ATLANTIS PRESS

Advances in Computer Science Research, volume 77

8th International Conference on Management and Computer Science (ICMCS 2018)

Optimization Analysis of Improved Association Rules Based on Decision Tree and Data Clustering

Qing Tan[1, a]

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

[a]edutanqing@163.com

Keywords: Association rule; Decision tree; Clustering; Optimization analysis; Data mining

Abstract. In this paper, we first analyze the optimization strategies and classical algorithms of association rules and clustering mining algorithms, from which we can see the basic ideas of these algorithms to solve the problem, and find the shortcomings of the traditional algorithms. Then, this article describes analysis of hierarchical clustering and association rule mining based on dynamic models. The paper presents optimization analysis of improved association rules based on decision tree and data clustering.The experimental results show that the proposed method is very effective.

Introduction

Adaptation of data mining technology is to specific data storage types. Different data storage methods will affect the implementation mechanism of data mining, target location, technical effectiveness and so on. It is not realistic to expect a general application pattern to find effective knowledge in all data storage modes [1]. Therefore, according to the characteristics of different data storage types, targeted research is popular and must be faced for some time in the future.

Based on the analysis of the architecture of the existing data mining system (prototype system), the basic process and functional components of the data mining system are systematically studied. These studies include the basic process of data mining to explore the system should have the main functional components and their interrelation, different source data types of data mining system functional components requirements; Different application targets require the functional components of the data mining system, and the realization mechanism of the main functional components of the data mining system.

A clustering algorithm is using dynamic model in hierarchical clustering. Chameleon is based on the shortcomings of CURE. Cure and its related schemes ignore the information about the aggregation interconnection of objects in two different clusters. ROCK and its related schemes emphasize the interconnection of objects, but ignore the information about the degree of approximation between objects.

C4.5 is a decision tree algorithm, which is an improved decision tree (decision tree is the organization between decision tree nodes like a tree, in fact is an inverted tree) core algorithm ID3, So basically knowing half the decision tree construction method can construct it. The method of constructing decision tree is to select a good feature and split point as the classification condition of the current node.

Because the massive set of things is stored in a large database, the classical FP-Growth algorithm traverses the subtraction schema base twice every time it generates a new FP-Tree. As a result, the system needs to repeatedly apply for local and database server resources to query the same amount of data. On the one hand, it reduces the efficiency of the algorithm, and on the other hand, it creates a high load on the database server [2]. It is not conducive to the normal operation of the database server.

Knowledge discovery from database, data analysis, data fusion and decision support. Primitive data is viewed as a source of knowledge, like mining from ore. Raw data can be structured, such as data in relational databases, or semi-structured, such as text, graphics, image data, or even heterogeneous data distributed over the network. The method of discovering knowledge can be mathematical or non-mathematical, deductive or inductive.

Integrated data warehouse data is the original scattered database data extraction, cleaning on the basis of systematic processing, aggregation and collation, must eliminate the source data inconsistency, to ensure that the information in the data warehouse is consistent global information about the whole enterprise. The data in the operational database is usually updated in real time, and the data changes according to the need. The data of the data warehouse is mainly used for the decision analysis of the unit. The operation of the data involved is mainly the query and loading of the data. Once a certain data is loaded into the data warehouse, it will normally be kept as a data file for a long time. There is almost no modification or deletion, that is, there is usually a large number of query operations and a small number of periodic loading (or refreshing) operations for the data warehouse.

Analysis of Hierarchical Clustering and Association Rule Mining based on Dynamic Models

There is no doubt that the introduction of constraints can accelerate the process of data mining. However, the introduction of constraints must solve the formal representation of constraints, such as the constraints suitable for data mining, and the use of constraints in a specific stage of data mining. This paper probes into the theoretical problems of data mining under the condition of temporal beam mining. In the temporal interval algebra space, two new temporal interval variable operations (temporal intersection T and temporal and UT) are defined. The data mining theory framework based on such temporal constraints.

In a transaction database, the storage space and execution time are better when the item sequence is closely correlated (a large number of item sequences are repeated or partially repeated). Under the moderate transaction database scale, ISS-DM can be the basis of the ideal association rule mining algorithm according to the generally high system configuration at present. The main problem of ISS-DM algorithm is that the ISS may increase linearly with the increase of database capacity. Therefore, the application will be limited [3].

Main idea: firstly, the data objects are clustered into a large number of relatively small subclusters by a graph partitioning algorithm, and then a condensed hierarchical clustering algorithm is used to find the real result clusters by repeatedly merging subclasses [4]. It considers not only the interconnection, but also the similarity between clusters, as is shown by equation (1), where P() is especially the characteristics within the clusters, to determine the most similar subclusters [5].

$$\mathrm{P(St+1S1,\ldots,St)}=\sum_{i=1}^{t}\mathrm{WiPSiSt}\ +1/\sum_{i=1}^{t}\mathrm{Wi} \tag{1}$$

Here, it overcomes the disadvantage of using information gain to select attributes with more values. Pruning during tree construction, I hate nodes with several elements hanging when I construct decision trees. For such nodes, simply do not consider the best, otherwise it is easy to lead to overfitting. For non-discrete data can be processed, this is actually a formula, see where the value of the continuous split. In other words, the continuous data is transformed into discrete values to be processed.

The implementation method is still using FP-Growth algorithm to construct a FP-Tree, but adding a field of: con-countin the node of each item prefix subtree. When mining conditional database Di, the domain records the number of transactions including II in the transaction represented by the path.

Another generalizable knowledge discovery method is Attribute-Oriented reduction method proposed by SimonFraser University in Canada. In this method, SQL language is used to represent data mining queries, to collect relevant data sets in database, and then a series of data generalization techniques are applied to the related data sets, including attribute deletion, concept tree promotion, and so on. Attribute threshold control, counting and other aggregation function propagation.

Reflect historical changes. The operating database (OLTP) is mainly concerned with the current data in a certain period of time, but the data in the data warehouse usually contains the older historical data, so it always includes a time dimension so that the trends and changes can be studied. A data warehouse system typically records information about a unit from a certain point in the past (such as when the data

warehouse system was started) to the present, through which, Can make quantitative analysis and forecast to the development course and future trend of the unit.

In addition, NFWARM algorithm is also a representative algorithm in matrix based association rule algorithm. These algorithms to a large extent solve the shortcomings of the classical algorithms and meet the needs of data development. However, with the development of data technology, the existing algorithms are faced with more data processing problems, and the multidimensional data processing is the most typical one [6].

If St 1 in the actual sequence is the same as the predicted value, it is recorded as normal, otherwise it is abnormal. It is also found that the normal system pattern is used to predict the behavior of the normal system, and the abnormal ratio of the sequence is much smaller than that obtained from the prediction of the abnormal behavior sequence.

The rules that are excavated must reflect the actual situation of the data. Although the rule cannot be 100% applicable, it must be within a certain degree of credibility. Practicality: the mining rules must be simple and usable, and aim at the mining target. There are no 100 rules, of which 50 are not related to business goals, and 30 users can't understand. Novelty: mining association rules can provide new valuable information for users. If they are known in advance by the user, the rules are of no value even in the right way.

Optimization Analysis of Improved Association Rules based on Decision Tree and Data Clustering

In order to solve the difficulty of determining the parameters and Minpts in the DBSCAN algorithm, it does not explicitly produce a data collection cluster, but calculates a cluster analysis method for automatic and interactive clustering analysis [7]. Therefore, the OPTICS method produces a cluster of density based clusters, and its time complexity is the same as that of DBSCAN.

Data mining is an unpredictable work; it is difficult to design all the problems in advance. Therefore, there is a need to constantly validate and modify errors. Even if some knowledge is correct, it may not be of interest to us. The accuracy of mining results depends not only on its trustworthiness, but also on whether it is useful to us. The use of constraints can help us to identify problems and adjust them in time so that all stages of knowledge discovery can evolve in the right direction.

Grid-based clustering method refers to the use of a multi-resolution network data structure. Firstly, the data space is divided into a finite number of cells, as is shown by equation (2), where all the processing of D is based on a single unit [8]. The commonly used methods are statistical information grid method (STINGM), wavelet transform based clustering method (WaveCluster) and clustering high-dimensional space method (CIQUE).

$$D(M_i, M_j) = \beta \cdot \left| a_i - a_j \right| + (1-\beta)\left| \Delta t_i - \Delta t_j \right| \tag{2}$$

To find out all the large itemsets that contain I, first find the node PiS of each item prefix subtree in the frequent item header table along the chain domain of the item I, and then go up along the parent pointer of each Pi to the root node. Make the con-count field of each item prefix subtree node on this path increase Pi.count. The root node does not increase. At the same time, a temporary frequent item header table, lTable. Each table item is composed of three fields: 1: 1 / item-name / 2 / node-link / 3) con-count.

CART is also a decision tree algorithm. In contrast to the conditional implementation of a multivariate classification cart with multiple subtrees under a node, only two subtrees are classified, which is a little easier to implement. Therefore, the decision tree generated by CART algorithm is a simple binary tree. This is very simple, is to see the K person around you (sample) that category of people accounted for more, that more than I am that many. The realization is to calculate the similarity of each training sample and select the similarity Top-K training sample to see which of the K samples more is.

Decision is a tree structure similar to a flowchart. It is similar to the concept of binary decision tree in data structure. Each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a class or class distribution. The top definition of the tree is the root node.

The data warehouse collects information about the subject matter of the entire organization and is therefore enterprise-wide. For data warehouses, constellation patterns are usually used because they can model multiple related topics; data marts are a subset of departments of the data warehouse, which is sector-wide for selected topics [9]. For data marts, star or snowflake patterns are popular because they are suitable for modeling a single topic.

The main idea of LDPIS algorithm is to constrain the itemset of each thing in the multidimensional transaction database by the dimension value in the transaction. In LDPI trees, only the frequent k-itemsets constrained by frequent dimensional predicate sets in the improved LD tree can form multidimensional frequent itemsets together with frequent dimensional predicates, so only the dimensional constraints represented by frequent dimensional predicate sets in LD should be considered. Because the bucket in LDPI tree corresponds to the frequent dimensional predicate set in LD, the candidate itemsets and frequent itemsets under the constraint of the corresponding frequent dimensional predicate sets can be stored in buckets.

System Experiments and Analysis

In order to mine this kind of database better, we study the association rules mining problem under temporal constraints. We begin with algebraic formalization of temporal interval lattice space and define two basic temporal interval operations [10]. Then they are applied to database filtering and temporal interval merging. We do this for two main purposes: one is to reduce the capacity of the data set by filtering the database; The second is that temporal interval fragments which may be generated by filtering are merged into disjoint mining time zone sets through temporal interval merging and each mining time zone is separately generated by memory calculus to generate association rules.

The basic idea of the algorithm is: first, the data space is divided into rectangular elements, corresponding to different levels of resolution, there are different levels of rectangular elements, these units form a hierarchical structure: each unit in the upper layer is divided into multiple lower-level units. The statistical information of the high level unit can be obtained by computing the lower layer unit, and the query of the statistical information is based on the top-down grid-based method. The main advantages of this method are that the grid structure is advantageous to parallel processing and incremental updating, and its computation is independent of query. In addition, its processing efficiency is very high, as is shown by equation (3), where X is computes the statistical information of the unit by scanning the database once.

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} A \\ A\Phi \end{bmatrix} S + N = \overline{A} S + N \tag{3}$$

NB considers that each feature is independent and who is not involved. So a sample (a collection of feature values, such as the word “data structure” and the word “file "), can be multiplied by the probability that all of its occurrence features are in a given category. For example, the probability of occurrence of “data structure” in class 1 is 0.5, “file “appears in class 1 with a probability of 0.3 , it may be considered that it belongs to class 1 with a probability of 0.5 * 0.5 * 0.3, as is shown by equation(4).

$$\begin{aligned} P^{(\beta)}(m|m) &= E\left\{\left[X(m) - \hat{X}^{(\beta)}(m|m)\right]\left[X(m) - \hat{X}^{(\beta)}(m|m)\right]^T\right\} \\ &= W_X^* \overline{P}^{(\beta)}(m|m) W_X \end{aligned} \tag{4}$$

If the con-count field of an item prefix subtree node is added Pi.count, and if no table item in lTable has the same item-name as Pi.item-name, then add a table item to lTable so that item-name is the same as con-count with Pi, At the same time, the node-link points to the address of Pi of the subtree node of the prefix. If a table item exists in lTable and its item-name is the same as Pi.item-name, then simply add

Pi. to the con-count field of the table item. Then, the con-count field of each table item in Table is counted. If its con-count field is larger than the minimum support degree given in advance, the table item will be retained; otherwise, the table item will be deleted.

Compared with the above method, it is a method that does not produce candidate mining frequent item sets. It constructs a highly compressed data structure, FP- growth, and compresses the original object database. It focuses on frequent pattern growth, avoiding high-cost candidates and achieving better efficiency.

The LDPI tree structure is used to count the candidate items in the bucket by the given thing. The bucket contains a set of ordinal numbers, which can be used to find the concrete things in the array of global objects. The LexiItemsTreeBuild algorithm is used to deal with them.

For large transaction databases, filtering databases with constraints is an important way to reduce I / O cost and improve host efficiency. Filtration efficiency can be expressed as Fefff (D-D'O / D). When the user is concerned about mining time zones that are only a small part of the time zones stored in the database, Fefff is bound to be very large, which makes the cost of I / O significantly lower. It is also possible to integrate frequent item sequences by scanning the database to memory once.

Summary

In this paper, the problem of improving the quality of association rules mining is summarized, and the methods of subjective and objective evaluation of association rules mining techniques are pointed out. Then the constrained data mining problem is discussed. On this basis, we formalize temporal interval, temporal constraint data mining space and temporal interval operation.

References

[1] Lu H et al, Effective data mining using neural netwo rks, IEEE Trans. on Knowledge and Data Eng.. 1996, 8(6).

[2] He Zhongsheng, Zhuang Yanbin, A frequent itemset Discovery algorithm based on Apriori & Fp-growth, Computer Technology and Development.2008,18: 45-46.

[3] Venter F et al, Knowledge discovery in databases using lattices, Expert Systems with Applications. 1997, 13(4): 259-264.

[4] Yang Yun, Luo Yanxia, Improved FP-Growth algorithm, Computer Engineering and Design.2010,31,(7),1508.

[5] Agrawal R et al, Parallel mining of association rules: Design, implementation, and experience, IEEE Transactions on knowledge and Data Engineering. 1996,Vol.8, No.6: 962-969.

[6] Li Xusheng, Wang Baobao, an improvement of Apriori algorithm in Mining Association rules, computer Engineering. 2002.7(28): 104-105.

[7] Raghavan P, Information retrieval algorithms: A survey. In Proc. 1997 ACM-SIAM Symp, Discrete Algorithms, New Orleans. USA. 1997: 11-18.

[8] Savasere A, Omiecinski E,Navathe S, An efficient algorithm for mining association rules, In: Proceedings of the 21st International Conference on VLDB.Zurich,1995:432-444.

[9] Han J et al, Discovery of multiple-level association rules from large databases, In Proc. 21st Int. Conf. Very Large DataBases. Zuiich, Swizerland. Sept. 1995: 420-431.

[10] Wu Shaoying, Qiao Mei, Lou Jia, A new algorithm for mining multidimensional association rules, Journal of Tianjin University of Technology. 2008.24(4):78-81.