Bigdata analysis 3-5 data reduction

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 reduction .

2

Complex data analysis of large-scale database content usually takes a lot of time.

Data reduction (subtraction) technology is used to help obtain a **condensed data set** from the original huge data set, and make this condensed data set maintain the integrity of the original data set, so that data analysis on the condensed data set is obviously efficient higher, and the results of analysis are basically the same as those obtained by using the original data set.

3

Data reduction work should satisfy some standard.

1. The time spent on data reduction should not exceed or "offset" the time saved by analysis on the reduced data
2. The data obtained by the reduction is much smaller than the original data, but can produce the same or almost the same analysis results

4

Data reduction technology includes dimensionality reduction and numerosity reduction.

And dimensionality reduction includes attribute subset selection, principal Component analysis and wavelet transform

Numerosity reduction could be divided into parametric and non-parametric methods.

non-parametric methods include data cube aggregation , clustering, sampling histogram etc.

5

First let’ s look at Dimension reduction

Remove irrelevant attributes and reduce the amount of data processed by data analysis.

6

Attributes subset selection Goal is to Find the smallest attribute subset and ensure that the probability distribution of the new data subset is as close as possible to the probability distribution of the original data set

For example, when digging into the classification rules of whether a customer will buy a player in a shopping mall, the customer's phone number is likely to be irrelevant to the mining task and should be removed.

7

**Step forward Attributes subset selection**

Starting from an empty attribute set (as the initial value of the attribute subset), each time a current optimal attribute is selected from the original attribute set and added to the current attribute subset. Until the optimal attribute cannot be selected or a certain threshold constraint is met.

8

**Remove gradually backward** Attributes subset selection

Starting from a full attribute set (as the initial value of the attribute subset), each time a current worst attribute is selected from the current attribute subset and eliminated from the current attribute subset. Until the worst attribute cannot be selected or a certain threshold constraint is met.

or **Combine forward selection and backward deletion**

9

**Decision tree (decision tree) induction**

Use **the decision tree induction method** to classify and induct the initial data to obtain an initial decision tree.

All attributes that do not appear on the decision tree are considered irrelevant attributes.

Therefore, delete these attributes from the initial attribute set to obtain an initial decision tree. A better subset of attributes.

There is also **Reduction based on statistical analysis**

10

Now let’s look at Data reduction by Data compression

11

Compression algorithm can be classified into **Lossless compression and Lossy compression .**

**Lossless compression:** Compressed data can be restored without losing any information.

For example: string compression

Have a broad theoretical foundation and sophisticated algorithms

**Lossy compression:** Only an approximate representation of the original data can be reconstructed.

For example: audio/video compression

Sometimes it is possible to reconstruct a fragment without decompressing the overall data

12

Data compression-use data encoding or transformation to obtain a compressed representation of the original data.

The two data compression methods commonly used in the field of data mining are both lossy:

**Principal component analysis (PCA)** assumes that the data to be compressed consists of N tuples or data vectors taken from k dimensions. Principal component analysis and search to obtain c “k-dimensional” orthogonal vectors that best represent the data, where c≤k. In this way, the original data can be projected into a smaller space to achieve data compression

**13**

We talked some technologies of dimensionality reduction and now let’s look at some technologies of numerosity reduction.

Numerosity reduction -use smaller data to represent data, or use shorter data units, or use data models to represent data to reduce the amount of data.

Numerosity reduction includes Data cube aggregation , clustering, sampling, histogram etc.

14

**Let’s look at the data cube aggregation**

A data cube is a multi-dimensional modeling and representation of data, composed of dimensions and facts.

* Dimension: attribute
* Facts: data

Data cube aggregation definition-gather n-dimensional data cubes into n-1 dimensional data cubes.

Using Data cube Aggregates for approximate query

In the diagram, 3 dimension data cube aggregate into 2 dimension, the men and women ’ sale amount aggregated together.

15

**Now let’s look at Data reduction method Discretization.**

**Attribute values could be Name type-e.g. value in an unordered set, Ordinal-e.g. value in an ordered set and Continuous value-e.g. real number.**

**Discretization technology**

**Reduce the number of values of a continuous (value) attribute by dividing the range of the attribute (continuous value) domain value into several intervals.**

16

Using Concept hierarchical generation to do the Data reduction

Concept hierarchy defines a set of mappings from low-level concept sets to high-level concept sets.

It allows data to be processed at various levels of abstraction, thereby discovering knowledge at multiple levels of abstraction.

**Use higher-level concepts to replace lower-level concepts** (such as the value of age in this example) to reduce the number of values.

Although some details disappeared in the process of data generalization,

But the generalized data obtained in this way may be easier to understand and more meaningful.

Data analysis on the reduced data set is obviously more efficient.

The concept hierarchy can be represented by a tree, and each node of the tree represents a concept of certain level.

17

**In this session we learned** Data reduction and related technologies**.**

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