Exploration of time-based data pretreatment technology

Wei jing[1] Yu wenqiang[2]

[1] North china institute of science and technology
[2] SK China Co. Ltd

Abstract

Among all the data processed by associated rules in this article, part of them is time data and it is also the key point of the analysis. To facilitate the concept hierarchy of the time data, this research brings out the (0, 1) standardized technology for time data normalization and use it to convert time data to numeric data skillfully. Then in the process of time data generalization, this research presets threshold value according to the data attributes, digging task requirement and other actual situation, also applies histogram analytical method of concept hierarchy technology to fine tuning on preset threshold value, then brings out histogram threshold fine tuning technology.

Keyword： Data Mining,Campus Network,Association Rule, (0, 1) Standardized

1. A data transformation - standardization technique

The data conversion is to convert the data into the form suitable for mining. In addition to smoothing, aggregation, data generalizability and attribute constructor, three main data normalization methods are as follows:

1.1. Minimum - maximum standardization

Minimum - maximum standardization is linear transformation of the original data. Assume that minA and maxA are respectively the minimum and maximum value of the attribute A. Minimum - maximum standardization is calculated by

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

A value of v is mapped to the interval [new_minA,new_maxA] v '

1.2. Z-score standardization (or zero - mean standardization)

For z-score standardization (or zero - mean standardization), the value of the attribute A is based on A's mean value and standard deviation standardization. A's value, v, is standardized to v ' and is calculated by the following formula:

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Wherein, respectively $\overline{A}$ and σ_A are attribute A's mean value and standard deviation. When the maximum and minimum values of the attributes are unknown, or as isolated points have control over the maximum - minimum standardization, this method is useful.

1.3. Decimal scaling standardization

The decimal scaling is standardized by moving the decimal point position of the attribute A. The decimal mobile digits depends on the maximum absolute value of A. A's value, v, is standardized to v ', is calculated by the following formula:

$$v' = \frac{v}{10^j}$$

Wherein, j is the smallest integer which meet max (| v '|) <1.

Below we will discuss the fourth data standardization methods - time data standardization methods.

2. The data reduction - the concept hierarchy generation method of data attributes: the histogram analysis

The histogram uses binning approximate data distribution which is a popular form of data reduction. The histogram of the attribute A divides A's data distribution into disjoint sets or barrel. The barrel is placed on the horizontal axis, while the height of the drum (and area) is the average probability of the values represented by the tub. If each barrel represent a single attribute value / frequency on the barrel it is called a single-barrel. Usually, the barrel is a continuous range with given attributes.

We can take advantage of some of the division rules to determine the barrels and the division of the property value, the rules include some of the following:

- Monospaced
- Equal depth(or equal height)
- V-optimal
- MaxDiff

3. Time data standardization technique discussion

Three existing data standardization methods were introduced earlier, next we're going to explore another time data standardization methods, here I call it the (0,1) standardization.

For example, there's a series of time data among the attribute data involved in the association rule mining analysis which is shown in Figure 4-1, and it is learned that the attribute data required in association rule mining is discrete, then it is needed to apply the discretization for the continuous time data before mining which is also called generalization, from it we will think of using the data concept hierarchy generation methods, where we use the histogram analysis. Like a time of 18:29, it looks like it's a little hard to use this form of data directly for the histogram analysis. To facilitate the generalization and data mining, we must first convert the data, below is the process of converting specific time data to numerical data, that is, the (0,1) standardized principle and implementation process.

(1)The range of time values is [0,24], the length of the interval [0,24] is set to 1, then the time number can be standardized to the interval [0,1] via the following formula.

$$H' = \frac{H}{24}$$

(Time value H standardized to H ')

(2) However, the time data is generally consisted of two parts, such as: hh: mm. And the two parts of the unit are not unified, front hh means hours and behind mm means minute. So first of all, we want them expressed in unified hours. Here firstly we present mm in hh, that is mm=1/60 hh (hour). The entire time hours hh: mm = hh + 1/60*mm. For example: 18:30 = 18 + 30/60 = 18.5 hours.

(3) use the formula from (1) to standardize the time value obtained from (2) to the interval [0,1], which will complete time data conversion task. Specific conversion formula is as follows:

$$H' = \frac{\mathrm{hh} + mm/60}{24}$$

(Where hh and mm are the hour and minute bit of time data, such as hh: mm)

Take advantage of the standardization of the above principle, the number of login time attribute is converted to numerical data [0,1], the results are shown in Fig.1 and Fig.2 as below.

login time
8:18
8:18
8:19
8:19
8:19
8:21
8:21
8:23
8:23

Fig.1:Login Time Attribute columns

login_t
0.35
0.35
0.35
0.35
0.35
0.35
0.35
0.35
0.35

Fig.2: (0,1) standardized time data

(0,1) standardization methods can also be used in the form: December 1, 1980 18:40:38 time data.

4. Time data generalization - histogram score points fine-tuning

After time data has been converted into easy-to-analyze numerical data, the next step is to do the concept hierarchy on the numerical data, here we use histogram analysis.

Usually the histogram analysis methods is firstly to determine the score points of the data value through the division rules in the histogram construction, and then to use the score point to generate histogram for analysis. But here we firstly determine the score points based on the characteristics of the analysis data itself, the environment and the purpose of the analysis task, then use the histogram analysis methods to fine-tune the score points, so that we can get a reasonable interval division beneficial for data mining. The specific methods and steps are as follows:

(1) According to the specific circumstances of the research questions and analyze the general rules of the relevant attributes, Preliminary division area is [00:00,8:00], (8:00,11:30], (11:30,14:30], (14:30, 18:30], (18:30,24:00], corresponding numerical data area after (0,1) standardization is [0,0.333], (0.333,0.479] (0.479,0.604] (0.604,0.771], (0.771,1]. Attention that the value of regional score points are retained to three decimal places but analysis data values are reserved two decimal places, and the purpose of doing so is to avoid the same data with the score point in using a histogram analysis can not be clearly allocated to one region.

(2) Set 0.333,0.479,0.604,0.771,1 to score points, using Excel the histogram tool of Analysis ToolPak to generate histogram shown in Fig.3.

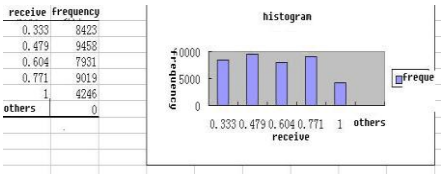

Fig.3:histogram with score point 0.333, 0.479, 0.604, 0.771, 1

(3) Adjust the first score point forwardly and rearwardly for half an hour, i.e. to 0.313,0.479,0.604,0.771,1 and to 0.354,0.479, 0.604,0.771,1 as a new score points respectively generate histogram is shown in Fig.4, Fig.5.

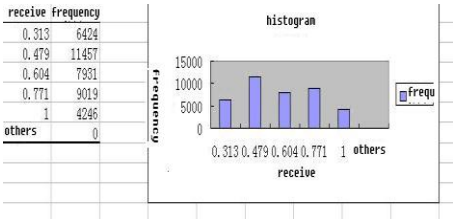

Fig.4: Histogram with first score point adjusted half an hour forwardly

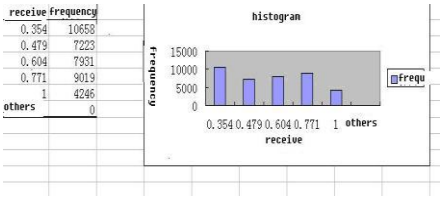

Fig.5:Histogram with first score point adjusted half an hour rearwardly

(4) Adjust the second score points back tone for half an hour to score points 0.333,0.5,0.604,0.771,1 to generate a histogram in Fig.6 below.

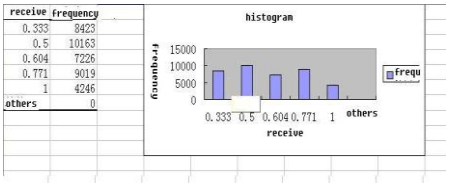

Fig.6:Histogram with second score point set back half an hour forwardly

(5) Adjust the third score points forwardly for half an hour to 0.333,0.479,0.583,0.771,1 to generate a histogram in Fig.7 shows.

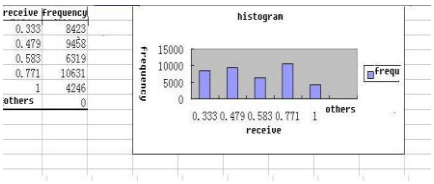

Fig.7:Histogram with third score point adjusted half an hour forwardly

(6) Adjust the fourth score points forwardly for half an hour to score points 0.333,0.479,0.604,0.75,1 to generate a histogram in Fig.8 below.

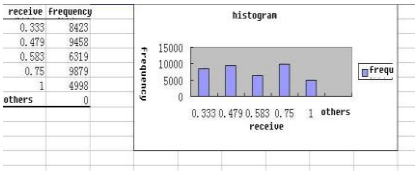

Fig.8:Histogram with fourth score point adjusted half an hour forwardly

(7) Compare histograms from (3) (4) (5) (6) with the one from (2), calculate the respective frequency difference, divide the respective frequency differences by the number of all data to get rate of difference and set a value which we called the biggest rate of difference, If the calculated rate of difference is less than the maximum rate of difference, then the score point is not needed for adjustment, otherwise adjust score points. Continue to adjust until all the rate of difference are all less than the maximum rate of difference, then the final score points is the division region.

(8) Here, we set the maximum rate of difference 0.075. All rate of difference from (6) are all lower than the maximum rate of difference, so finalize the score point as the original score points 0.333, 0.479, 0.604, 0.771, 1.

In summary, the time data can be cleverly converted to numerical data by the (0,1) standardization technology for time data standardization. When the numerical data is applied for the concept hierarchy, we firstly determine the score points based on the characteristics of the analysis data itself, the environment and the purpose of the analysis task, then use the histogram analysis methods to fine-tune the score points, so that we can get a reasonable interval division beneficial for data mining and come up with a score point tuning technology. This provides a lot of convenience to the future the time-based association rule mining analysis.

5. References

[1] (plus) Jiawei Han, Micheline Kamber, Fan, X. Meng, translation, data mining concepts and techniques, Machinery Industry Press,pp.70-91，152-171, 2008.

[2] [U.S.] Pang,-Ning Tan, Michael Steinbach, Vipin Kumar, translated Fanming Fan Hongjian, etc. Introducatipn to Data Mining, ,pp. 2,2006.

[3] Lv Zhifang, association rule mining library data processing applications (thesis), China Ocean University, 2008.