ATLANTIS PRESS
Advances in Intelligent Systems Research, volume 130
6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016)

Research on Data Mining Framework Based on Improved Sequential Association Rule Discovery

Qing Tan[1, a]

[1]College of Information Technology, Luoyang Normal University, Henan Luoyang, 471934, China

[a]edutanqing@163.com

Keywords: Sequential association rule; Data mining; Apriori algorithm; Clustering; FP tree

Abstract. This paper firstly analyzes the shortcomings of sequential association rule discovery technology, and proposes the improvement method to make up the deficiency. Then, the paper discusses the data mining method based on association rules. The paper presents research on data mining framework based on improved sequential association rule discovery. This novel method can make use of frequent itemsets to generate the required association rules, according to the user set the minimum credibility of the choice, the generation of time sequence association rules.

Introduction

Two different algorithms in data mining, clustering and association rule mining algorithms have strong influence in the technical data, the association rules algorithm of traditional Apriori algorithm due to produce a large number of frequent itemsets set in the calculation, this caused a great impact to the implementation of the algorithm of time and space, and the clustering algorithm can get through clustering frequent itemsets and does not generate a null set to reduce the influence of time and space.

Apriori algorithm is a classical algorithm of association rules in data mining. The algorithm through data mining frequent itemsets Boolean association rules are needed, and the association rules mining algorithm is the core of finding frequent itemsets.

Neural network is a network system which is composed of a large number of simple neurons which are connected by a certain rule. It can simulate the structure and function of the human brain, use some kind of learning algorithm to learn from the training samples, and the knowledge stored in the connection between each unit of the network [1]. Neural networks mainly include forward neural network, backward neural network and self-organizing network. In the field of data mining, it is mainly used to extract the classification rules from the forward neural network. Neural network algorithm in the literature, and then put forward a lot of deformation, including the replacement of the error function, the dynamic adjustment of the network topology, learning rate and the dynamic adjustment of parameters.

Agrawal and other integrated machine learning and database technology, the three types of data mining goals that classification, association and sequence as a unified rule discovery problem to deal with. They give a unified mining model and several basic operations in the rule discovery process, which can solve the problem of how to map the data mining problem into the model and find the rules through the basic operation. The data mining framework based on rule discovery is a common method in data mining research. The paper presents research on data mining framework based on improved sequential association rule discovery.

Analysis of Improved Sequential Association Rule Mining Algorithm

As the name implies, the element that forms a transaction sequence is a transaction. For example, a customer for a certain period of time is in the supermarket to buy goods in the record sequence. Event sequence elements are events. For example, in the wireless communication network in the fault sequence, the user interfaces interaction behavior sequence.

Clustering analysis is a set of physical or abstract data set partitioning process into a number of categories, with high similarity between arbitrary each category after clustering in two data samples with low similarity between different categories of data sample (application) is one of the main data mining technology application, it can be used as an independent tool to use the unknown class label data set is divided into multiple categories, observe the characteristics of data samples in each category, and further analysis of specific categories [2].

Assuming there is no more than two transactions occurring at the same time, the number of purchases of items is not considered in the transaction, and only the item that is concerned about a project is purchased or not. An item set equation (1): i is a non-empty set of items; a is sequence of ordered queues consisting of a number of projects, including a sequence of relationships: if there is an integer:

$$1 \le i_1 < i_2 < \cdots < i_n \le m, \; a_1 \subseteq b_{i1}, a_2 \subseteq b_{i2}, \cdots a_n \subseteq b_{in}, \; \langle a_1 a_2 \cdots a_n \rangle \angle \langle b_1 b_2 \cdots b_n \rangle \tag{1}$$

Classification and prediction: example: will the country according to the classification of climate, according to the gasoline consumption quota will auto classification, said: the decision tree model, classification rules, neural network can be used to predict some unknown or missing numeric values, clustering analysis: example: clustering of WEB log data to find the same user access patterns, outlier analysis shows.

The constructing process of FP tree can be described as: first, the creation of the tree root node, using the "null" mark [3]. Scanning the transaction data set DB, each transaction in the project in accordance with the degree of support in descending order, and creates a branch for each transaction. Generally, when a transaction to consider increasing the branching, counts along each node of the common prefix increased 1, to follow the prefix node created after the project and links. For the convenience of the tree traversal, create a list of frequent items, so that each project through a nod node pointer to its location in the tree [4]. FP tree mining process can be described as: from the beginning of the length frequent project 1, project of its base structure conditions and conditions of the FP tree, and recursively in the tree for mining, as is shown by equation (2).

$$\overline{\bigcup_{j=-\infty}^{\infty} V_j} = L^2(R), \quad \bigcap_{j=-\infty}^{\infty} V_j = \{0\} \tag{2}$$

Data mining and data analysis of the two closely linked relationships with cyclic recursive, data analysis results need further data mining and data mining to guide decision-making, to assess the value of the process also need to adjust the prior constraint again for data analysis.

The mining of association rules is divided into two steps: (1) to find out all frequent itemsets; (2) to generate strong association rules from frequent itemsets. And its overall performance is determined by the first step. In the search for frequent item sets, the most simple and basic algorithm is the Apriori algorithm. It is an original algorithm for mining frequent item sets of R.Agrawal and R.Srikant. The name of the algorithm is based on the fact that the algorithm uses a priori knowledge of the frequent item sets.

Clustering algorithm of traditional K-means algorithm has some limitations in the process of implementation, there are serious deficiencies also is the treatment of isolated points, and in the educational management data is likely there will be a large number of isolated [5]. The clustering algorithm of particle swarm optimization has a good processing ability for multi dimension data, but when it comes to a large number of data, it will cause a lot of space consumption due to the complexity of processing methods and processing steps [6].

The basic properties of the Apriori algorithm is an arbitrary subset of frequent itemsets are frequent. The algorithm uses the iterative method of layer by layer searching for frequent itemsets mining, using K itemsets mining results (k+1) itemsets, the algorithm firstly calculates all items containing only 1 elements set the frequency of the one-dimensional frequent itemsets L1; and then began circulating processing, by the L1 L2 L2 L3 by mining, mining, until no more frequent itemsets.

Circulation process of the algorithm: layer search database to compute support of candidate itemsets, and the minimum support degree were compared, the largest find the itemsets dimension k, as is shown by equation(3) [7].

$$a_3\sum_{i=0}^{n}x_i y_i + a_4\sum_{i=0}^{n}y_i^2 + d_2\sum_{i=0}^{n}y_i - \sum_{i=0}^{n}y_i y_i' = 0$$
$$a_3\sum_{i=0}^{n}x_i + a_4\sum_{i=0}^{n}y_i + Nd_2 - \sum_{i=0}^{n}y_i' = 0 \quad (3)$$

The project set, or item set (Itemset), is a collection of various items. Set I={i1, i2,... , im} is a collection of projects, transaction database D={t1, t2,... Tn} is composed of a series of transactions with a unique identity TID, each transaction Ti (i=1,2,... N) are corresponding to a subset of the I. Let I1 I support project set I1 on D (support) is the percentage of I1 contains the transaction accounted for in D, i.e..

$$\text{support}(I1)=\|\{t\in D|I1\subseteq t\}\|/\|D\|$$

$$Ck'=\{X\cup X'|X,X'\in Lk\text{-}1,|X\cap X'|=k\text{-}2\} \quad (4)$$

Temporal association rules algorithm is a global optimization method to simulate the process of biological evolution, the inferior initial solution through a set of genetic operators (selection, cross - breeding, recombination and mutation, mutation), in the solution space according to certain rules of random iterative search to the optimal solution of the problem of direct. The genetic algorithm in data mining is the main application fields are: (1) it is combined with BP algorithm to train the neural network, and then extract the rules from the network; (2) the design of systems, such as encoding, trust design of distribution function and improvement of genetic algorithm. The problem of genetic algorithm for data mining is: (1) the algorithm is more complex, (2) the premature convergence of convergence to local minimum is not solved.

Data Mining Framework by Improved Sequential Association Rule Discovery

Data mining technology is considered as an optimization process of the problem [8]. Kleinberg et al. established a theoretical system to determine the value of the model in the framework of micro economics [9]. They believe that if a knowledge model is valid for an enterprise, then it is interesting. Interesting pattern discovery is a new optimization problem, which can be based on the basic objective function, to provide a special algorithm perspective on the value of the "data mining", to derive the optimization of enterprise decision-making.

FIS-ES algorithm is the extension of traditional set operations, the maximum frequent itemset extended set operator generation method based on FIS-ES algorithm through from the database to determine whether meet the minimum support requirement of frequent items, and delete the frequent item until a proper subset of read cycle operation records in the database so far. This algorithm can compress the search space and improve the efficiency of data mining by retaining only the maximal frequent item sets, as is shown by equation(5).

$$x^2(t) = \sum_{j=1}^{n+1}c_j^2(t) + 2\sum_{j=1}^{n+1}\sum_{k=1}^{n+1}c_j(t)c_k(t) \quad (5)$$

Data integration processing need to consider the following questions: (1) from a number of data sources in the data table through the same primary key to the natural connection, each table in the primary key to match each other, or cannot connect. (2) redundancy, which is often a problem in data integration, so before the connection to each table in the field of artificial selection, and the use of natural connection, to prevent the redundant field generated. (3) The data value of the conflict

detection, from different data sources may be different attribute values, so to check the data table in connection with the type of field and if there is the same record and other issues.

Evaluation model of data mining is to extract the data model, using intelligent methods: according to some kind of interest measure, recognition knowledge representation really interesting model, knowledge representation and knowledge representation: the use of visualization technology, providing the mined knowledge to the user (the first 4 is the data preprocessing step), according to the pretreatment of original data from multiple business a database or data warehouse, the structure and rules of them may be different, which will lead to the original data is very messy, not available, even in the same database, there may be repeated and incomplete data, in order to make these data can meet the requirements of data mining, improve the efficiency and get clear results, must be preprocessed data.

System Experiments and Analysis

The general sequential Association Apriori-Gen method is mainly aimed at the problem 1. The main characteristics of the Apriori-Gen method is: (two times with Apriori method) 1 with time as the conversion identifies the transaction database D to customer identification number as the SD sequence database, sequence mode for each customer only corresponds to a set of 2 projects that the SD sequence database should be solved by Apriori method K project set. 3 is candidate sequence formed by the K item set. 4 by the candidate sequence application Apriori method for large sequence.

Classification is one of the most important objectives and tasks in data mining. The purpose of classification is to learn a classification model (called classifier), which can map the data items in a database to a given class [10]. To construct a classifier, a training sample data set is needed as the input. Because data mining is the process of mining knowledge from the source data, this kind of knowledge must also come from the source data, should be the source of the data filtering, extraction (sampling), compression, and the concept of extraction, as is shown by equation(6).

$$k_E(x)=\begin{cases}\frac{1}{2}c_d^{-1}(d+2)(1-x^2), & if \quad x<1\\ 0, & if \quad x\geq 1\end{cases} \tag{6}$$

Aiming at the inherent defects above Apriori algorithm, frequent pattern tree (Frequent Pattern tree FPtree) algorithm. This algorithm can't generate candidate itemsets under the condition of mining frequent itemsets. The divide and conquer method, in the completion of the first scan of the database, to provide frequent itemsets compressed into a FPtree database but at the same time, retain information associated with the item set, then FPtree will differentiate into condition database, mining on these conditions database.

Given a transaction database, the GSP algorithm requires multiple times to scan the transaction database, the basic framework of GSP algorithm for mining sequential patterns are as follows: the database of every item in the support of the first pass, which determine the number of data sequences for each item included in the transaction database.

Conclusions

The paper presents research on data mining framework based on improved sequential association rule discovery. K- center point method, its basic strategy is: first, for each cluster randomly select a representative object, the remaining objects according to the distance distribution and its representative objects to a cluster recently, then repeated with non-representative objects instead of representative objects, with improved clustering quality. This method can effectively deal with small data sets. FP-DFS algorithm, based on the business address index table to reduce the transaction Apriori optimization algorithm and other algorithms and Apriori algorithm to improve the search efficiency, but there are also a narrow range of applications and other issues.

References

[1] YiJie Chen, "The Development of the Commodity Flow Analysis System Based On Association Rule Mining", IJACT, Vol. 4, No. 13, pp. 430 ~ 436, 2012.

[2] Yuan Wang, Lan Zheng, "Endocrine Hormones Association Rules Mining Based on Improved Apriori Algorithm", JCIT, Vol. 7, No. 7, pp. 72 ~ 82, 2012.

[3] Jiebing Liu, Baoxiang Liu, Jianming Liu, Huanhuan Chen, "Association Rule Mining Algorithm Based On Fuzzy Association Rules Lattice and Apriori", JCIT, Vol. 8, No. 8, pp. 399 ~ 406, 2013.

[4] Yan Wang, Ying Wu, Weichao He, "Development of Classification Models for Predicting Happiness: A Data Mining Approach", JDCTA, Vol. 10, No. 3, pp. 1 ~ 10, 2016.

[5] Xiaoyan Wan, "Research on Data Mining Technology of Association Rule", JCIT, Vol. 8, No. 6, pp. 628 ~ 635, 2013.

[6] JIANG Yu-ting, Shao Kai, "The study on the Bank Customer Model Based on the Improved Data Mining", AISS, Vol. 5, No. 7, pp. 955 ~ 962, 2013.

[7] ZHANG Changzheng, WANG Shuo, "Application of Data Mining in Urban Traffic Accidents Governance Based on Association Rules", AISS, Vol. 4, No. 19, pp. 169 ~ 176, 2012.

[8] Somboon Anekritmongkol, Kulthon Kasamsan, "The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model", IJACT, Vol. 3, No. 1, pp. 58 ~ 67, 2011.

[9] Noor Diana Ahmad Tarmizi, Farha Jamaluddin, Azuraliza Abu Bakar, Zulaiha Ali Othman, Suhaila Zainudin, Abdul Razak Hamdan, "Malaysia Dengue Outbreak Detection Using Data Mining Models", JNIT, Vol. 4, No. 6, pp. 96 ~ 107, 2013.

[10] Yan Hai, HongLing Han, Guiming Lu, "Data Mining based on Rough Set and Decision Tree Optimization", JDCTA, Vol. 6, No. 12, pp. 480 ~ 489, 2012.