Data Sensitivity Measurement and Classification Model of Power IOT based on Information Entropy and BP Neural Network

Qianyi Zhang 1*, Chi Zhang 1, Jiaming Ni 1, Xuqiang Wang 2 and Yao Zhang 2

1 State Grid Tianjin Information & Telecommunication Company, 300010, Tianjing, China.
2 State Grid Tianjin Electric Power Company, 300010, Tianjing, China.
*Corresponding author’s e-mail: qianyi.zhang@tj.sgcc.com.cn, 5910586540@163.com.

Abstract. For the problems of privacy data protection caused by massive data sharing in the construction of power Internet of things, a data sensitivity measurement and classification model based on information entropy and BP neural network is proposed. Firstly, a recognition matching algorithm is proposed to identify the sensitive level of attributes in the dataset, and the information entropy is used to determine the weight of attributes sensitivity level, so as to calculate the sensitivity measurement value of records in the dataset; finally, the BP neural network is used to output the data classification results. The experimental results show that the model can achieve accurate measurement and classification of data, with low incorrect judgment rate and error rate.

1. Introduction

With the comprehensive promotion of power Internet of things (IOT) and the orderly construction of data centre, the integration and penetration of various data and the extensive data sharing of cross business become popular. How to ensure data privacy and provide efficient data algorithm model is particularly important. After international and domestic network security laws and regulations have been issued, various industries have launched research on data security and privacy protection.

Many scholars have put forward some effective privacy protection models and methods based on data disruption, data anonymity and other strategies [1-4]. However, in practical application, it is still restricted by the privacy data types and complex privacy application scenarios. As privacy is an abstract concept, it is difficult to form a clearly definition, which makes it difficult to identify privacy data [5]. Hence, there have been many effective research results on data privacy measurement. Li et al. [6] proposed a privacy metric based on calculating the distribution value of sensitive attributes by k-anonymity model. Gkountouna et al. [7] were also based on the principle of anonymity, a predictive binary tree graph was constructed to measure the risk of privacy.

At present, the research on encryption, desensitization, leakage prevention and other aspects of power data security has been carried, but the research on data sensitivity classification of power IOT using natural language processing technology has not been paid attention. The data of power IOT is huge, and usually classified manually with low efficiency. For the problems that the sensitive data of power IOT cannot be automatically recognized and classified, a data sensitivity measurement and classification model for structured datasets is proposed in this paper. Based on the data sensitivity classification word
library, a recognition matching algorithm is proposed to identify the sensitive level of attributes, such as top-secret\(L_4\), confidential\(L_5\), secret\(L_6\), classified\(L_7\) or public\(L_8\). Then the sensitive weight of attributes in the dataset is calculated by using the Shannon information entropy, and sensitivity measurement values of records are calculated, thus constituting the sensitivity measurement vector of records. Finally, the classification results of records in dataset are output through BP neural network, realizing the automatic and intelligent classification of data.

2. Preliminary

2.1. Information entropy

The information entropy \[8\] derived from the thermodynamics, which is proposed by C.E.Shannon. In thermodynamics, thermal entropy is a physical quantity that represents the confusion degree of molecular state, while Shannon uses the concept of information entropy to describe the uncertainty of information sources.

Suppose that there are \(n\) states in a system \(X\), denoted as \(X\{x_1, x_2, \cdots, x_n\}\), \(p(x_i) (i = 1, 2, \cdots, n)\) denotes the appearing probability of \(x_i\) in system \(X\), The Shannon entropy \(H(x)\) of system \(X\) is defined as:

\[
H(x) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i))
\]

where \(0 \leq p(x_i) \leq 1\) and \(\sum_{i=1}^{n} p(x_i) = 1\). When \(p(x_i) = 0\), \(0 \log(0) = 0\).

According to Shannon's theory of information entropy, the greater the information entropy is, the higher the disorder degree of information is, and the less information it contains; the smaller the entropy is, the lower the disorder degree of information is, and the more information it contains.

2.2. BP neural network

BP neural network \[9\] is a kind of multilayer feedforward network trained by error back propagation with BP algorithm. The basic idea of the algorithm adopts gradient search technology to minimize the mean square error between the actual output value and the expected output value of the network. The network is divided into input layer, hidden layer and output layer. All neurons were connected with the neurons in the next layer, but there was no connection between the neurons in the same layer. The BP neural network with four input layer neurons, five hidden layer neurons and three output layer neurons is shown in Figure 1.

![Figure 1. Structure of BP neural network](image)

The advantage of BP neural network is that it can learn and store a large number of input-output relations, without the requirement of mathematical relationship in advance. It consists of two processes of signal forward propagation and error back-propagation. In forward propagation, the input signal acts on the output node through the hidden layer and generates the output signal through nonlinear transformation. If the actual output does not match the expected output, it will turn into the error back-propagation process. Error back propagation is that the output error is transmitted to the input layer gradually through the hidden layer, and the error is allocated to all units in each layer. The error signal obtained from each layer is used as the basis for adjusting the weight of units.
3. Construction of Model for sensitivity measurement and data classification

The power is with large amounts of data and complex data types in the dataset, which contains unlimited information value. But how to measure the sensitivity degree of these information is not clearly defined. There are many factors will affect the determination of data sensitivity classification, such as the diversity and relevance of the application scenarios of the information contained in each record, as well as the differences of subjective factors. Without setting the weight of sensitivity measurement elements, a data sensitivity measurement and classification model based on information entropy and BP neural network is proposed this paper. The basic framework consists of three models, as shown in Figure 2.

The first is attributes sensitivity level determination model. Based on the data sensitivity classification library, it uses the minimum edit distance algorithm to realize the sensitivity level identification of field attributes, including top-secret level($L_a$), confidential level ($L_b$), secret level ($L_c$), classified level ($L_d$) or public level($L_e$), so as to obtain the sensitivity level distribution of attributes in the records.

The second is sensitivity measurement model. Because the sensitivity degree of attributes with the same sensitivity level in each record is not the same. In this paper, the information entropy is used to determine the weight of attributes in the same sensitive level, and finally the sensitivity measurement values of each record under each sensitive level are calculated, thus constituting the sensitivity measurement vector of each record.

The third is BP data classification model. With the self-learning and storage ability of BP neural network, the training dataset is extracted to train the BP neural network, and the data classification model is established. Finally, the classification results of records to be tested are obtained though BP neural network.

Figure 2. Model structure

3.1. Determination of attribute sensitivity level

In order to accurately classify the data sensitivity, a data security classification library is established according to the provisions of the Data Classification Regulation of China State Grid Corporation, which listed the words of five sensitive levels including top-secret($L_a$), confidential($L_b$), secret($L_c$), classified($L_d$) or public($L_e$) respectively. On this basis, an attribute recognition matching algorithm based on the minimum edit distance is proposed, which matches the attribute string in the dataset with the strings set of the word library, so as to determine the sensitivity level of the attribute.

For two strings to be matched, the index string $U = (u_1, u_2, \ldots, u_N)$ and the target string $V = (v_1, v_2, \ldots, v_M)$, the minimum edit distance (MED) between two strings is defined as the minimum cost of converting one string to the other, given three basic operations; insertion, deletion and substitution, and their associated costs.

A two-dimensional matrix, the penalties matrix $C(0, \ldots, N)(0, \ldots, M)$ is used to hold edit distance values, where $N$ and $M$ are lengths of the strings, $U$ and $V$. MED between $U$ and $V$ is computed as follows:

**Step1.** Initializes a penalties matrix $C(N, M)$, $C_{0,0} = 0$;

**Step2.** Initialize the first-row elements of the penalties matrix:
\[ C_{0,j} = C_{0,j-1} + D(v_j), \quad 0 < j \leq M; \]  
\[ C_{i,0} = C_{i-1,0} + I(u_i), \quad 0 < i \leq N; \]

**Step3.** Initialize the first column elements of the penalties matrix: \[ C_{i,0} = C_{i-1,0} + I(u_i), 0 < i \leq N; \]

**Step4.** Update elements of the penalties matrix:
\[
C_{i,j} = \min \{ C_{i,j-1} + D(v_j), C_{i-1,j} + I(u_i), C_{i-1,j-1} + S(u_i, v_j) \}, \\
0 < i \leq N, 0 < j \leq M.
\]

where \( S(u_i, v_j) \) is the penalty of substituting the element \( u_i \) in \( U \) with the element \( v_j \) in \( V \), \( I(u_i) \) is the penalty of inserting the element \( u_i \) in \( U \), \( D(v_j) \) is the penalty of deleting the element \( v_j \) in \( V \), \( C_{N,M} \) is the minimum penalties of converting the string \( U \) to the string \( V \). Though comparing the minimum penalties with the threshold, the matching between two strings is completed.

### 3.2. Data sensitivity measurement

In this paper, the information entropy measurement matrix is established for each attribute sensitivity level of the records in the dataset. Suppose the number of attributes contained in the top-secret level \( L_a \) is \( b \), and the sensitivity measurement results for records with number of \( n \) are calculated by an attribute information entropy measurement matrix \( X_a(n \times b) \). The steps are as follows.

**Step1.** The attribute information measurement matrix \( X_a \) is established for \( n \) regularized records, in which there are \( b \) attribute records belong to sensitivity level \( L_a \).

\[
X_a = \begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1b} \\
    x_{21} & x_{22} & \cdots & x_{2b} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & \cdots & x_{nb}
\end{bmatrix}
\]

\( x_{ij} \) in matrix \( X_a \) is the value corresponding to the \( j \)th attribute in the \( i \)th record after normalization, with the value of 0 or 1.

**Step2.** The elements in matrix \( X_a \) are transformed, as shown in the formula (6).

\[
x_{ij}^* = \frac{1}{\sum_{i=1}^{n} x_{ij}}
\]

then

\[
X_a^* = \begin{bmatrix}
    x_{11}^* & x_{12}^* & \cdots & x_{1b}^* \\
    x_{21}^* & x_{22}^* & \cdots & x_{2b}^* \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1}^* & x_{n2}^* & \cdots & x_{nb}^*
\end{bmatrix}
\]

**Step3.** According to the information entropy formula (1), the information entropy of each attribute \( j \) is calculated

\[
E_j = -k \sum_{i=1}^{n} x_{ij}^* \ln(x_{ij}^*)
\]

where \( k = \frac{1}{\ln(n)} \) \( j \) is the number of attributes. \( j = 1, 2, \cdots, b \). When \( x_{ij}^* = 0 \), \( E_j = 0 \).

**Step4.** Calculate of the attribute weights

\[
W_j = \frac{1 - E_j}{b - \sum_j^{b} E_j}
\]

The sensitivity measurement value of a single record \( d_i \) under the sensitivity level \( L_a \) is calculated:

\[
D_{ai} = \sum_{j=1}^{b} W_j x_{ij}^*, \quad i = 1, 2, \cdots, n.
\]

The higher the value of \( D_{ai} \), the higher the sensitivity of the record \( d_i \) under sensitivity level \( L_a \).

**Step5.** Repeat steps 1-4 to calculate sensitivity measurement value of a single record \( d_i \) under five sensitivity levels \( L_a, L_b, L_c, L_d, L_e \), the sensitivity measure vector of \( d_i \) is generated as follows:

\[
F(d_i) = \{ D_{ai}, \cdots, D_{ei} \}.
\]

### 3.3. Data classification

In this paper, the final classification result of privacy data is obtained based on BP neural network. The number of nodes in the input layer is five, which corresponds to the number of sensitive levels. The number of nodes in the output layer is also five, with output value 0 or 1. The output measured as top-secret level is \{1,0,0,0,0\}, and as public level is \{0,0,0,0,1\}. In this way, the output vector corresponds
to five privacy levels respectively. The tangent function and logarithmic function of sigmoid type are used as transfer functions of hidden layer and output layer respectively.

In the training process of BP neural network, 20% of the records in the training dataset are randomly selected, and the sensitivity measurement vector value of the training data is obtained by the sensitivity measurement method based on information entropy in Section 3.2, and the training samples were formed after normalization. The specific training process was shown in Figure 2. After 131th training of Neural network, the training target is achieved, which can meet the needs of the model.

4. Experiments and results

4.1. Training of BP neural network

The test of model focused on data classification based on BP neural network, mainly for neural network training test and accuracy test of classification. During the testing, 5000 data records was used as the training dataset, with training number of 1000, the learning efficiency of 0.1, and the target error of 0.000001. According to the formula $N = \sqrt{I + O + \alpha}$, the number of hidden layer nodes was preliminarily selected, where $I$ and $O$ are the number of input nodes and output nodes respectively, $\alpha$ is adjustment parameter. Finally, the number of hidden layer nodes was nine. The error curve of the training process is shown in Figure 3.

![Figure 3. Training error curve of BP neural network](image)

4.2. Classification accuracy experiments

For the accuracy test of data sensitive classification of this model, the error rate of misjudgment [10] is used to measure the deviation between the output of the model and the actual classification results. The formula is as follows.

$$\varepsilon = \sum_{i=1}^{k} e|D_i - D_i'| / k$$

(11)

where $k$ is the number of records to calculate the error rate of misjudgment, $e$ is the natural logarithm, $D_i$ and $D_i'$ are the actual sensitivity level and output of model for the record with index number of $i$, $|D_i - D_i'| = 0,1,2,3,4$.

The testing dataset included 500 records to verify the classification accuracy of the model. The test results were shown in Table 1. It proved that the overall accuracy of the data classification can reach 96.4%.

| Sample sensitive level | Samples number | Misjudgment number | Misjudgment rate | Misjudgment deviation rate |
|------------------------|----------------|--------------------|------------------|---------------------------|
| $L_a$                  | 100            | 2                  | 2%               | 1.89%                     |
| $L_b$                  | 100            | 3                  | 3%               | 6.12%                     |
| $L_c$                  | 100            | 4                  | 4%               | 8.65%                     |
| $L_d$                  | 100            | 3                  | 3%               | 7.44%                     |
| $L_e$                  | 100            | 6                  | 6%               | 13.37%                    |
| Total                  | 500            | 18                 | 3.6%             | 7.49%                     |
5. Conclusion

For the problems that the sensitive data of power IOT cannot be automatically recognized and classified, an intelligent and automatic data sensitivity measurement and classification based on information entropy and BP neural network is proposed. The model can effectively identify and classify the sensitivity of any structured dataset, and it is also suitable for large-scale structured data sets. Next, we will explore the algorithm and model of adaptive privacy protection for large-scale data based on sensitive data identification and classification.

References

[1] Sweeney L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5):557-570.
[2] Machanavajjhala A, Kifer D, Gehrke J. L-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):3
[3] Dwork C, McSherry F, Nissim K. Calibrating noise to sensitivity in private data analysis[J]. Proceedings of the VLDB Endowment, 2006, 7(8):637-648
[4] Dwork C, Roth A. The algorithmic foundations of differential privacy[M]. Boston: Now Publishers Inc. 2014
[5] Guzman J A D, Thilakaratna K, Seneviratne A. Security and privacy approaches in mixed reality: a literature survey[J]. arXiv: 1802.05797v2[cs.CR], 2018
[6] Li N, Li T, Venkatasubramanian S. Closeness: a new privacy measure for data publishing[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(7):943-956.
[7] Gkountouna O, Terrovitis M. Anonymizing collections of tree-structured data[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(8):2034-2048.
[8] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(4):379-423.
[9] Zhang G, Zhang Z, He Y. Review on BP neural network applied in the textile and clothing Field[J]. Shandong Textile Science & Technology, 2013.
[10] Yu Yihan, Fu Yu, Wu Xiaoping. Metric and classification model for privacy data based on Shannon information entropy and BP neural network [J]. Journal on Communications, 2018, (012):10-17.