big data analysis 3-3

1 Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering

School of Computer Science, in Beijing Institute of Technology, in this session, we will discuss Data Cleaning Technology.

2

**Data Cleaning Technology can help to solve three kinds of problems, repeat record cleaning, Missing value cleaning and Eliminating noise data.**

**First let’s look at the Repeat record cleaning**

3

Duplicate records will lead to erroneous analysis results, so it is necessary to remove duplicate records in the data set to improve the accuracy and speed of the analysis.

1 what are the Causes of duplicate values

1.1 Integrate data from multiple data sources will result the duplicate values

1.2 When inputting, some data is input repeatedly

2Duplicate values are combined after inference

2.1 Delete completely duplicate records

2.2 When merging different tables, some redundant attributes (such as time) is added

4

* It is necessary to compare the related attributes of two records. According to **the similarity of each attribute and the weight of the attribute**, the similarity of the record is obtained.
* If the similarity exceeds a certain threshold, it is considered as a duplicate record.

5

Now let’s look at the **Missing value cleaning.**

6

Causes of missing values could be

Equipment abnormal, or When entering, some data is not taken seriously and not entered.

Missing values must be inferred and added. We can Ignore this record; Use default; Use attribute average;

Use the average of similar samples and Predict the most likely value

7

The third **Data Cleaning Technology is Eliminating noise data**

8

Data noise can be Eliminated by

1. **Binning/split bin algorithm**
2. Clustering Algorithm
3. Regression algorithm.

First let’s look at Binning **algorithm**

9

* **Binning:**

Put the data to be processed into some boxes according to certain rules, examine the data in each box, and use a certain method to process the data in each box separately.

* **Box:**

A sub-interval divided by attribute value. If an attribute value is within a certain sub-interval range, it is said to put the attribute value in the "box" represented by this sub-interval.

* **The main issues:**

How to divide the box;

Data smoothing method, that is, how to smooth the data in each box, which means use what methods to decide the value to present each box.

In the example here, there are 4 different bins, grey, yellow green and purple, each age value is put to certain range, like age 10 is put into range from 10 to 16, the grey box. And so, on

10

The ways of how to **Sort the record set according to the size of the target attribute value before binning, include**

* **Equal depth binning**
* **Equal-width binning**
* **User-defined interval**

11

**Equal depth bin method** (unified weight):

Bins are divided according to the number of rows of records,

each bin has the same number of records, and the number of records in each bin is called the weight of the bin, also known as the depth of the bin.

In this example, we have 16 records, and we decide the box depth is 4 , after binning, we have 4 boxes results and each box contain 4 records

12

**Equal-width binning method,** evenly distributed over the entire attribute value interval, that is, the interval range of each box is a constant, called the box width.

In the example, Set the interval range (the width of the box) to RMB 1,000, then the result is also 4 boxes, but the records in the boxes are different from the Equal depth bin method.

13

After binning, we should choose a value to represent the boxes, which is called smooth. there are mainly 3 ways to smooth:

**Smooth by average：**Average the data in the same box value and replace all the data in the box with the average value.

**Smooth according to the boundary value：**Replace each data in the box with a boundary value with a smaller distance.

**Smooth according to the median：**Take the median value of the box and use it to replace all the data in the box**.**

14

Second way of Eliminating Data noise is **Clustering Algorithm. Now let’s look at Clustering Algorithm**

15

**Cluster:**

A collection of data objects. All objects in the same cluster are similar, and objects in different clusters are quite different.

**Clustering:**

Grouping a collection of physical or abstract objects into different clusters, finding and clearing those values (outliers) that fall outside the clusters. These isolated points are regarded as noise.

**Unusual data is discovered through cluster analysis**:

similar or adjacent data are aggregated to form clusters, and those data objects outside these clusters are naturally considered as abnormal data.

**Features:** Directly form clusters and describe the clusters without any prior knowledge.

16

The third way of Eliminating Data noise is **Regression Algorithm. Now let’s look at Regression algorithm**

17

**Regression:**

Find the **pattern** of change between two related variables, which are x and y in the diagram, and smooth the data by fitting the data to a function, that is, using the fitting function to smooth the data.

method:

**Linear regression (simple regression):** Use straight line modeling to treat one variable as a linear function of another variable.

For example: Y=aX+b, where a and b are called regression coefficients, and the coefficients a and b can be obtained by the least square method.

18

**In this session we general introduced** Data Preprocessing.

**thank you for your attention, if you have any question, feel free to contact me.**