Big Data Cleaning Model of Multi-Source Heterogeneous Power Grid Based On Machine Learning Classification Algorithm

Zhangchi Ying, Yuteng Huang, Ke Chen*, Tianqi Yu
State Grid Zhejiang Electric Power Corporation Information & Telecommunication Branch, Hangzhou, Zhejiang, 310016, China
*Yingzhangchi@zj.sgcc.com.cn

Abstract—Aiming at the low cleaning rate of the traditional multi-source heterogeneous power grid big data cleaning model, a multi-source heterogeneous power grid big data cleaning model based on machine learning classification algorithm is designed. By capturing high-quality multi-source heterogeneous power grid big data, weight labeling of data source importance measurement, data attributes and tuples, and constructing Tan network based on the idea of machine learning classification algorithm, the data probability value is finally used to complete the classification and cleaning of inaccurate data. Experiments show that the model based on machine learning classification algorithm can effectively improve the imprecise data cleaning rate compared with the traditional model to solve multi-source heterogeneous imprecise data cleaning.

1. Introduction
The performance of multi-source heterogeneous power grid big data analysis depends on the quality of data. The correct use of high-quality data can help make better prediction and decision-making, as well as more reliable data analysis[1]. The success of big data analysis of multi-source heterogeneous power grid largely depends on how the data is cleaned up, integrated and transformed. If the data is not cleaned up correctly, no matter how complex the intelligent analysis algorithm can not obtain the ideal analysis results[2]. Multi source heterogeneous power grid big data has a wide range of sources, often including several data sources, and these data sources usually come from homogeneous or heterogeneous databases, file systems and service interfaces, which reduces the reliability of data. Assuming that uncertain big data is applied to the subsequent analysis process without processing, the data analysis results are often affected by problems such as noise data, missing data values and data conflict, resulting in inconsistent or inaccurate data analysis results[3]. However, with the development of big data analysis technology, a large number of data with complex structure, multi view and multi-source description bring new challenges to data cleaning[4]. Based on the success of machine learning in the fields of pattern recognition, information retrieval and data mining, it provides a new solution for data processing, but the traditional statistical machine learning methods can not be simply used in dealing with multi-source heterogeneous data. Because the traditional statistical machine learning methods assume that the data to be processed come from the same feature space and have the same distribution. Therefore, this paper designs a multi-source heterogeneous power grid big data cleaning model based on machine learning classification algorithm, and committed to improving the big data cleaning rate of multi-source heterogeneous power grids.
2. Big data cleaning model of multi-source heterogeneous power grid based on machine learning classification algorithm

2.1. Capturing high-quality multi-source heterogeneous power grid big data

The uncertainty and inconsistency of data is a common problem in real life. Imprecise data refers to the data containing incompleteness, uncertainty, inconsistency and noise. The biggest problem faced by data analysis is to extract useful information from incomplete and fuzzy data, because imprecise data means the existence of conflicting versions of the same data, resulting in unreliable final results\(^5\). Even very insignificant wrong data will reduce the performance of mining analysis algorithm at a very high rate. Therefore, in order to obtain correct results, these imprecise data must be identified and eliminated\(^6\). In multi-source data cleaning, it must be considered that each data source may involve different data fields and different data forms. Therefore, there are many reasons for inaccurate data.

There are generally two ways to eliminate imprecise data in multi-source heterogeneous environment: one is to merge data into the same data source through data integration or data fusion, and remove imprecise data during and after data fusion. The other is a cross data source data cleaning method, which cleans inaccurate data for each data source at the same time by formulating a unified data cleaning standard. Compared with the first method, the cross data source data cleaning method is difficult to clean and the cleaning efficiency is uncertain. However, if the cleaning model is appropriate, the cleaning efficiency is greatly improved compared with the first method. Although data integration and data fusion technology have been developed for many years, due to the great differences between different data sources in coding, naming, data type and semantics, integrating data into the same data source through data extraction, data conversion and data reprint (i.e. ETL process) often can not get ideal results, especially when the data scale is large. When the data type is complex, the difficulty of data integration can be imagined.

For the cleaning of inaccurate data across data sources, the biggest problem is that the data cleaning process is not well controlled, the effect of data cleaning of each data source cannot be guaranteed, and the data quality problems caused by merging data sources will not be cleaned\(^7\). At the same time, due to different knowledge fields of multiple data sources, different data types and data representation, the cost of data repair will be very high. In reality, the scale of data sources is becoming larger and larger. When facing multiple data sources, how to improve the execution efficiency of the algorithm and reduce the complexity of data cleaning, and find the most useful data from many data is a challenge that cleaning needs to deal with. Therefore, in multi-source heterogeneous environment, how to accurately capture high-quality data is the key to clean inaccurate data.

2.2. Processing multi-source heterogeneous power grid big data based on machine learning classification algorithm

The processed metadata features are collected to form a multi feature collection. The main method of processing a large number of data sets is to establish and check the network graph model formed by mutual relations\(^8\). Graph model is a method of knowledge representation, learning and reasoning of the structure and relationship between data based on probability framework, and this method can well describe the uncertainty of data. Due to the dependence between various attributes in the data source, the data source model is represented as a Bayesian network, and the relationship between attributes is naturally captured through Bayesian network structure learning and the probability distribution of input attributes and tuples. Supposing a group of random variables \(U = \{X_1, X_2, ..., X_n, c\}\) and \(c\) is class variables, the value range is \(\{c_1, c_2, ..., c_m\}\), \(m\) is the total number of classes, \(\{x_1, x_2, ..., x_n\}\) is the attribute value of \(\{X_1, X_2, ..., X_n\}\) showing the characteristics of classification, and \(n\) is the number of attributes of classification. Tan classifier assumes that the structure of Bayesian network composed of attribute nodes \(\{X_1, X_2, ..., X_n\}\) is a tree, and each attribute variable has no more than one attribute
parent node except the parent class. Class node is the parent node of each attribute node, and a tree is formed between attribute nodes as the maximum weight span tree.

In general, when constructing Tan network, the weight between attributes is to calculate the mutual information of attributes between calculation variables\([9]\). Mutual information refers to the correlation degree between two random variables, that is, the weakening degree of uncertainty of another random variable after a random variable is given. Mutual information \(i(x, y)\) is defined as:

\[
i(x, y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

In formula (1), \(p(x, y)\) is the joint distribution of variables \((x, y)\), and \(p(x)\) and \(p(y)\) are the edge distribution respectively. The mutual information between attributes is the relevance of attributes. The attribute relevance values calculated by different class attributes are also different. Considering the addition of class variable attributes of Tan network, the mutual information formula of a classification attribute needs to be redefined. Therefore, the calculation formula of mutual information of Tan network is:

\[
i_{ij} = (C_i, C_j | C) \quad (2)
\]

In formula (2), \(C_i, C_j\) is an attribute variable and \(C\) is a class variable. The calculation of classification result probability is to transfer the query condition attribute into the classification model, and then calculate the probability value under different types of attributes. The classification attribute value with the maximum probability value is the final classification result, that is, the set with the maximum probability of imprecise data, that is, calculate its joint probability distribution:

\[
P(C_1, C_2, \ldots, C_n) = \prod nP(|\text{arg max}(C_i)) \quad (3)
\]

In formula (3), \(P\) is the weight of the attribute. Through formula (3), the processing of multi-source heterogeneous power grid big data based on machine learning classification algorithm is realized.

2.3. **The big data cleaning model of multi-source heterogeneous power grid is obtained**

According to the above analysis, for a given task, some data sources in the multi-source environment may be irrelevant or redundant. Therefore, selecting a group of data sources related to the task from the multi-source environment is of great significance to improve the efficiency and accuracy of data cleaning\([10]\). By capturing this association relationship, we can judge the importance of the data source. Reduction algorithm based on attributive importance:

- **Input**: Decision tables \(S=(U, Q, V, F)\);
- **Output**: Simplest attribute set

Stept 1: The discernibility matrix of the decision table is calculated, and the core attribute in the discernibility matrix is assigned to the attribute set obtained after attribute reduction, that is, \(\text{Red} = \text{Core}\);

Stept 2: Removing the core attribute in the discernible recognition matrix and reduce all remaining attribute combinations;

Stept 3: Calculating the occurrence frequency of each conditional attribute, arrange all attribute frequencies in descending order, select the attribute with the highest attribute frequency as \(A_1\), \(\text{Red} = \text{Red} U (A_1)\), and delete the combination item containing conditional attribute \(A_1\) from all combination items of the variable matrix.

Stept 4: Judging whether the discernible matrix is empty. If the discernible matrix is not empty, turn to step 3. If the discernible matrix is empty, it ends. \(\text{Red}\) is the final reduction result. The big data reduction process of multi-source heterogeneous power grid is shown in Figure 1.
As shown in Figure 1, the algorithm has guiding significance for the research of attribute reduction, but when there are many important attributes in the decision table, the algorithm is more complex, and the attribute frequency is used as the only standard of attribute importance, but when the attribute frequency is the same, there is no way to measure the attribute importance with the same attribute frequency. Thus, the big data cleaning model of multi-source heterogeneous power grid can be obtained. If the calculation expression is set to $\tau$, the calculation formula is as shown in formula (4).

$$\tau = \frac{\sqrt{\text{ETS}}}{\sqrt{\sum_{i=1}^{n} \left(\frac{1}{m_i}\right)^2}}$$ (4)

In formula (4), $L$ refers to the induced value of multi-source heterogeneous power grid big data; $m$ refers to the feature similarity of regulatory data in power operation process. Through formula (4), the big data cleaning model of multi-source heterogeneous power grid can be obtained. According to the big data cleaning model of multi-source heterogeneous power grid, the conditional attributes are subdivided, and then the attribute reduction set is directly output by comparing the relatively positive regions after removing the conditional attributes. After the improvement of the algorithm, the algorithm mainly calculates and compares the relative positive region from the classification of conditional attributes through the cumulative removal of conditional attributes, so as to judge whether the core attributes and all important attributes are removed. Finally, the qualified conditional attributes are added to the attribute reduction set, and the final attribute reduction set is output.

3. Experiment

3.1. Experimental preparation

In order to verify the accuracy of the data cleaning process in this paper, prototype experiments are carried out based on Hadoop cluster, hive cluster and SQOOP cluster. Hadoop is a distributed file
system. Using Hadoop cluster can process large-scale data in parallel. Hive is mainly responsible for mapping structured data files into database tables one by one, and can also provide the function of SQL query, and then convert SQL statements into MapReduce tasks and upload them to the cluster to realize. Sqoop is mainly responsible for data transmission on Hadoop (hive) and traditional data databases (Mysql, PostgreSQL...), and can also realize the transfer of data, such as Mysql, Oracle, The data in the relational database such as Postgres is transferred to HDFS of Hadoop, and the data of HDFS can also be transferred to the relational database. The data processing flow is as follows:
(1) Importing the data from relational database into Hadoop cluster through sqoop.
(2) The data is processed preliminarily through hive, and then the data in the previous step is further cleaned up through pig.
(3) After the preprocessed data is obtained, HRSC (hierarchical reduction classification cleaning) strategy and ARCC (attribute reduction joint cleaning 1 strategy) are called to further process the data.
(4) After cleaning up, export it to the relational database through sqoop again; The virtual machine parameters of the experimental environment are shown in Table 1.

### Table 1 Parameters of virtual machine in experimental environment

| Name                       | CPU (nucleus) | Memory (G) | Quantity (PCs.) |
|----------------------------|---------------|------------|-----------------|
| Hadoop virtual machine1    | QEMU Virtual CPU, 4 | 126        | 1               |
| Hadoop virtual machine2    | QEMU Virtual CPU, 4 | 64         | 1               |
| Hadoop virtual machine3    | QEMU Virtual CPU, 4 | 32         | 1               |
| Mysql database virtual machine1 | QEMU Virtual CPU, 4 | 8          | 1               |
| Mysql database virtual machine2 | QEMU Virtual CPU, 4 | 8          | 1               |
| Hive virtual machine1      | QEMU Virtual CPU, 4 | 8          | 1               |
| Hive virtual machine2      | QEMU Virtual CPU, 4 | 8          | 1               |
| Hive virtual machine3      | QEMU Virtual CPU, 4 | 8          | 1               |
| Sqoop virtual machine1     | Xeon E5-2407,4 | 8          | 1               |
| Sqoop virtual machine2     | QEMU Virtual CPU, 2 | 8          | 1               |
| Sqoop virtual machine3     | Xeon E7-4850*2,12 | 8          | 1               |

As shown in Table 1, a multi-source heterogeneous power grid big data set is selected as the sample data. The main content of the experiment is to test the cleaning rate of two cleaning models for multi-source heterogeneous power grid big data, so as to evaluate the cleaning model with higher cleaning quality. In this experiment, five data sources are set up to clean the multi-source heterogeneous power grid big data using the design in this paper and the traditional model, and their respective cleaning rates are recorded by MatalB software to obtain the experimental results.

#### 3.2 Analysis and conclusion of experimental results

The experimental results are shown in Table 2.

### Table 2 Comparison results of model cleaning rate

| Number of experiments | data source | The cleaning rate of this model | Cleaning rate of traditional model |
|-----------------------|-------------|---------------------------------|-----------------------------------|
| 1                     | data source 1 | 88.77%                          | 65.24%                            |
| 2                     | data source 2 | 89.52%                          | 64.20%                            |
| 3                     | data source 3 | 89.32%                          | 65.11%                            |
| 4                     | data source 4 | 88.97%                          | 64.38%                            |
| 5                     | data source 5 | 89.20%                          | 64.09%                            |
The following conclusions can be drawn from table 2: the cleaning rate of the model designed in this paper is significantly higher than that of the traditional model. Through the verification results, it is proved that the functions of the designed cleaning model can meet the overall design requirements, and can be widely used in multi-source heterogeneous power grid big data cleaning.

4. Conclusions
In this paper, aiming at the problem of a large number of inaccurate data in multi-source heterogeneous data environment, this paper proposes a data cleaning strategy of hierarchical reduction and classification cleaning. Firstly, the importance of data sources is measured to fundamentally reduce irrelevant or redundant data sources, which greatly reduces the workload of data cleaning. Then, the data attributes and tuples are weighted by data density, and the core tuples and edge tuples are retained according to the weight to reduce outliers. Finally, based on the idea of machine learning classification algorithm, the augmented tree Bayesian Tan network is constructed through attribute weight, and the probability value of Tan network is used to classify imprecise data and accurate data. Compared with the current data cleaning technology, the cleaning strategy of machine learning classification in this paper has higher cleaning efficiency. Experimental analysis shows that the hierarchical reduction algorithm proposed in this paper can effectively extract the required relevant data, especially for the complex environment of multi-source heterogeneous data. In the correlation reduction algorithm, it can identify the relevant attributes in multiple data sources, and the reduction effect of data source layer, data surface layer and data layer is also prominent.

References
[1] Basgalupp M P , Barros R C , AGCD SM, et al. (2020) An extensive experimental evaluation of automated machine learning methods for recommending classification algorithms[J]. Evolutionary Intelligence, (3).
[2] Penumuru D P , Muthuswamy S , Karumbu P . (2020) Identification and classification of materials using machine vision and machine learning in the context of industry 4.0[J]. Journal of Intelligent Manufacturing, 31(5):1229-1241.
[3] Jancarczyk D , Berna M , Boczar T . (2019) Classification of Low Frequency Signals Emitted by Power Transformers Using Sensors and Machine Learning Methods[J]. Sensors, 19(22):4909.
[4] Ma X, Sha J, Wang D, et al. (2018) Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J]. Electronic Commerce Research and Applications, 31: 24-39.
[5] Basiri M E , Abdar M , Cifci M A , et al. (2020) A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques[J]. Knowledge-Based Systems, 198.
[6] Rohini M , Surendran D . (2019) Classification of Neurodegenerative Disease Stages using Ensemble Machine Learning Classifiers[J]. Procedia Computer Science, 165:66-73.
[7] Gudivada V, Apon A, Ding J. (2017) Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations[J]. International Journal on Advances in Software, 10(1): 1-20.
[8] Corrales D C, Ledezma A, Corrales J C. (2020) A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks[J]. Applied soft computing, 90: 106180.
[9] Quatrini E , Costantino F , Gravio G D , et al. (2020) Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities[J]. Journal of Manufacturing Systems,56:117-132.
[10] P R10 P , Papenberg B , Tracht K . (2021) Classification of assembly operations using machine learning algorithms based on visual sensor data[J]. Procedia CIRP, 97:110-116.