 JIOT, vol.1, no.1, pp.1-7, 2019

Electrical Data Matrix Decomposition in Smart Grid

Qian Dang[1], Huafeng Zhang[1], Bo Zhao[2], Yanwen He[2], Shiming He[3, *] and Hye-Jin Kim[4]

Abstract: As the development of smart grid and energy internet, this leads to a significant increase in the amount of data transmitted in real time. Due to the mismatch with communication networks that were not designed to carry high-speed and real time data, data losses and data quality degradation may happen constantly. For this problem, according to the strong spatial and temporal correlation of electricity data which is generated by human's actions and feelings, we build a low-rank electricity data matrix where the row is time and the column is user. Inspired by matrix decomposition, we divide the low-rank electricity data matrix into the multiply of two small matrices and use the known data to approximate the low-rank electricity data matrix and recover the missed electrical data. Based on the real electricity data, we analyze the low-rankness of the electricity data matrix and perform the Matrix Decomposition-based method on the real data. The experimental results verify the efficiency and efficiency of the proposed scheme.

Keywords: Electrical data recovery, matrix decomposition, low-rankness, smart grid.

1 Introduction

As the development of smart grid and energy internet [Tsoukalas and Gao (2008)], the amount of transmitted data in real time significantly increase. Due to the mismatch with communication networks that were not designed to carry high-speed and real time data, data losses and data quality degradation may happen constantly.

For this problem, the most common data recovery methods [Tu, Lin, Wang et al. (2018); Meng, Rice, Wang et al. (2018)] are used, such as mean, regression, interpolation and deep learning [Zeng, Dai, Li et al. (2018); Xiang, Li, Hao et al. (2018)]. According to the strong spatial and temporal correlation of electricity data which is generated by human's actions and feelings, some work takes the weather information as aid to recover electrical data via collective matrix factorization [Han, Dang, Zhang, et al. (2018)]. However, the weather information is quietly different for different locations, which can only be used to recover the electrical data of one location.

[1] Information & Communication Corporation, State Grid Gansu Electric Power Company, Lanzhou, 730050, China.

[2] State Grid Gansu Electric Power Corporation, Lanzhou, 730050, China.

[3] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, China.

[4] Business Administration Research Institute, Sungshin W. University, 02844, Korea.

[*] Corresponding Author: Shiming He. Email: shuiqiao9999@163.com.

Inspired by Matrix Decomposition or Matrix factorization (MF) [Tikk (2008); Hoyer (2004)], we treat the electricity data as a low-rank matrix where the two dimensional are day and user. We divide the low-rank electricity data matrix into the multiply of two small matrices and use the known data to approximate the low-rank electricity data matrix and recover the missed electrical data. Based on the real electricity data, we perform the Matrix Decomposition-based method on the real data. The experimental results verify the efficiency and efficiency of the proposed scheme.

The remainder of this paper is organized as follows. Section 2 introduces the system model. Section 3 presents electrical data matrix factorization. Section 4 provides simulation results and analyses. In the end, we conclude this work in Section 5.

2 System model

Generally, the value of the smart meter is the the cumulative power consumption of user on each day. The minus of two consecutive values is the power consumption on one day. We take the power consumption on each day as the electricity data. The electricity data is generated by human's actions and feelings, which has a strong spatial and temporal correlation. At the same time the human's actions and feelings has periodicity. Therefore, we treat the electrical data as a matrix $\mathbf{X} \in \mathbb{R}^{H \times N}$. In the electrical matrix, there are N uses and H days. The electrical data matrix contains the data within a H times measurement for N users. An element x_{ij} represents the power consumption of user j on the ith day, as shown in Fig. 1.

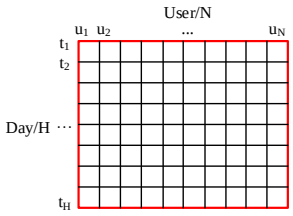

Figure 1: The matrix of electrical data (The row is time. The column is user.)

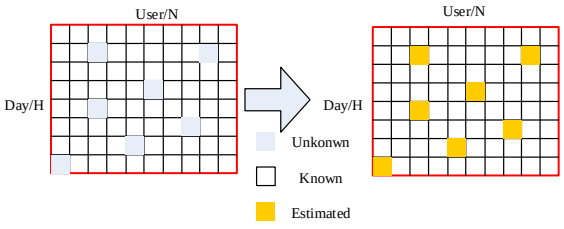

Figure 2: The recovery of electrical data

The electrical matrix has many lost elements. The subset Ω of matrix is the known set, where the elements x_{ij}, $(i, j) \in \Omega$ are known. As shown in Fig. 2, the recovery task is to estimate the lost or unknown element in the matrix by the spatial and temporal correlation and periodicity of the data in order to minimize the recovery error, which is usually

defined as squared error $(x_{ij} - \hat{x}_{ij})^2$, where x_{ij} is the real value and $\hat{x}_{ij}$ is the estimated value. The low-rankness of the electricity data matrix is analyzed [Han, Dang, Zhang et al. (2018)].

3 Electrical data Matrix Factorization

MF techniques approximate a low rank matrix $\mathbf{X}$ as a product of two much smaller matrices:

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}$ is an $H \times K$ and $\mathbf{V}$ is a $N \times K$ matrix.

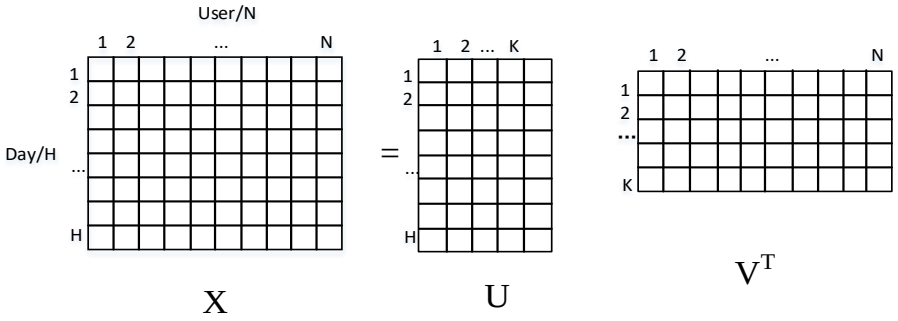

Figure 3: The concept of matrix factorization

$\mathbf{X}$ has many unknown elements which cannot be treated as zero. The subset of its entries $x_{ij}, (i, j) \in \Omega$ are known. The subset Ω can be formed with randomly selected entries of the matrix, and the sampling operator $P_{\Omega} : \mathbb{R}^{H \times N} \rightarrow \mathbb{R}^{H \times N}$ is defined by

$$[P_{\Omega}(\mathbf{X})]_{ij} = \begin{cases} x_{ij}, (i, j) \in \Omega \\ 0, otherwise \end{cases} \tag{2}$$

For this case, the approximation task can be defined as follows. Let $\mathbf{U} \in \mathbb{R}^{H*K}$ and $\mathbf{V} \in \mathbb{R}^{N*K}$. Let u_{ik} denote the elements of $\mathbf{U}$, and v_{jk} the elements of $\mathbf{V}$. Let $\mathbf{U}_{i*}$ denote a row of $\mathbf{U}$, and $\mathbf{V}_{j*}$ a row of $\mathbf{V}$. We can calculate the dot product of the two vectors corresponding to $\mathbf{U}_{i*}$ and $\mathbf{V}_{j*}$ as Equation 2 to get the estimation of x_{ij}.

$$\hat{x}_{ij} = \mathbf{U}_{i*}\mathbf{V}_{j*}^{\mathrm{T}} = \sum_{k=1}^{K} u_{ik}\mathrm{v}_{jk} \tag{3}$$

$\mathbf{X}$ can be recovered by solving the optimization problem.

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \frac{1}{2} \| P_{\Omega}(\mathbf{X} - \mathbf{U}\mathbf{V}^{\mathrm{T}}) \|_{\mathrm{F}}^2 \\ &= \min_{\mathbf{U},\mathbf{V}} \frac{1}{2} \sum_{(i,j)\in\Omega} e_{ij}^2 = \min_{\mathbf{U},\mathbf{V}} \frac{1}{2} \sum_{(i,j)\in\Omega} (x_{ij} - \hat{x}_{ij})^2 \end{aligned} \tag{4}$$

where e_{ij} denotes the training error on the (i, j)-th element. Problem (5) states that the optimal $\mathbf{U}$ and $\mathbf{V}$ minimizes the sum of squared errors only on the known elements of $\mathbf{X}$. we can use a simple incremental gradient descent method to find a local minimum, where one gradient step intend to decrease the square of prediction error. We compute the gradient of $\frac{1}{2}e_{ij}^2$:

$$\frac{\partial}{\partial u_{ik}}\frac{1}{2}e_{ij}^2 = -e_{ij}\cdot v_{jk} \quad , \quad \frac{\partial}{\partial v_{jk}}\frac{1}{2}e_{ij}^2 = -e_{ij}\cdot u_{ik} \tag{5}$$

Having obtained the gradient, we can now formulate the update rules for u_{ik} and v_{kj} as follows:

$$u_{ik}^{'} = u_{ik} + \eta\cdot e_{ij}\cdot v_{jk} \quad , \quad v_{jk}^{'} = v_{jk} + \eta\cdot e_{ij}\cdot u_{ik} \tag{6}$$

where η is a small value that determines the rate of approaching the minimum. To avoid over fitting, a regularized MF by penalizing the square of the Euclidean norm of weights is introduced.

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} L'(\mathbf{U},\mathbf{V}) \\ &= \min_{\mathbf{U},\mathbf{V}} \frac{1}{2}\| P_{\Omega}(\mathbf{X}-\mathbf{U}\mathbf{V}^{\mathrm{T}})\|_{\mathrm{F}}^2 + \frac{\lambda}{2}(\|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2) \end{aligned} \tag{7}$$

where $\|\bullet\|_{\mathrm{F}}$ represents Frobenius norm. The first two terms in the objective function are used to control the error in the matrix factorization process. The last item is the Euclidean paradigm of the factorized sub-matrix. The regularization penalty term prevents the matrix item from appearing negative values.

The objective function is not conjointly convex for all variables $\mathbf{U},\mathbf{V}$. We solve it by gradient descent. The partial derivative of the variable is used as a gradient.

$$\begin{aligned} \frac{\partial L'(\mathbf{U},\mathbf{V})}{\partial \mathbf{U}_{i*}} &= (\hat{x}_{ij} - x_{ij})\mathbf{V}_{j*}^{\mathrm{T}} + \lambda\mathbf{U}_{i*} \\ \frac{\partial L'(\mathbf{U},\mathbf{V})}{\partial \mathbf{V}_{j*}} &= (\hat{x}_{ij} - x_{ij})\mathbf{U}_{i*} + \lambda\mathbf{V}_{j*} \end{aligned} \tag{8}$$

Having obtained the gradient, we can now formulate the update rules as follows:

$$\begin{aligned} \mathbf{U}_{i*} &= \mathbf{U}_{i*} - \eta\frac{\partial L'(\mathbf{U},\mathbf{V},\mathbf{T})}{\partial \mathbf{U}_{i*}} = \mathbf{U}_{i*} - \eta((\hat{x}_{ij} - x_{ij})\mathbf{V}_{j*}^{\mathrm{T}} + \lambda\mathbf{U}_{i*}) \\ \mathbf{V}_{j*} &= \mathbf{V}_{j*} - \eta\frac{\partial L'(\mathbf{U},\mathbf{V},\mathbf{T})}{\partial \mathbf{V}_{j*}} = \mathbf{V}_{j*} - \eta((\hat{x}_{ij} - x_{ij})\mathbf{U}_{i*} + \lambda\mathbf{V}_{j*}) \end{aligned} \tag{9}$$

where η is a small value that determines the rate of approaching the minimum. All above, the stochastic gradient descent (SDG) Algorithm of MF for recovery is shown as Algorithm 1.

Algorithm 1 SDG Algorithm of MF for recovery

Input：$\mathbf{X}$，Error threshold ε

Output： $\mathbf{U},\mathbf{V}$

1.Random initialization $\mathbf{U} \in \mathbb{R}^{H*K}, \mathbf{V} \in \mathbb{R}^{N*K}$

2. η is the step，t is the number of iteration which is set to 1

3.While (t<M and L_t-L_{t+1}> ε)

4. for all $\omega_{ij} \neq 0$

$\hat{x}_{ij}=\mathbf{U}_{i*} \times \mathbf{V}_{j*}^{\mathrm{T}}$

$\mathbf{U}_{i*} = \mathbf{U}_{i*} - \eta\lambda\mathbf{U}_{i*} - \eta(\hat{x}_{ij} - x_{ij})\mathbf{V}_{j*}^{\mathrm{T}}$;

$\mathbf{V}_{j*} = \mathbf{V}_{j*} - \eta\lambda\mathbf{V}_{j*} - \eta(\hat{x}_{ij} - x_{ij})\mathbf{U}_{i*}$;

$L_{t+1} = \sum_{\omega_{ij}\neq 0} \| \hat{x}_{ij} - x_{ij} \|^2$

$t = t + 1$

5.Return $\mathbf{U},\mathbf{V},\mathbf{T}$

4 Simulation

The real electrical data comes from Lanzhou power system company with 160 users in Jiuquan of the Lanzhou province from August 1, 2016 to August 31, 2017. Except for the lost data, we can get the available real data of 160 users in 385 days which are all known. The real data is treated as a matrix $\mathbf{X} \in \mathbb{R}^{385\times160}$. There are 160 uses and 385 days.

The root mean squared error (RMSE) is used to evaluate the recovery accuracy, which is defined as:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{(i,j,k)\notin\overline{\Omega}} (x_{ijk} - \hat{x}_{ijk})^2}{|\overline{\Omega}|}} \tag{14}$$

where $\overline{\Omega}$ is set of the entries on which the values are unknown, $|\overline{\Omega}|$ is the number of unknown entries. If the RMSE is smaller, the recovery accuracy will be higher. We compare our scheme with the Average filling (AVG) recovery on different sample ratios. The sample ratio is the ratio of the number of known elements to the number of all elements in the electrical data. The higher the ratio, the more known elements, the more information we know, and the fewer elements we need to recover. We set the sampling ratios from 85% to 97.5%, increasing at 2.5% intervals.

Fig. 4 shows the RMSE of MF and Average filling (AVG) with different sample ratios. With the all sample ratio, the recovery accuracy of CP is better than that of MF. The reason is that MF uses more periodicity information. And as the sample ratio increases, the RMSE decreases because the more information is known, the more potential relationships will be provided to help improve recovery accuracy.

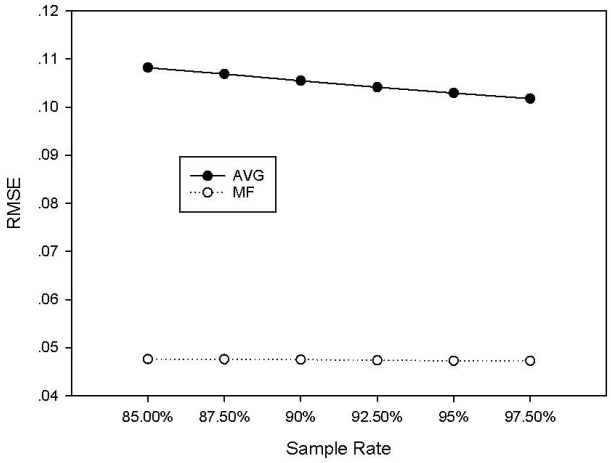

Figure 4: The RMSE with different sample ratios

5 Conclusion

According to the strong spatial, temporal correlation and periodicity of electricity data, we treat them as a low-rank matrix where the dimensional are day and user. We perform the matrix decomposition-based method on the real data. The experimental results on real data verify the recovery accuracy efficiency of the proposed scheme.

Acknowledgement: This work was supported by the Science and Technology Projects of State Grid Gansu Electric Power Corporation (52272315000X).

References

Han, X.; Dang, Q.; Zhang, H.; He, Y. W. (2018): Electrical data recovery with weather information via collective matrix factorization. *International Conference on Applications and Techniques in Cyber Intelligence*, pp. 1-10.

Hoyer, P. O. (2004): Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, no. 5, pp. 1457-1469.

Meng, R. H.; Rice, S. G.; Wang, J.; Sun, X. M. (2018): A fusion steganographic algorithm based on faster R-CNN. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1-16.

Tikk, D. (2008): Investigation of various matrix factorization methods for large recommender systems. *IEEE International Conference on Data Mining Workshops*, pp. 1-6.

Tsoukalas, L. H.; Gao, R. (2008): From smart grids to an energy internet: assumptions, architectures and requirements. *Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*, pp. 94-98.

Tu, Y.; Lin, Y.; Wang, J.; Kim, J. U. (2018): Semi-supervised learning with generative adversarial networks on digital signal modulation classification. *Computers, Materials &*

Continua, vol.55, no.2, pp. 243-254.

Xiang, L. Y.; Li, Y.; Hao, W.; Yang, P.; Shen, X. B. (2018): Reversible natural language watermarking using synonym substitution and arithmetic coding. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 541-559.

Zeng, D. J.; Dai, Y.; Li, F.; Sherratt, R. S.; Wang, J. (2018): Adversarial learning for distant supervised relation extraction. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 121-136.