Research on classification query optimization algorithm in data stream

Zhou hong[1, a], Wang bin[1,b]*(Author for correspondence),Fu chunyan[1,c],Zhi yuan[1,d],and Xue jiamei[1,e]

[1]College of Information Science & Electronic Technology of jiamusi university,

Jiamusi, china

[a]zhouhong0857@163.com,[b]jmsuwang@163.com,[c]jmsfu@126.com,[d]jmszhiyuan@126.com,[e]xuejiameixzy@163.com

Keywords: data stream; classification; query

Abstract. The classification query method of data stream can not only improve the efficiency of data stream query, also achieve data stream query in the best matching state. The difficulty of classification query of data stream difficulty is how to achieve data matching in the optimal matching degree, the traditional classification query method for data stream is the method based on keyword matching, the effect on a single condition is better, but when there are more query conditions, query efficiency is low and matching degree is poor. To this end, a classification query optimization method on data stream is proposed based on improved TFIDF algorithm, the information entropy between data characteristics and the information entropy within characteristic are viewed as weighting factors of data classification query, nonlinear mapping ability of neural network is adopted to realize weight calculation and the fuzzification of TFIDF algorithm, so as to solve classification query problems of data streams. With actual database to process classification query, experimental results show that, the proposed algorithm for classification query on data stream have greatly improved query efficiency, which has good application value.

Introduction

Data stream is a major achievement since the development of computer technology. The data processing technology related to data stream, especially the mature of classification query technology of data stream, which makes the management and query of information possible [1-3]. In the past few decades, the classification query technology on data stream is at the stage of rapid development, different characteristics have been continuously improving [4-6]. However, in the modern mass data information, the classification query on data stream is always a serious bottleneck faced by management and development of modern data [7,8]. The multi polarization of data characteristics, massive amount of data, fuzzification of data information, making the modern classification query on the data flow becomes more and more difficult [9,10].

Related principle of classification query optimization algorithm in data stream

Description of improved TFIDF method. Defined under the given probability distribution $P=(p_1, p_1, ..., p_n)$, information entropy is defined for the data stream transmission as:

$$H(P) = -\sum_{i=1}^{n} p_i \bullet \log_2 p_i \tag{1}$$

If the collection of data in one data stream is D, and classified as k classes according to the type of data, it is denoted as $C_1, C_2, ..., C_k$, the definition containing probability distribution of characteristics t as $P=(n_1/N, n_2/N, ..., n_k/N)$.

In the classification querying process in data streams, higher data distribution uniformity index including a feature, shows that the distribution entropy H_{ac} between each class after classification is larger, namely system contribution is less. Therefore, the improved TFIDF method need to combine distribution entropy between the various classes and individual information entropy within a class together, forming a new influencing parameter, the new information entropy factor between the classes is defined as:

$$\begin{cases} a(H_{ac}) = 1 - \dfrac{H_{ac}}{\max(H_{ac}) + l} \\ \max(H_{ac}) = \log_2 k \end{cases} \tag{2}$$

Among them:

$\max(H_{ac})$ — After feature extraction, characteristic items are correlated, the maximum value of information entropy between each feature class;

k — The number of categories

l — Constant coefficient

Therefore, combining information entropy between classes and within class of characteristics items, TFIDF weighted method is defined as:

$$\begin{cases} W_{ik}(d) = \dfrac{IDF_1}{IDF_{const}} \times a(H_{ac}) \\ IDF_1 = tf_{ik}(d) \times \log(\dfrac{N}{n_k} + 0.01) \\ IDF_{const} = \sqrt{\sum_{i=1}^{n} (tf_{ik}(d))^2 \times [\log(\dfrac{N}{n_k} + 0.01)]^2} \end{cases} \tag{3}$$

By improving definition through the formula, categories contribution of characteristic in data of data stream can be well reflected.

Data weight calculation of data stream based on neural network

BP neural network principle. BP (Back Propagation) neural network is a multilayer feedforward network trained by error back propagation algorithm, and one of the most widely used neural network models. The basic idea is: through training by a large number of data, a set of optimal weights are obtained, afterwards, for a set of specific information, according to the optimal weights which are trained before, the predicted output information can be acquired. BP neural network have good input and output nonlinear mapping relationship, can achieve self learning and updating, which is distinguishable from probabilistic and statistical method. The BP neural network model used in the paper as shown in Figure 1:

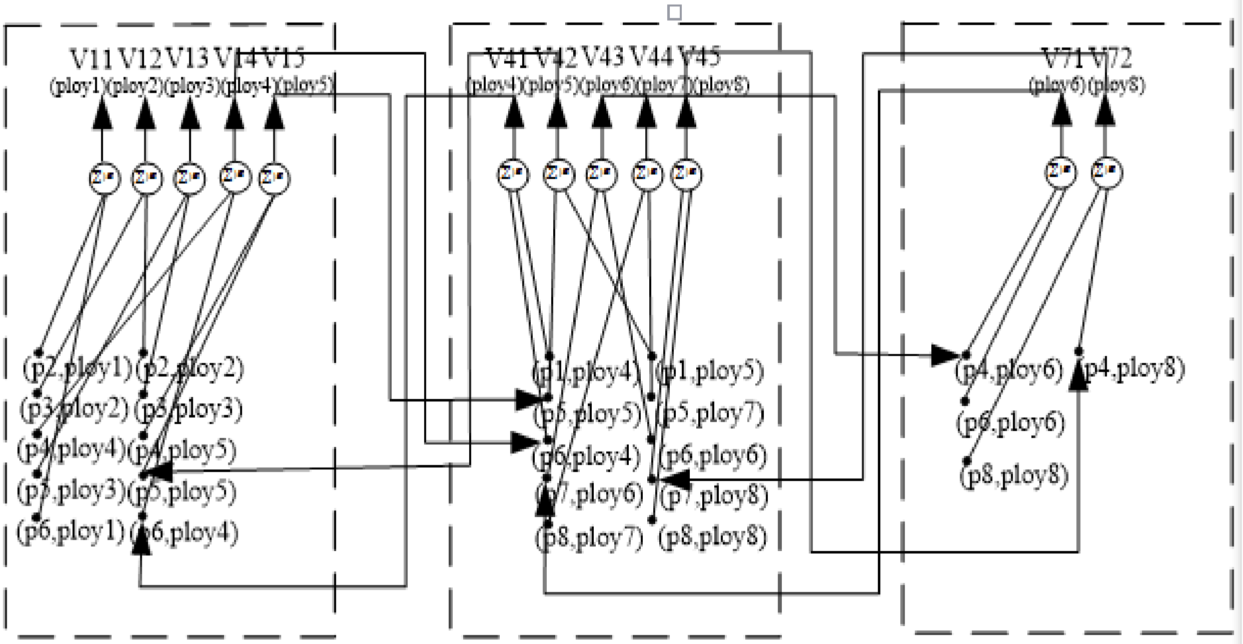

Fig.1 BP neural network model

The relationship between input and output is defined as:

$$y_j = \sum_{i=1}^{n} x_i * a_{ij} + \delta 1 \tag{4}$$

$$z_k = \sum_{j=1}^{m} y_j * b_{jk} + \delta 2 \tag{5}$$

Among them:

$x_i (i = 1, 2, , n)$ — The element of input layer, n is the number of elements in the input layer;

$a_{ij} (i = 1, 2, , n; j = 1, 2, , m)$ — The weighted value from input layer to the hidden layer;

m — The number of elements in hidden layer;

$\delta 1$ — The threshold value from input layer to the hidden layer;

$y_j (j = 1, 2, , m)$ — The element value of hidden layer;

$b_{jk} (j = 1, 2, , m; k = 1, 2, , p)$ — The weighted value from hidden layer to the output layer;

p — The number of elements in hidden layer;

$z_k (k = 1, 2, , p)$ — The value of the output element;

$\delta 2$ — The threshold value from hidden layer to the output layer.

BP neural network used for weight calculation.BP neural network is used for the weight calculation, by inputting a large number of training data of data stream to neural network to obtain the characteristics collection of weight. And then the database of data characteristics in data stream is built, afterwards, through the weighted calculation for new data, results are compared with characteristics database, if they meet the characteristics, appropriate weight will be given.

In this paper, BP neural network is used for the weight calculation, has the following advantages:

(1) the algorithm has strong anti-jamming ability;

(2) the algorithm is capable of adapting to various environments;

(3) does not require complicated statistical model.

Experiments and results analysis

Experimental environment description. In order to test the effect of the proposed algorithm for the classification query in data stream, the data characteristics of 15 dimension actual data stream was adopted for testing.

The data characteristic parameters of data streams used in experiment is shown in table 1.

Table 1 the data characteristics of data stream

Data dimension	Data characteristics	Normalization factor	Data coincidence	The vector degree of the data
1	5	0.351	0.741	0.645
2	7	0.624	0.753	0.123
3	5	0.411	0.789	0.258
4	4	0.641	0.751	0.512
5	6	0.951	0.953	0.751
6	8	0.364	0.285	0.319
7	6	0.842	0.258	0.154
8	2	0.642	0.621	0.153
9	5	0.641	0.423	0.419
10	6	0.157	0.128	0.127
11	4	0.652	0.328	0.137
12	8	0.321	0.285	0.151
13	4	0.512	0.318	0.341
14	6	0.624	0.313	0.788
15	4	0.621	0.131	0.246

Experiment results and analysis. On the basis of data analysis model built in Table 1, through the comparison experiment of classification query over data streams with traditional keyword and classification query over data streams with the method proposed in the paper, the query condition is set to 15, which reflect the 15 dimensions of data in data streams from multiple aspects, namely the classification query situation of most complex data stream.

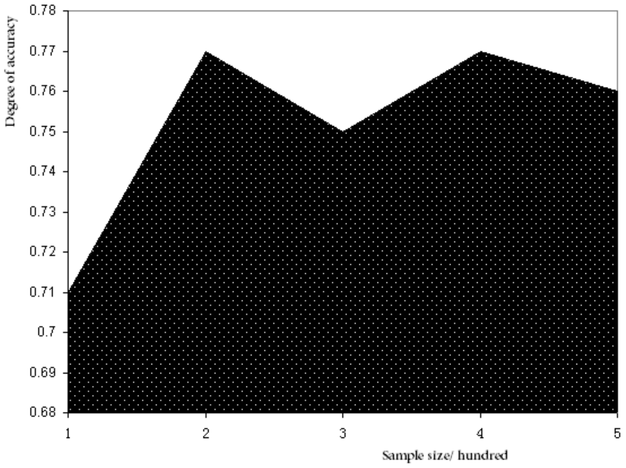

Fig. 2 the query results of traditional algorithm

As can be seen from Figure 2, with the traditional algorithm to process classification query in data stream, when in the face of multidimensional data processes query at the same time, it was unable to extract the data characteristic in data stream of multi-dimension effectively, meanwhile, the data was deeply integrated, so the query results is very poor, among the classification query results, the result which need to be queried may be already included, but because of the limitations of the traditional methods, so the queried data cannot be used, it still require the user to select the needed data from a large number of query results, which did not reach the expected results.

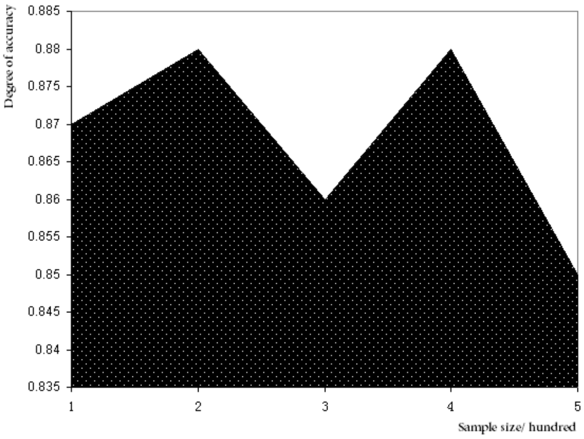

Fig.3 the query results of proposed algorithm

As can be seen from Figure 3, under the classification query condition of the most complex data stream, the classification query method for data stream proposed in the paper can still deeply integrate all query conditions and extract the optimal query condition, and then effective classification query can be conducted for the data of data stream, finally the classification query results also show that the classification results is very good, has the ability for classification query under complex conditions in data stream.

Conclusion

When there are more query conditions, query efficiency of traditional classification query method is low and matching degree is poor. To this end, a classification query optimization method on data stream is proposed based on improved TFIDF algorithm, the information entropy between data characteristics and the information entropy within characteristic are viewed as weighting factors of data classification query, nonlinear mapping ability of neural network is adopted to realize weight calculation and the fuzzification of TFIDF algorithm, so as to solve classification query problems of data streams. With actual database to process classification query, experimental results show that, the proposed algorithm for classification query on data stream have greatly improved query efficiency, which has good application value.

Acknowledgement

This research was supported by the National Science Foundation of Jiamusi University(Grant No. L2012-073）, the National Science Foundation of Jiamusi University(Grant No. Lz2014-005）all support is gratefully acknowledged.

References

[1] Cai Junren, Yu Jianjia. Optimization algorithm for YFilter stream query on DTD data [J]. Computer engineering and design, 2012, 33 (2): 811-814.

[2] Zhang Xiaolin, Cui Min, Tan Yuesheng. Optimization algorithm for LazyDFA XML stream query on DTD data [J]. Computer engineering and applications, 2009, (28): 131-132.

[3] Sun Ling, Zhou Lin. data flow optimization of dynamic grid based on repairing particle swarm optimization algorithm [J]. Laboratory research and exploration, 2011, 30 (6): 208-212.

[4] Zhang Xiaolin, Cui Min, Tan Yuesheng. Optimization algorithm for XPath query in the XML data stream based on LazyDFA [J]. Computer engineering and applications, 2008, 44 (28): 125-127.

[5] Qian Jiangbo, Xu Hongbing, Dong Yisheng et al. optimization algorithm of the data stream window connection based on the minimum spanning tree, [J]. Computer research and development, 2007, 44:1000-1007.

[6] Zuo Liyun, Ma Yingjie. Multi query optimization algorithm based on data stream processing model [J]. Journal of computer engineering and science, 2009, 31 (3): 71-74.

[7] Zhou Hong. Parallel query optimization technique on data stream based on genetic algorithm [J]. Journal of Jiamusi University: Natural Science Edition, 2008, 26 (4): 500-503

[8] Hu Ping. Optimization study on the wear leveling data flow control algorithm of flash based database [J]. Computer software and Application 2014, (10): 98-99.

[9] Xiong Liyan, Zhang Shenghui. The Optimization of the WRR Algorithm in Multi-Class Real-Time Data Scheduling [J]. Computer engineering and science, 34 (7): 35-38

[10] Yu Xingjiang, Tao Yang. Classified Optimization Scheduling Algorithm Based on Multi-QoS Attributes [J]. Computer Engineering, 2009, 35 (5): 31-33.