Mathematical classification process modeling for massive data with small differences

GOU Ge

Math and Finance Department of Sichuan university of Arts and Science, Dazhou Sichuan 635000

Keywords: data with small differences; mathematical classification; associated decision tree;

Abstract. Mathematical classification methods of massive data with small differences are studied. During mathematical classification process of data with small differences, assuming the amount of data is too large, the correlation between the data would be reduced, which makes it difficult to perform accurate mathematical classification. In order to avoid these shortcomings, mathematical classification methods of massive data with small differences based on associated decision tree is proposed. Associated decision-making calculation is performed for massive data with small differences to obtain the correlation between all of the data. According to the data relevance, associated decision tree is built to obtain the mathematical classification model for massive data with small differences. Experimental results show that the proposed algorithm utilized for mathematical classification of massive data with small differences, can effectively improve the accuracy of the mathematical classification of the data, so as to meet customer requirements.

1 Introduction

With the rapid development of computer information processing technology, the utilization of mathematical classification methods for data classification has become a major data management method [1]. Mathematical classification method for data is the core of data management technology. With this method, it is possible to perform mathematical classification for a large number of data and provide the basis for data management [2]. The mathematical classification method for data have important value in the field of data management, and attracts many experts' attention [3]. Therefore, the mathematical classification method for data has become a hot issue being researched in the field of data management. At current stage, major mathematical classification methods for data include the mathematical classification method for data based on support vector machine algorithm, the mathematical classification method for data based on K-means clustering algorithm and the mathematical classification method for data based on information gain algorithm [4]. Among them, the most commonly used is the mathematical classification method for data based on support vector machine algorithm. Since the application scope of mathematical classification is more widely, it catch the attention of many scholars, there is much room for development and application [5-6].

2 Principles of mathematical classification method for massive data with small differences

2.1 clustering process for associated data

The collection composed of massive data with small difference is set to be described by $V=\{v_1,v_2,\cdots,v_q\}$, wherein, v_l is used to describe the l-th data in the collection, the attribute corresponding to above data can be described by $K=\{k_1,k_2,\cdots,k_p\}$. Fuzzy clustering method is utilized to classify all data with small differences. The specific method is described below:

The following formula is adopted to describe the fuzzy clustering objectives of data with small difference:

$$L_p(Y,B)=\sum_{k=1}^{q}\sum_{l=1}^{e}y_l^p f_{kl}^2(z_k,b_l)$$

（1）

The state parameters of massive data with small differences can be defined by $e,q,p,d=1$, clustering centers can be described by $B_{(d)}=(b_1,b_2,\cdots,b_e)$, the following formula can be carried out to update all the fuzzy cluster center:

$$y_{kl}=\frac{1}{\sum_{m=1}^{e}\left[\frac{f_{kl}}{f_{km}}\right]^{\frac{2}{p-1}}} \text{ 当} f_{kl}\neq 1$$

$$y_{kl}=0 \text{ 当} e_{kl}=0, l\neq m$$

$$y_{kl}=1 \text{ 当} e_{kl}=0$$

（2）

The following equation is employed to calculate the average of massive data with small differences:

$$b_l=\frac{\sum_{k=1}^{q}y_{kl}}{\sum_{k=1}^{q}y_{lm}}$$

（3）

b_d is compared to $b_{(d+1)}$, assuming that the above data are consistent with the following constraints, it is capable to achieve fuzzy clustering process of massive data with small differences:

$$\left|b_d-b_{(d+1)}\right|\leq\varphi$$

（4）

During the mathematical classification process of massive data with small differences, the range of the objective function in fuzzy clustering is shrinking, which can effectively avoid involved into a local minimum during iterative process, so as to gain cluster structure of massive data with small differences.

2.2 The establishment of associated decision tree for data with small differences

The collection composed of massive data with small differences is denoted as $Z=\{(z_k,a_k)\,|\,k=1,2,\cdots,total\}$, the k -th data in the collection can be described by $z_k=(z_{k1},z_{k2},\cdots,z_{kf})$. According to the following formula to calculate mathematical classification expectations of massive data with small differences:

$$K(q_1,q_2,\cdots,q_p)=-\sum_{l=1}^{p}\frac{q_l}{total}\log_2(\frac{q_l}{total}) \quad (5)$$

The following formula is adopted to calculate information gain ratio of attributes of massive data with small differences:

$$E(C_h)=K(q_1,q_2,\cdots,q_h)-G(C_h)$$

$$u(C_h)=-\sum_{u=1}^{s}\frac{q_u}{total}\times nd\left(\frac{q_u}{total}\right)$$

$$I\_t(C_h)=\frac{I(C_h)}{u(C_h)}$$

(6)

The following formula is utilized to build the mathematical classification decision tree for massive data with small differences:

$$C_k = MIN + \frac{MAX - MIN}{Q} \times k \quad (7)$$

In the above formula, $k = 1, 2, \cdots, Q$.

The data with maximum gain ratio of massive data with small differences is viewed as a branch of the decision tree to build a decision tree for massive data with small differences, so as to realize the mathematical classification for data with small differences.

3 Experimental results and analysis

In order to verify the effectiveness of the proposed method, an experiment is needed to be conducted, the experimental environment is Visual C ++ 6.0. The amount of all data with small differences in the database is set to 100,000. The number of types of data attributes is 290.

10 data with small difference of different attributes are randomly selected from the above data, the details of the selected data are as follows:

Table 1 table of different attribute data

serial number of data attribute	Data attribute	Data size
1	data of steel bar strength	13
2	production data of steel bar	7
3	range of application data of steel bar	5
4	suitable temperature data of steel bar	15
5	usage data per unit area of steel bar	31

The distribution of 5 data attributes in the above table was analyzed, the following figure can be obtained:

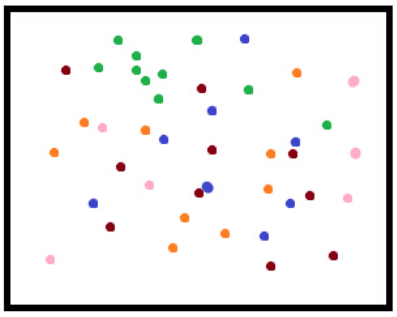

Figure 1 distribution diagram of different attribute data

In the figure, each color represents a data attribute.

When the small mass difference data existed in the database, respectively, different methods were adopted to perform mathematical classification for massive data with small differences, the results obtained can be described by the following table:

Table 2 experimental data table of different attribute data when the amount of data is large

Number	The accuracy of support vector machine algorithm（%）	The accuracy of K-means clustering algorithm（%）	The accuracy of information gain algorithm（%）	The accuracy of the associated decision tree algorithm（%）
1	78	88	89	96
2	77	85	88	98
3	79	87	87	98
4	76	85	88	97
5	77	87	89	99

6	78	86	90	96
7	74	83	88	98
8	77	84	89	98
9	78	88	90	96
10	75	83	88	98

According to the table, it can be learned that in the situation of a large amount of data, with traditional algorithms and proposed algorithm to perform mathematical classification for massive data with small differences respectively, the proposed algorithm can acquire much better accuracy, demonstrating the superiority of the proposed algorithm for mathematic classification of data applied under the situation of large amount of data.

Conclusion

This paper presents a mathematical classification methods of massive data with small differences based on associated decision tree. Associated decision-making calculation is performed for massive data with small differences to obtain the correlation between all of the data. According to the data relevance, associated decision tree is built to obtain the mathematical classification model for massive data with small differences. Experimental results show that the proposed algorithm utilized for mathematical classification of massive data with small differences, can effectively improve the accuracy of the mathematical classification of the data, so as to meet customer requirements.

References

[1] Niu Peng, Wei Wei. Classification of hyperspectral remote sensing images with dynamic support vector machine ensemble [J]. Journal of computer applications, 2010, 30(6):1590-1593.

[2] Ouyang Sen, Song Zhengxiang, Wang Jianhua, Chen Degui, Geng Sanying. A Power Quality Signals De-Noising Algorithm Based on Signals' Multi-Scales Correlation and the Wavelet Transform Theory [J]. Transactions of China electrotechnical society, 2003, 18(3):112-116.

[3] Han Tong. Tracking and managing the number of concurrent users to improve the efficiency of database system [J]. Information Technology & Informatization, 2011.5:49-50.

[4] Jiang Mei. Comparison of SciFinder Scholar database and CA on CD database [J]. Modern information, 2006.3:83-86.

[5] Ke Huaming, Wang Jinliang, Chen Chaozhen. Research of BP Neural Network Classification with Optimization of Genetic Algorithm for Remote Sensing Imagery Based on Principal Component Analysis [J]. Journal of Yichun University, 2010, 32(4):1-4.

[6] Hu Zhaoling, Li Haiquan. Study on the Extraction of Texture Features and Its Application in Classifying SAR Images [J]. Journal of China University of Mining & Technology, 2009, 38(3): 422-427