Given subsets of \( M \) as \( P_\Omega(M) \), find an optimal solution as \( M' \),

\[
\begin{align*}
\text{minimize} & \quad \| M' - M \|_F \\
\text{subject to} & \quad P_\Omega(M') = P_\Omega(M)
\end{align*}, \quad (2)
\]

Where \( \| \cdot \|_F \) represents the Frobenius norm.

When \( k \geq 2 \), multi-attributes sensor data reconstruction problem is considered. Construct a three-order tensor using the sampled incomplete matrices. The three modes represent sensor time stamp \( I_t \), sensor ID \( I_{ID} \), and attributes \( I_a \) respectively: \( T \in R^{I_t \times I_{ID} \times I_a} \). The multi-attributes sensor data reconstruction problem is defined as follows:

Given subsets of \( T \) as \( P_\Omega(T) \), find an optimal solution as \( T' \),

\[
\begin{align*}
\text{minimize} & \quad \| T' - T \|_F \\
\text{subject to} & \quad P_\Omega(T') = P_\Omega(T)
\end{align*}, \quad (3)
\]

III. MATRIX SENSOR DATA RECONSTRUCTION

A. Formulate the single attribute sensor data problem

Let \( M \) denote received single attribute sensor data matrix, \( M \in R^{I_t \times I_{ID}} \). Due to low rank structure feature, the missing values in \( M \) can be recovered through rank minimization.

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(X) \\
\text{subject to} & \quad x_{i,j} = m_{i,j}, (i, j) \in \Omega
\end{align*}, \quad (4)
\]

where the elements of \( M \) in the set \( \Omega \) are given while the remaining elements are missing. \( \text{rank}(X) \) denotes the rank of matrix \( X \). For simplicity, the constraint condition can be summarized using \( P_\Omega(X) = P_\Omega(M) \). However, the program (4) is a NP-hard problem, hence it cannot be easily used in practice. A widely used alternative is the convex relaxation:

\[
\begin{align*}
\text{minimize} & \quad \| X \|_*, \quad (5)
\text{subject to} & \quad P_\Omega(X) = P_\Omega(M)
\end{align*}
\]

where \( \| X \|_* \) is the nuclear norm of matrix \( X \), that is, the sum of singular values of \( X \). The nuclear norm minimization problem (5) is the general model of matrix completion [10].

For convenience, define sampling matrix \( B \):

\[
b_{i,j} = \begin{cases} 
 1, & \text{if (i, j) } \in \Omega \\
 0, & \text{otherwise}
\end{cases}, \quad (6)
\]

Obviously, \( B \) is a \( n \times t \) binary index matrix, which indicates if a data point in \( M \) is missing. We now define the matrix sensor data reconstruction problem.

**Definition 1** \( B \) is the sampling matrix and \( M \) is the incomplete sensor data matrix that is to be recovered. Then the missing values in \( M \) can be effectively estimated by solving the below convex optimization problem,

\[
\begin{align*}
\text{minimize} & \quad \| X \|_*, \\
\text{subject to} & \quad B \cdot X = B \cdot M
\end{align*}, \quad (7)
\]

Where \( (\cdot) \) denotes the element-wise production of matrix.

Problem (7) is a typical convex optimization problem. It can be transformed to a semidefinite programming problem and solved using interior point methods. In section IV, to solve problem (7) we use CVX, a package for specifying and solving convex programs [11] [12] as contrast experiment.

In this paper, a matrix rank minimization algorithm, named ADMC, is proposed based on ADMM model to solve problem (7).

The alternating direction method of multipliers is a convex optimization algorithm dating back to the early 1980's. It has attracted attention again recently due to the fact that it is efficient to tackle large scale problems [13].

B. Rephrasing the matrix sensor data reconstruction problem as an ADMM model

In order to apply the ADMM method to problem (7), we need to transform it into the ADMM form. We first rephrase (7) as below:

\[
\begin{align*}
\text{minimize} & \quad \| X \|_* + (1/2 \lambda) \| B \cdot X - B \cdot M \|_F^2, \\
\text{subject to} & \quad X = Z
\end{align*}, \quad (8)
\]

where parameter \( 0 < \lambda < 1 \), it controls the fit to equality constraint. Consider a continuation technique for decreasing values of \( \lambda \), problem (8) equals to problem (7).

By introducing a new variable \( Z \), problem (8) becomes:

\[
\begin{align*}
\text{minimize} & \quad \| X \|_* + (1/2 \lambda) \| B \cdot Z - B \cdot M \|_F^2 \\
\text{subject to} & \quad X - Z = 0
\end{align*}, \quad (9)
\]

The augmented Lagrangian of (9) is:
\[ \mathcal{L}_\rho(X, Z, Y) = \|X\|_s + (1/2\lambda) \|B \cdot Z - B \cdot M\|_F^2 + (\rho/2) \|X - Z + U\|_F^2 \] (10)

where \(U\) is the scaled dual variable [13].

(1) **Update step for the X-variable**

Before giving the update step for \(X\)-variable, the following definition and theorem are needed [14].

**Definition 2** Assume that the singular value decomposition of the matrix \(X\) is given by \(X = U \text{diag}(\sigma)V^T\), where \(\sigma\) is the singular vector of \(X\), \((\cdot)^T\) is the transpose operator, \(U\) and \(V\) are orthogonal matrices. For any \(\tau > 0\), the matrix shrinkage operator \(\mathcal{D}_\tau(\cdot)\) is defined as:

\[ \mathcal{D}_\tau(X) = U \Sigma_\tau V^T, \]

where \(\Sigma_\tau = \text{diag}(\max(\sigma - \tau, 0))\). (11)

**Theorem 1** For any \(\tau > 0\), \(X := \mathcal{D}_\tau(Y)\) is a closed form solution for the following optimization problem:

\[ \min_X f(X) = \tau \|X\|_s + (1/2) \|X - Y\|_F^2. \] (12)

According definition 2 and theorem 1, the update of \(X\) is obtained:

\[ X^{k+1} = \mathcal{D}_{\frac{1}{\rho}}(U^k - Z^k). \] (13)

(2) **Update step for the Z-variable**

Before giving the update step for \(Z\)-variable, a proposition is given.

**Proposition 1** Let \(Z_{\text{right}} = (1/\lambda)(B \cdot M) + \rho(X + U)\) \(Z_{\text{left}} = (1/\lambda)B + \rho I\). Then \(Z := Z_{\text{right}} / Z_{\text{left}}\) is the closed form solution for the following optimization problem

\[ \min_Z (1/2\lambda) \|B \cdot Z - B \cdot M\|_F^2 + (\rho/2) \|Z - X - U\|_F^2, \] (14)

where \((/\cdot)\) represents element-wise division of matrix and \(I\) is a matrix that all its entries are one.

**Proof:** Suppose \(Z^*\) is the optimal solution to (14), if and only if

\[ 0 = (1/\lambda)(B \cdot Z^* - B \cdot M) + \rho(Z^* - X - U) \]

which is equivalent to

\[ ((1/\lambda)B + \rho I) \cdot Z^* = (1/\lambda)(B \cdot M) + \rho(X + U) \]

Thus \(Z^* = Z_{\text{right}} / Z_{\text{left}}\).

According to proposition 1, the update for variable \(Z\) is obtained:

\[ Z^{k+1} = Z_{\text{right}} / Z_{\text{left}}. \] (15)

**C. ADMC algorithm**

**Algorithm 1** ADMC algorithm for matrix sensor data reconstruction

1: Given \(B, M, \lambda, \rho, c, \lambda^*\)
2: Initialize \(Z^0 = U^0 = 0, k = 0\)
3: for \(k = 0, 1...\) do
4: \(X^{k+1} = \mathcal{D}_{\frac{1}{\rho}}(U^k - Z^k)\)
5: Calculate: \(Z_{\text{right}} = (1/\lambda)(B \cdot M) + \rho(X^{k+1} + U^k)\)
6: \(Z_{\text{left}} = (1/\lambda)B + \rho I\)
7: \(Z^{k+1} = Z_{\text{right}} / Z_{\text{left}}\)
8: \(U^{k+1} = U^k + X^{k+1} - Z^{k+1}\)
9: \(\lambda^{k+1} = \max(c, \lambda^*, \lambda^*)\)
10: end for
11: return \(X^k\)

**IV. MULTI-ATTRIBUTES SENSOR DATA RECONSTRUCTION**

Tensor is the higher-order generalization of vector and matrix. It may better represent the practical data structure. For example, the sensor nodes in Internet of things can sense multiple attributes simultaneously. Using tensor-based model to represent the multiple attributes sensor data can take full advantage of the correlations between attributes. That may further improve the accuracy of data reconstruction. In this section, tensor based methods are proposed to solve multi-attributes sensor data reconstruction problem.

In this paper, tensors are denoted with calligraphic font [15]. An \(N\)-order tensor is defined as \(\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times ... \times I_N}\). The "\(\text{unfold}\)" operation along the \(k\)th mode on a tensor \(\mathbf{X}\) is defined as \(\text{unfold}_k(\mathbf{X}) := \mathbf{X}_{(k)} \in \mathbb{R}^{I_k \times (I_1 \times I_2 \times ... \times I_{k-1} \times I_{k+1} \times ... \times I_N)}\). The opposite operation "\(\text{fold}\)" is defined as \(\text{fold}_k(\mathbf{X}_{(k)}) := \mathbf{X}\).

**A. Formulate the multi-attributes sensor data reconstruction as a convex optimization problem**

Assume \(\mathbf{T}\) is the received \(n\)-order tensor sensor data. Due to low rank structure feature, the missing data in \(\mathbf{T}\) can be recovered by rank minimization just like matrix case. Generalize the matrix completion algorithm to higher order tensors by solving the following optimization problem [16]:

\[ \min_{\mathbf{X}} \|\mathbf{X}\|_s \]

subject to \(P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{T})\). (16)

Now, we define the muti-attributes sensor data reconstruction problem.
Definition 3 Let $B$ denote binary sampling tensor and $T$ denote the incomplete tensor of multi-attributes sensor data. Then the missing values in $T$ can be effectively estimated by solving the below convex optimization problem,

$$\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{N} \| X_{(i)} \|_1, \\
\text{subject to} & \quad B \cdot X = B \cdot T
\end{align*}$$

where $(\cdot)$ denotes the element-wise production of tensor.

In the following part of this section, an ADMM based multi-attributes sensor data reconstruction algorithm, named ADTMC, is proposed to solve the problem (17).

B. Rephrasing the multi-attributes sensor data reconstruction problem as an ADMM model

In order to apply the ADMM model to problem (17), it need to be transformed into ADMM form.

Relax problem (17) to unconstrained formulation:

$$\begin{align*}
\min_X & \quad \sum_{i=1}^{N} \| X_{(i)} \|_1 + (1/2 \lambda) \| B \cdot X - T \|_F^2, \\
\text{subject to} & \quad Y_i = X_i, \quad i = 1, \ldots, N
\end{align*}$$

where the parameter $\lambda$ allows a tunable tradeoff between rank minimization and accuracy fitness. By decreasing value of $\lambda$, (18) equals to problem (17).

Introduce $N$ new tensor-valued variables, $Y_1, \ldots, Y_N$. Let $Y_{i(\cdot)} = X_{(\cdot)}$, $i \in \{1, \ldots, N\}$. With these new variables, the problem (18) can be rephrased as follows:

$$\begin{align*}
\min_{X, Y} & \quad \sum_{i=1}^{N} \| Y_{i(\cdot)} \|_1 + (1/2 \lambda) \| B \cdot X - T \|_F^2, \\
\text{subject to} & \quad Y_i = X_i, \quad i = 1, \ldots, N
\end{align*}$$

The augmented Lagrangian of program (19) is:

$$\begin{align*}
\mathcal{L}_\rho(Y, X, U) = & \sum_{i=1}^{N} \| Y_{i(\cdot)} \|_1 + (1/2 \lambda) \| B \cdot X - T \|_F^2 + \\
& + (\rho/2) \sum_{i=1}^{N} \| Y_i - X + U_i \|_F^2
\end{align*}$$

where $U_i, i=1, \ldots, N$ is the scaled dual variable.

(1) Update step for the $Y_i$-variables

Variable $Y_i$ can be solved independently by the matrix shrinkage operator introduced in section III. So the update for $Y_i$ becomes:

$$Y_{i}^{k+1} = \text{fold}_i(J_D_{\rho}(X^{k} - U_i^{k})).$$

(2) Update step for the $X$ variable

Before giving the update step for $X$ variable, the following proposition is given.

Proposition 2 Let $X_{\text{right}} = (1/\lambda)(B \cdot T) + N \rho (\overline{Y} + \overline{U})$, $X_{\text{left}} = (1/\lambda)B + N \rho 1$, $\overline{Y} = (1/N) \sum_{i=1}^{N} Y_i$, $\overline{U} = (1/N) \sum_{i=1}^{N} U_i$. Then $X^* := X_{\text{right}}/X_{\text{left}}$ is the closed form solution to:

$$\begin{align*}
\min_{X} & \quad (1/2 \lambda) \| B \cdot X - T \|_F^2 + (\rho/2) \sum_{i=1}^{N} \| Y_i - X + U_i \|_F^2
\end{align*}$$

where $1$ is a tensor that all its entries is one.

Proof: Suppose $X^*$ is the optimal solution to (21), if and only if:

$$0 = (1/\lambda)(B \cdot X^* - T) + \rho (N \cdot X^* - \sum_{i=1}^{N} Y_i)$$

Substitute $\overline{Y}$ and $\overline{U}$ into the above equation and sort it:

$$((1/\lambda)B + N \rho 1) \cdot X^* = (1/\lambda)(B \cdot T) + N \rho (\overline{Y} + \overline{U})$$

Thus $X^* := X_{\text{right}}/X_{\text{left}}$.

According to Lagrangian function and proposition 2, the update for $X^{k+1}$ is:

$$X^{k+1} := X_{\text{right}}/X_{\text{left}}.$$

C. ADMAC algorithm

Algorithm 2 ADMAC algorithm

1: Given $B, T, \lambda, \rho, c_{\lambda}, \lambda^*$
2: Initialize $X^0 = U^0 = 0, i = 1, \ldots, N, k = 0$
3: for $k = 0,1, \ldots$ do
4: for $i = 1, \ldots, N$ do
5: $Y_i^{k+1} = \text{fold}_i(J_D_{\rho}(X^{k} - U_i^{k})).$
6: end for
7: $\overline{Y}^{k+1} = (1/N) \sum_{i=1}^{N} Y_i^{k+1}$, $\overline{U} = (1/N) \sum_{i=1}^{N} U_i^{k}$
8: $X_{\text{right}} = (1/\lambda)(B \cdot T) + N \rho (\overline{Y} + \overline{U})$
9: $X_{\text{left}} = (1/\lambda)B + N \rho 1$, $X^{k+1} := X_{\text{right}}/X_{\text{left}}$
10: for $i = 1, \ldots, N$ do
11: $U_i^{k+1} = U_i^{k} + Y_i^{k+1} - X^{k+1}$
12: end for
13: $\lambda^{k+1} = \max(c_{\lambda}, \lambda^k, \lambda^*)$
14: end for
15: return $X^{k}$
V. PERFORMANCE EVALUATION

In this section, the performance of proposed algorithms are evaluated using real datasets.

The original datasets are from data sensing lab [17]. The data in data sensing lab are gathered by around 50 sensor motes distributed at the O'Reilly Strata Conference venue in Santa Clara in February 2013. And the datasets contain two attributes: temperature and humidity, which share the same selecting entries.

A. Matrix sensor data reconstruction experiment

For the performance evaluation, in our experiment, we first get complete raw matrix sensor data \( X \) from data sensing lab including temperature and humidity data. Using the random binary sampling matrix \( B \), we get the incomplete sensor data \( M = B \cdot X \).

In ADMC algorithm, we choose \( \lambda = 1, \lambda^* = 10^{-8}, c=1/4 \). \( \rho \) is any positive number and it's value affects the speed of convergence of the ADMM algorithm mildly. In our experiment we choose \( \rho = 0.1 / \text{std}(y) \) [18], where the \( y \) is a vector including all the sampling values and \( \text{std}(y) \) is the standard deviation of the observed values \( y \).

In this experiment, the ADMC algorithm is compared with interior point method based CVX.

Fig.1 shows that for temperature our algorithm can effectively reconstruct the loss data using only 30\% random sampling ratio, with error ratio all less than 5\%. Besides, it can be seen that our proposed convex optimization based algorithm almost has the same reconstruction accuracy with CVX, which indicates our algorithm's effective performance. However, it can be seen that our algorithm run much faster than CVX. This is because that CVX uses interior point methods, which have very high computation complex, to solve convex optimization problems. Whereas ADMC belongs to first order methods.

B. Multi-attributes sensor data reconstruction experiment

In this section, we use the sensor data to constitute a three-order tensor, where the three modes represent sensor time stamp, sensor node ID, and sampling attributes (including temperature and humidity), respectively. In our experiment, the tensor is of size \( 20 \times 100 \times 2 \).

In this experiment, we first get complete raw tensor sensor data \( X \) from data sensing lab. Using the random binary sampling tensor \( B \), we then get incomplete sensor data \( T = B \cdot X \).

In ADMAC algorithm , the parameters are set to be the same with those in ADMC, i.e., \( \rho = 0.1 / \text{std}(y), \lambda = 1, c = 0.25, \lambda^* = 10^{-8} \). In this experiment, the ADMAC algorithm is compared with EM-based Tucker decomposition algorithm [19]. It can be justified by the fact that the original data is approximately low rank \( \text{rank-(2,2,2)} \) in the sense of Tucker decomposition. As a contrast, we use the correct rank \( \text{Tucker[2,2,2]} \) and higher rank \( \text{Tucker[5,5,2]} \) to do Tucker decomposition experiment.

Fig.2 shows that the proposed tensor data recovery algorithm performs as good as Tucker decomposition of correct rank-(2,2,2) when sampling ratio more than 30\%, at that point they both get error ratio less than 5\%. However, a slightly larger rank-(5,5,2) Tucker decomposition can get worse performance. This means, using Tucker decomp-
sition to reconstruct tensor missing data accurately should first get exact \( n \)-rank. However, this will be intractable in practical. Especially the tensor is incomplete.

VI. CONCLUSION

In this paper, in order to solve the missing data problem in Internet of things, we first present an ADMC algorithm based on matrix completion method. Then for the multi-attributes sensor data, we provide a tensor-based method to estimate missing values. An effective solution is proposed using ADMM method. The experiments illustrate that our algorithm can achieve an effective reconstruction with less than 5% error with 30% random sampling values using the real sensor data.

The proposed multi-attributes sensor data recovery method assumes that the tensor is low-rank in all modes simultaneously. This can be not always true in real application. So in the future work, how to solve the data missing problem when the original tensor is low-rank only in certain modes will be meaningful.

ACKNOWLEDGMENT

This work is supported by the Fundamental Research Funds for the Central Universities (N140404015, N140404013, and N150402004) and the National Natural Science Foundation of China (61271350).

REFERENCES


