Data Desensitization Method of Electricity Information
Shuai YANG¹, Kun-lun GAO², Ting ZHAO² and Jiang-qi CHEN²

¹School of Electrical and Electronic Engineering, 
North China Electric Power University, Beijing, China
²Advanced Computing and Big Data Technology Laboratory of SGCC 
Global Energy Interconnection Research Institute, Beijing, China

Keywords: Fourier transform, Data desensitization, Clustering analysis, Data dimension reduction.

Abstract. With the explosive growth of electricity information, its potential value and security problem is becoming more and more attractive. To ensure the safety and reliability of electricity data at different application levels, this paper proposes two kinds of data desensitization methods based on Fourier transform theory by studying the appropriate processing measurements, connotation and adjustable parameters. Specifically, recoverable and irreversible desensitization methods are proposed, which are used in the typical scene of electric data mining—load clustering. Experiments show that both types of methods can support clustering analysis effectively.

Introduction

With the further development of smart grid, higher degree of digitization, informatization and intelligence in power system are developing fast, and it stimulates the increase of the data sources. The current Electricity Information Collection System has already satisfied information collection from nearly 400 million meters. In the new situation of explosive growth in power system data, it is necessary to explore the data processing technology to satisfy the data asset management demand and large data analysis and mining requirements.

The value of electricity information is to explore relationship or law in data, meet the power enterprise demands about improving the quality, benefit and efficiency, promote the optimal allocation of power resources and efficient service. At present, the researches mainly include electricity users segment[1-2], bad data identification[3], short-term load forecast[4], demand management[5], etc. Because of the high value of electricity data, the problems of sensitive data sharing are of concern. Therefore, to achieve the balance of electricity data mining and privacy risk avoiding, data desensitization should be applied when storing, transmitting and sharing electricity data.

To meet the demands of information acquisition and the requirements of external data application, this paper proposed two kinds of data desensitization methods based on Fourier transform in a different perspective compared with traditional solution using in big data. The paper develops as follows. The concept and form of discrete Fourier transform are introduced and the feature extraction method based on time-frequency transform is studied; recoverable and irreversible data desensitization method are proposed based on different low-pass filters, meanwhile, the recovery principle in the former and data dimensional reduction characteristics in the latter are explained; in the typical application of electric data mining—cluster analysis, it is verified that two kinds of data desensitization methods can support model analysis effectively by determining the optimal number of clustering using selected appropriate clustering evaluation index.

Discrete Fourier Transform

Fourier Transform Theory

Fourier transform is a method of analyzing signals, which can analyze the components of the signal or synthesize the signals by using these components. Many waveforms can be used as components
of signals, such as sine wave, square wave, sawtooth wave, and so on, Fourier transform uses sine wave as the component of signal usually. Fourier transform is divided into Continuous Fourier transform (CFT) and Discrete Fourier transform (DFT). CFT is mainly used for theoretical research, while the real signal is usually processed by Fast Discrete Fourier Transform (FDFT) which is implemented by computer in practical application. The DFT formula for signal $x(n)$ with length $N$ is:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}, \quad k = 0, ..., N-1$$

(1)

The Inverse Discrete Fourier Transform (IDFT) formula is:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N}, \quad n = 1, ..., N-1$$

(2)

The sequence of transformations at both ends (time and frequency domains) is finite in form, in fact, both sets of these sequences should be considered as the main sequence of discrete periodic signals. When using DFT to process a finite discrete signal, it should be regarded as a periodic signal via periodic continuation. FDFT is usually used to compute DFT efficiently, and the computational time complexity can be reduced from $O(N^2)$ to $O(N\log N)$.

**Eigenvalues Extraction based on DFT**

One-dimensional signal $x(n)$ with length $N$ is transformed into spectrum $X(k)$ by DFT. It is a complex signal with length $N$ in form, but it should be considered as a period of infinite signal with on connotation, that means $X(N+k) = X(k)$. When $x(n)$ is a real signal, its spectrum $X(k)$ has the following characteristics: $X(0)$ is a real number; $X(N/2)$ is real number if $N$ is even number; $X(k)$ and $X(N-k)$ are conjugate complex numbers. These explain why $N$ number of real numbers which correspond to the real and imaginary parts of the Fourier coefficients are sufficient to reproduce the original $x(n)$ with length of $N$. Therefore, these real numbers can be considered as eigenvalues for further analysis.

The frequency range of spectrum $X(k)$ is arranged from $[0, N-1]$. To analyze and discuss frequency domain data more intuitively, section $[(N-1)/2, N-1]$ should be shifted to section $[-(N-1)/2, 0]$. The “spectrum” mentioned below is the frequency spectrum after this shift. The Low-frequency component retains the macroscopic characteristics of the data and can reflect the whole state of the data while the high frequency component retains the microscopic characteristics and can reflect the details of data. For most natural signals, Fourier spectra often have the characteristics of low amplitude of high frequency part and high amplitude of low frequency part. To preserve more macroscopic features after data processing, the principle of keeping low frequency components and removing high-frequency components is often adopted.

**Data Desensitization of Electricity Information**

**Outline of Desensitization**

Data desensitization is referring to the process of dealing with secret or privacy information, such as personal identification information and trade confidential data, to achieve data deformation or data fuzzification. It makes it impossible for a malicious attacker to obtain sensitive information from the desensitized data directly to protect the confidentiality and privacy. Generally, desensitized data should have most of the characteristics of the original data, as they will still be used for developing or testing occasions. According to different desensitization rules, data desensitization is mainly divided into the recoverable and the irreversible. The true value in data processed by the former only
can accessed by authorized administrators or users by a specific method, reducing the risk of sharing and moving important data.

Any real signal can be reproduced perfectly by corresponding DFT coefficients. Thus, we apply specific deformation to its Fourier coefficients for achieving data desensitization. The low-pass filter filters the high-frequency components of the signal and retains the low-frequency components.

Although the weakening extent of each frequency signal is different for different filters, all filters provide a smooth form of the signal by removing short-term fluctuations and keeping long-term trends. Since the low-pass filter retains the characteristics and trends of the signal, it meets the desensitization requirements of the electrical data. There are many low-pass filters, such as ideal low-pass, Gaussian low-pass and Butterworth low-pass filter. The frequency response curve of Butterworth Low-pass filter is the most smooth as shown in Figure 1. By selecting different filters to modify the Fourier coefficients of electricity data, we made controllable data desensitization reality. The specific implementation of recoverable and irreversible desensitization methods based on different filters are explained below.

![Image of filter types](image)

Recoverable Desensitization Method

By using low-pass filters to modify the Fourier coefficients, controlling adjustable parameters, controllable data desensitization can be realized. However, not all Fourier modifying methods are reversible. Because of the most smooth frequency response curve of Butterworth low-pass filter, the modified spectrum does not appear zero value. Furthermore, the inverse filter can be used to restore this modification so that desensitization data can be recovered.

The steps of recoverable data desensitization is the following:

1) data desensitization
   a) Using DFT to deal with power load curve data \( x(n) \) with length \( N \), the spectrum of the daily curve can be obtained, which is consist of \( N \) frequency coefficients:

   \[
   X(k) = \sum_{n=0}^{N-1} x(n)e^{-\frac{2\pi k n}{N}}, \quad k = 0,...,N-1
   \]  
   (3)

   b) Modifying spectrum by second order Butterworth low-pass filter with cut-off frequency \( \sigma \):

   \[
   \tilde{X}(k) = X(k) \cdot \left(1 + \frac{k^2}{\sigma^2}\right)^{-\alpha}, \quad k = 0,...,N-1
   \]  
   (4)

   c) The desensitized data can be obtained after transforming modified spectrum by IDFT:

   \[
   \tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k)e^{\frac{2\pi k n}{N}}, \quad n = 0,...,N-1
   \]  
   (5)

2) Desensitized data recovery
a) Using DFT to process desensitized data $\tilde{x}(n)$:

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n)e^{-j2\pi kn/N}, \quad k = 0, \ldots, N-1$$

(6)

b) Obtaining the cut-off frequency and type of low-pass filter through secure channels, modified spectrum is restored by inverse filtering method:

$$X(k) = \tilde{X}(k) \cdot (1 + \frac{k^2}{\sigma^2})^{-2}, \quad k = 0, \ldots, N-1$$

(7)

c) Transforming the restored spectrum by IDFT, the restored data can be obtained:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N}, \quad n = 1, \ldots, N-1$$

(8)

Applying method above to a selected power load curve with different cut-off frequency, the curve before and after the desensitization and the load curve before and after recovery as shown in the Figure 2.

![Figure 2. Load curve recoverable desensitization and recovery.](image)

We can see that the smaller $\sigma$ is, the higher degree of desensitization is. Meanwhile, it is by the proposed method that the desensitization data can be restored to the original signal perfectly after obtaining associated arguments through security channel. Since not involving data compression, we can select modified frequency coefficients as eigenvalues for desensitization access of electrical data, also we can select desensitized data directly for security access.

**Irreversible Desensitization Method**

If the selected low-pass filter makes the inverse filter failure, the data desensitization method is irreversible. One of the simplest and most effective way is to use the ideal low-pass filter for data desensitization. This filter is physically impossible, but it can cut off the signal completely higher than the cut-off frequency, while transmit the signal without distortion below the cut-off frequency. Because of this characteristics, the corresponding Fourier coefficient is set to zero. So that this deformation is irreversible, then achieving irreversible data desensitization.

The steps of irreversible data desensitization is the following:
a) Using DFT to deal with power load curve data \( x(n) \) with length \( N \), the spectrum of the daily curve can be obtained, which is consist of \( N \) frequency coefficients:

\[
X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, \quad k = 0, \ldots, N-1
\]  

(9)

b) Modifying spectrum by ideal low-pass filter with cut-off frequency \( \sigma \):

\[
\tilde{X}(k) = \begin{cases} 
X(k), & |k| \leq \sigma \\
0, & |k| > \sigma 
\end{cases}, \quad k = 0, \ldots, N-1
\]  

(10)

c) The desensitized data can be obtained after transforming modified spectrum by IDFT:

\[
\tilde{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}(k) e^{j2\pi kn/N}, \quad n = 0, \ldots, N-1
\]  

(11)

When the cutoff frequency is chosen to be \( \sigma \), it is equivalent to the fact that we retain the frequency coefficients as the eigenvalue from the 0-band radius \( \sigma \) as shown in Figure 3. One way to store, transfer, and share desensitized data is to access defined eigenvalues so as to achieve data compression or data dimensional reduction.

![Figure 3. Data dimensional reduction characteristic.](image)

**Study on Desensitization Effectiveness for Clustering Analysis**

**Load Clustering**

The electrical data desensitization is intended to support data analysis when secure data storage, transmission and sharing. Data analysis functions fall into two basic types: statistical analysis and classification analysis. For the former, its evaluation index is defined as the mean, standard deviation, discrete coefficient and other statistical error; for the latter, its evaluation index is defined as the accuracy of the classification model. The spectrum of desensitization data retains the zero frequency, which is equivalent to the sum of the time domain data, ensuring the mean value of non-negative electricity data unchanged, so there's not much room for discussion on statistical index. Since the load curve clustering belongs to the category of classification analysis, we will discuss the effectiveness of proposed data desensitization method for load clustering.

The load curve has certain regularity, so load clustering analysis is an effective means for the load model to be practical. The traditional clustering algorithms include hierarchical clustering, partitioned clustering, density-based clustering, grid-based clustering, model-based clustering algorithm, etc[7-8]. The k-means clustering algorithm is a common dynamic clustering algorithm. Its implementation process is as follow: Select the clustering center to classify samples initially, then judge whether the clustering is reasonable according to the clustering criteria, if not, modify the cluster until it is reasonable. Though k-means is sensitive to the initial value setting compared with other clustering
algorithms, the algorithm has the characteristics of simplified calculation and faster convergence speed. Therefore, we selected this algorithm for load clustering analysis.

**Clustering Evaluation Index**

For load clustering, how to evaluate the clustering effectiveness is the key. There are two interrelated evaluation criteria usually: 1) In clustering results, shorter intra-cluster distance and longer inter-cluster are better; 2) If clustering result is more consistent with the artificial judgment, it’s better. The former is an index based on the objective function, and the latter is based on the artificial judgement. People tend to evaluate the quality of a work based on adjacent degree between expectation and result. Relatively speaking, the index based on artificial judgment are more suitable for evaluating the clustering result. However, in fact, there are fewer data based on manual labeling especially the volume of electricity data is so large. So we summarized several index based on objective functions including Compactness Index(CPI), Separation Index(SPI), Davies-Bouldin Index(DBI), Dunn Validity Index(DVI). Because SPI and CPI only consider one dimension of the clustering effect and computing complexity of DVI is high, this paper chooses DBI as the clustering evaluation index to determine the optimal clustering number.

DBI takes into account the intra-class compactness and the inter-class separation:

\[
DBI = \frac{1}{k} \sum_{i=1}^{k} \left( \frac{CPI_i + CPI_{max}}{w_i - w_j} \right)
\]

where \(T_i\) is the data volume of the i-th class, \(k\) is the number of clusters, \(w_i\) is the i-th cluster center, \(x_j\) is the j-th data point of the i-th cluster. The lower DPI is, the better clustering effect.

**Data Desensitization Clustering Example**

**Original Data Clustering.** Select 2000, 96-point load curve data for experiment. If neglecting the machine error, desensitization data processed by proposed recoverable method can be restored without missing any information by obtaining associated parameters safely. It can be concluded that the clustering result of recoverable desensitization data is the same as the original data clustering result.

The daily load curve data is classified using k-means clustering algorithm. Observe the clustering effect and the operation time. The relationship between clustering number and DBI is shown in Figure 4. Clustering number \(k=5\) is the minimum point of the relation curve, that is, the clustering result is the best when \(k=5\). The relationship between the clustering number and operation time is illustrated in Figure 5. As the clustering number increases, the operation time is increasing. When \(k=5\), time \(t=0.0372s\).

![Figure 4. Clustering number and DBI.](image)

![Figure 5. Clustering number and operation time.](image)
The results of the original data clustering are shown in Figure 6. These load curves can be classified into 5 classes, the main types are single-peak type, peak averting type, low load rate type, high load rate type and slight fluctuation type. The peak value, peak period and base load of each class cluster are different.

**Recoverable Desensitization Data Clustering.** Experiment with the same data as above. The core of the recoverable desensitization is the use of a second-order Butterworth low-pass filter to modify the Fourier coefficients. Different cutoff frequency corresponds to different data desensitization degree. To measure desensitization degree, we defined the reconstructed error \( ERR \), which means the relative error between the desensitization data and the original data:

\[
ERR = \frac{\sum_{i=1}^{M} \| x_i - \tilde{x}_i \|}{\sum_{i=1}^{M} \| x_i \|}
\]

(14)

where \( M \) is the number of electricity data samples, \( x_i \) is the original value of the i-th sample, \( \tilde{x}_i \) is the desensitization value of the i-th sample.

Select different cut-off frequency, calculate the corresponding reconstruction error \( ERR \) as shown in Figure 7. We can see that the \( ERR \) decreases with the increasing of cut-off frequency \( \sigma \). When \( \sigma = 24 \), \( ERR \) does not exceed 5%. Because of the corresponding relationship between the load curve turning point and the actual switch off, it can be considered that the turning point has been blurred when \( ERR \) reaches 5%. To balance the desensitization effect and clustering effect, the cutoff frequency \( \sigma \) is selected as 24, and the relationship between clustering number and DBI is shown in Figure 8.

![Figure 6. Original data clustering results.](image)

![Figure 7. ERR and cut-off frequency \( \sigma \).](image)

![Figure 8. Clustering number and DBI.](image)

We can see clustering number \( k = 2 \) is the minimum point of the curve. When the \( k \) is 2, exaggerated SPI will make DBI become small, then the two clusters, further data mining is not available in the result with two clusters except for the difference of amplitude. Therefore, when clustering load curve, discarding the result of \( k = 2 \). In fact, \( k = 4 \) is the most reliable clustering number.

Because different cutoff frequencies have different smoothing effects on data, the DBI value may be changed to some extent. The results of the recoverable desensitization data clustering are shown in
Figure 9. The main types of load curves are single-peak type, peak averting type, stable type and slight fluctuation type. The stable type in this experiment is actually composed of high and low load rate types in the result of the original data clustering. Therefore, the results are reasonable, indicating that proposed recoverable data desensitization method can support the cluster analysis effectively.

Irreversible Desensitization Data Clustering. Experiment with the same data as above. According to the data dimensional reduction of the irreversible desensitization method, the cutoff frequency is chosen as \( \sigma \), then the \( 2\sigma + 1 \) number of real eigenvalue can be classified as the eigenvector of the load curve, so as to achieve desensitization. Select different cut-off frequency, calculate the corresponding ERR as shown in Figure 10. From the results, with the number of eigenvalues increasing, ERR is reduced, when the number of eigenvalues is 19, ERR does not exceed 10%. On the one hand, we can reduce the data storage space, on the other hand, we can reduce the calculation time of Euclidean distance between vectors and improve the efficiency of the algorithm. Therefore, the eigenvector is as short as possible.

To balance desensitization effect and clustering effect, ERR is limited to less than 10%, and the number of eigenvalues is at least 19. The desensitized load curves are classified with eigenvector. The relationship between clustering number and DBI as shown in Figure 11. Discarding the clustering result of \( k=2 \), and \( k=5 \) is the best clustering number, which is consistent with the original data clustering result. The relationship between the clustering number and operation time is illustrated in Figure 12. Its trend is the same as the original data clustering. When \( k=5 \), time \( t=0.0111s \). The computing time is 70.16% shorter than original data clustering.
The results of the irreversible desensitization data clustering are shown in Figure 13. These load curves can be classified into 5 classes, the main types are single-peak type, peak averting type, low load rate type, high load rate type and slight fluctuation type. It is similar to the original data clustering results.

Table 1. Confusion matrix - origin data vs. desensitized data.

|                   | Original data clustering | Irreversible desensitization data clustering |
|-------------------|--------------------------|---------------------------------------------|
|                   | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| Cluster 1         | 591       | 12        | 7         | 12        | 14        | 591       | 12        | 7         | 12        | 14        |
| Cluster 2         | 12        | 411       | 3         | 0         | 3         | 12        | 411       | 3         | 0         | 3         |
| Cluster 3         | 0         | 0         | 76        | 0         | 0         | 0         | 0         | 76        | 0         | 0         |
| Cluster 4         | 7         | 0         | 2         | 732       | 0         | 7         | 0         | 2         | 732       | 0         |
| Cluster 5         | 0         | 0         | 0         | 0         | 118       | 0         | 0         | 0         | 0         | 118       |

Based on the original data clustering results, we applied the Confusion Matrix to measure how well irreversible desensitization data clustering results match the original as shown in Table 1. It can be seen that the accuracy of the five cluster clusters are 96.89%, 97.16%, 86.36%, 98.39% and 87.41%. The high accuracy proves that the irreversible desensitization method can support the clustering model analysis effectively.

Summary

With the incoming wave of big data, the huge commercial value of large data in power system can be mined and available. Meanwhile, some thorny problems appear in privacy and sensitive information protection. In this paper, two kinds of data desensitization methods for model analysis were proposed by selecting different cut-off frequency, type of filter based on Fourier transform theory, and we used them in the typical application of electricity data mining--load clustering analysis. The experimental results show that the two kinds of data desensitization methods can effectively support cluster analysis: the recoverable has high flexibility; the irreversible can reduce data dimensions reduction and save computational resources. The proposed data desensitization methods can not only protect the sensitive and private information from leakage, but also achieve sharing, analysis and mining of large data effectively.

Acknowledgement

This paper is supported by research project of State Grid Corporation of China SGTYHT/15-JS-191 project.
References

[1] Li Zhiyong, Wu Jingying, Wu Weilin, etc. Power Customers Load Profile Clustering Using the SOM Neural Network, J. Automation of Electric Power Systems, 2008, 32(15):66-70.

[2] Zhou Kai-le, Yang Shan-lin. An improved fuzzy C-means algorithm for power load characteristics classification, J. Power System Protection and Control, 2012(22):58-63.

[3] Jiang De-long, Wang Ke-wen, Wang Xiang-dong. Clustering method of fuzzy equivalence matrix to bad-data detection and identification, J., Power System Protection and Control, 2011, 39(21):1-6.

[4] Liao Ni-huan, Hu Zhi-hong, Ma Ying-ying, etc. Review of the short-term load forecasting methods of electric power system, J. Power System Protection and Control, 2011, 39(1):147-152.

[5] Tian S, Wang B, Jing Z. Key Technologies for Demand Response in Smart Grid[J]. Proceedings of the Csee, 2014, 34(22):3576-3589.

[6] Bian Chaoyi, Zhu Shaomin, Zhou Tao. Implementation and evaluation of big data desensitization system based on format-preserving encryption, J. Telecommunications Science, 2017(03).

[7] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2006, 5(4):1-18.

[8] He Ling, Wu Ling-da, Cai Yi-chao. Survey of Clustering Algorithms in Data Mining, J. Application Research of Computers, 2007, 24(1):10-13.