ATLANTIS PRESS

Advances in Computer Science Research, volume 59

7th International Conference on Education, Management, Computer and Medicine (EMCM 2016)

Construction of Multidimensional Data Knowledge Base by Improved Classification Association Rule Mining Algorithm

Qing Tan[1, a]

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

[a]edutanqing@163.com

Keywords: Multidimensional data; Knowledge base; Association rule; Classification; FP-Growth

Abstract. Firstly, this paper analyzes the advantages and disadvantages of the existing association rules mining algorithm, and gives the method of multidimensional data analysis. In order to remedy the defects, this paper discusses and improves clustering and classification algorithm for constructing intelligent knowledge base. The paper presents construction of multidimensional data knowledge base by improved classification association rule mining algorithm. The effectiveness of the proposed method is analyzed by an example.

Introduction

Artificial intelligence is one of the most controversial areas in computer science, but it still keeps strong life. Machine learning should be fully researched and developed, and data mining technology inherits the thought of machine learning to solve problems [1]. Expert system (Expert System) was once thought of artificial intelligence development toward the practical direction of the most promising technology, but this technology has gradually shown a large investment, strong subjectivity and narrow application range of some fatal weakness. For example, knowledge acquisition is widely regarded as a bottleneck problem in expert system research.

A concept hierarchy defines a mapping sequence, the low level of concept mapping to more general high-level concepts, multidimensional data model (data cube) makes possible the observation data from different angles, and the concept of hierarchical offers from different levels of data observation ability; combined with the characteristics of both. We can define various OLAP operations in multidimensional data model, for users from different angles, different levels of observation data provides flexibility.

In recent years, more and more attention has been paid to the rule extraction from neural networks. This mainly has the following two tendencies: (1) the rule extraction of network structure decomposition; (2) the nonlinear mapping relation extraction rule. The future development of neural network can further reduce the complexity of the algorithm; improve the understanding of the extracted rules and the applicability of the algorithm.

The source data after cleaning and conversion to be suitable for mining data sets, data mining in this form with a fixed data set to complete the knowledge extraction, finally to appropriate knowledge model for further analysis and decision work. From this narrow point of view, we can define data mining is the process of refining knowledge from a particular data set. Data mining as an important step in KDD view, we can easily focus on the focus of research, effective solution to the problem.

Data warehouse is new data processing architecture, it is a theme oriented, integrated, relatively stable, historical data set, which provides the integrated information for enterprise decision support system. To understand the system response problem in enterprise decision analysis, data warehouse can provide faster response rate than traditional transaction database. Two is to solve the decision analysis of the special needs of the data. Decision analysis needs comprehensive and correct integration data, which is not provided directly by traditional transaction database. Three is to solve the decision analysis of the special operation requirements of the data. Decision analysis is a professional user rather than a general salesman, the need for the use of professional analysis tools, the results of the analysis also in a commercial intelligence way performance, which is not provided by the transaction database. The paper

presents construction of multidimensional data knowledge base based on improved classification association rule mining algorithm.

Multidimensional Data Analysis Based on Association Rule Mining Algorithm

Generalized knowledge discovery method and implementation technology there are many, such as data cube, attribute oriented reduction. Data cubes have other aliases, such as "multidimensional database", "implementation view", "OLAP", etc.. The basic idea of this method is to realize the calculation of some commonly used aggregation functions, such as counting, summation, averaging, maximum, etc., and store these implementation views in multidimensional database [2]. Since a lot of aggregation functions often need to repeat the calculation, then store the recomputed results in the multidimensional data cube will ensure a quick response, and can flexibly provide different data view angles and different levels of abstraction.

The advantages of the FP-Growth algorithm: (1) a large database can be effectively compressed into smaller than the original database of high density structure, avoids the repetition of the overhead of scanning the database, (2) the FP-Tree algorithm based on recursive mining take growth strategy pattern, this thesis creatively put forward the method of mining without candidate itemsets in long, frequent itemsets mining efficiency is good.

Apriori algorithm generates a large number of intermediate itemsets. Apriori_gen function is generated by Lk-1 candidate Ck, the resulting Ck by a set of K itemsets. Obviously, the number of candidate K itemsets generated by K increases exponentially. As of 1 frequent itemsets number is 104, the length of the number of candidate itemsets will reach 2 5*107, if you want to generate a longer rule, which requires the number of candidate itemsets generated will be unimaginable, like astronomical figures, as is shown by equation(1).

$$\overline{P}^{(\eta)}(m|m) = \overline{P}^{(\eta)}(m,M) = \prod_{s=1}^{M}\left[I - \overline{K}^{(\eta)}(m,s)\Psi_w(m,s)\right]\overline{P}^{(\eta)}(m,0) \tag{1}$$

The purpose of data reduction is to reduce the size of the data mining, but it will not affect (or basically not affect) the final mining results. The existing data reduction includes: (1) data aggregation; (2) reducing dimension, eliminate redundant attributes through correlation analysis; (3) data compression; (4) a block of data reduction, using clustering or parameter model to replace the original data.

The algorithm can obtain large itemsets by computing the support of each item separately. If the support degree of the project set is greater than the minimum support, then the large itemsets. Keep all large itemsets, exclude all items that do not meet minimum support, as is shown by equation (2).

$$\begin{aligned} f_2^0(m,n) &=< f_1(x,y), \phi(x-2m, y-2n) > \\ f_2^1(m,n) &=< f_1(x,y), \psi^1(x-2m, y-2n) > \\ f_2^2(m,n) &=< f_1(x,y), \psi^2(x-2m, y-2n) > \\ f_2^3(m,n) &=< f_1(x,y), \psi^3(x-2m, y-2n) > \end{aligned} \tag{2}$$

Two commonly used methods of data generalization, attribute deletion and attribute generalization rules 1, attribute deletion: having a large number of different attribute value on the initial working relationship, with the following conditions, should be removed using the property, this property no generalization operators (such as the property does not define the concept hierarchy related) this property is said to high-level concepts with other attributes [3].

FP-DFS algorithm in which the author takes FP-tree as the basic data structure, firstly, and it is a new search strategy and pruning strategy, the D database for the FP-tree compression in memory, and then press the reverse order one by one treatment concentration of the project, and it is at each iteration starts with a project all the frequent itemsets. So as to improve the search efficiency of the algorithm, reduce the memory footprint, the complexity of the algorithm is low.

Rough set method for classification is found to be inaccurate or noise data internal relation for discrete attributes, can also be used for feature reduction and correlation analysis. Rough set reduction and expert system has been used in many applications in feature. The fuzzy set method provides convenience for handling at high abstraction level. In general, it is involving the use of fuzzy logic in the rules of the system: (1) based on the attribute value into the fuzzy value; (2) for a given sample, you can use a number of fuzzy rules; (3) a combination of the above obtained and get a return value system.

Multidimensional databases always provide data views at different levels of abstraction. For example, the data can be stored for a week, but also the formation of monthly data in the end of the month, monthly data and annual data [4]. About multidimensional data model operation, has been well studied, many documents may be associated with data warehouse, as is shown by equation (3). In fact, this kind of model, especially its completeness operation (such as drilling, drilling, etc.) can become the basis of generalized knowledge discovery.

The range of attributes in the specific background knowledge is reasonably segmented to form an alternative discrete value or interval set. For example, sales of age EAGE mentioned above, can be abstracted into {[20, 29], [30, 39], [40, 49], [50, 59]} {or} young, middle-aged and elderly; VALUE can be abstracted into {[01000], [1000, 2000), [20003000), [30004000), [40005000), all low, {or}, high} [5].

$$\hat{X} = \mu_X + C_X H^T (HC_X H^T + C_V)^{-1}(Z - H\mu_X) \tag{3}$$

The common data warehouse data model: (1) star pattern: in this model, the data warehouse includes a large number of redundant data contains and does not contain the center table, the table attached to a group of small, dimension tables around the central fact table shows a ray. Example: sales data warehouse star mode, this mode contains central fact table sales, which contains four dimensions time, item, branch and location. (2) Snowflake pattern: it is a variant of the star pattern, in which some dimensional tables are standardized, thus further decomposing the data into additional tables. Example above and it is only some of the dimensions to extend.

At the end of the first scan, the algorithm knows which items are frequent, that is, the frequent 1 itemsets, and each frequent set of the 1 is formed a frequent sequence of 1. The frequent K sequence set Lk (k+1) can generate candidate sequence set Ck+1 (k+1), the candidate sequence set in each candidate sequence contains the same number of items, and the number of items were higher than the corresponding set of frequent sequences of seed number 1 Lk items [6]. In each candidate (k+1) sequence and the count, when the entire candidate sequences (k+1) have been produced, according to each candidate algorithm (k+1) to determine which candidate sequence counting (k+1) sequence formed frequent (k+1) sequence, and as the next seed set. When the candidate sequence set generated by a seed set Lk is empty, the algorithm ends.

Clustering and Classification Algorithm for Constructing Intelligent Knowledge Base

The multi-dimensional constraint mechanism, the second step two basic steps for association rules in the constraints, there are certain rules of redundant data by calculating the association rules algorithm is obtained, while the removal does not meet the minimum support threshold when the data set has lost rules may, by making a corresponding the multi-dimensional constraint mechanism, the smallest reduction algorithm itself influence on the acquisition of rules.

Split clustering algorithm is to divide the data set into several subsets, that is, given a set of examples x, including N data objects, and to generate a number of K clusters. Commonly used segmentation based clustering methods have K means method and K one center method, CLARA method and CLARANS method, as is shown by equation (4) [7].

$$\mathrm{P(St+1S1,\ldots,St)} = \sum_{i=1}^{t} \mathrm{WiPSiSt} + 1/\sum_{i=1}^{t} \mathrm{Wi} \tag{4}$$

Clustering knowledge reflect the differences between similar things common property characteristic knowledge of different things and characteristics of knowledge type. The most typical classification method is based on decision tree classification. It is a kind of supervised learning method which is constructed from the instance set. The method first forms a decision tree according to the training subset (also called window). If the tree cannot give the correct classification of all objects, select a few exceptions to join the window and repeat the process until the correct decision set is formed. The end result is a tree, the leaf node is the name of the class, the middle node attribute with branches, the branches of a possible value should attribute.

The core of ID3 algorithm is: the choice of attributes in decision tree nodes at all levels, by calculating the information gain to select attributes, so that at each non leaf node when tested, can be tested on the largest category information record [8]. The specific method is: to detect all the attributes, attributes the maximum information gain selection decision tree node, by the properties of the different values of establishing branches, then a subset of a recursive call to each branch of the decision tree method to establish branch nodes, until all the subset contains only one class of data. Finally, a decision tree is obtained, which can be used to classify the new samples.

Classification technology is a kind of guided learning (Supervised Learning), that is, each training sample data objects already has class identification, through learning can form the expression of data objects and class identification between the corresponding knowledge. In this sense, the goal of data mining is to classify the source data according to the sample data and classify the data.

System Experiments and Analysis

FP-Growth algorithm shortcomings and improved methods, the algorithm takes a growth pattern recursive strategy, although the generation of candidate itemsets is avoided. But in the process of mining, if the number of a large aggregate a lot, and obtained by the original FP-Tree database of many branches, and branch length is too long, the algorithm needs to construct a large number of conditional FP-Tree, not only time-consuming and takes up a lot of space, the mining efficiency is not good, and the use of recursive algorithm efficiency low.

The algorithm only considers single dimensional Boolean association rules mining, but in practice, multi-dimensional, quantitative, multi-layer association rules may occur. At this time, the algorithm is no longer applicable, need to be improved, and even need to re design algorithm.

The shortcomings of Apriori algorithm: First: in every step of generating candidate item sets combined cycle produces too much, do not rule out should not participate in the combination of the second elements; each support calculation of itemsets, of all the records in D database were scanned again, if it is a large the database, the scan will greatly increase the overhead of a I/O computer system, as is shown by equation (5).

$$u_i = \frac{1}{\sqrt{\lambda_i}} A v_i \quad i = 1,2,\cdots,r \tag{5}$$

The improved algorithm for the candidate centralized computing projects support more efficient can be calculated by using the subset function. This method requires advance to set the threshold to control the degree of support, and the need to traverse the database repeatedly, so the algorithm complexity is exponentially increasing [9]. Therefore, if there is a n project, then there are 2n possible frequent itemsets, which constitutes a set of I on the possible solution space [10].

On the basis of the original database FP-Tree, using Apriori algorithm mining, mining process does not construct conditional FP-Tree. The mining process is still using the divide and conquer strategy, soon after the compression of the D database into a set of conditions for each condition database, database associated with a frequent item. If there were n a large itemset, the database can be divided into n D Di (i=1 database,... N, and Di), database is a database connection condition item sets Ii. Then using the Apriori algorithm for mining conditions of each database Di, and it is all with Ii as the tail of the large itemsets.

Decision tree method, in many machine learning books or papers can find a detailed introduction of such methods. ID3 algorithm is the most typical decision tree classification algorithm, after the improved algorithm including ID4, ID5, C4.5, C5.0, etc.. These algorithms are studied and developed from the aspect of machine learning, which is difficult to adapt to large training samples. This is the decision tree application to the direction of data mining must face and solve the key issues.

Summary

The paper presents construction of multidimensional data knowledge base based on improved classification association rule mining algorithm. The improved method assumes a model for each cluster to find the best fit of the data to the given model. At present, the research focuses on the use of probabilistic statistical models for conceptual clustering and neural network technology for self-organizing clustering. One of the main problems that it needs to solve is still how to apply to large database clustering applications. Recent research tends to explore the use of a variety of techniques of integrated clustering methods to solve large-scale database or high-dimensional database clustering mining problems.

References

[1] S.Balasubramaniam, V.Kavitha, "A Survey on Data Retrieval Techniques in Cloud Computing", JCIT, Vol. 8, No. 16, pp. 15 ~ 24, 2013.

[2] Pei Yin, Hongwei Wang, Wei Wang, "Extracting Features for Sentiment Classification: in the Perspective of Statistical Natural Language Processing", AISS, Vol. 4, No. 15, pp. 33 ~ 41, 2012.

[3] Hongxin Wan, Yun Peng, "Clustering and Evaluation on Electronic Commerce Customers Based on Fuzzy Set", IJACT, Vol. 5, No. 3, pp. 199 ~ 206, 2013.

[4] Sheng-Chu Su, Chien-Hung Lin, Tzu-Chin Chao, "Data Mining Technique on Bank Service Satisfaction Research", JCIT, Vol. 8, No. 6, pp. 845 ~ 854, 2013.

[5] Srivastava J, Cooley R, Deshpande M, et al. Web usage mining: discovery and application of usage patterns from web data. SIGKDD Explorations, 2000, 1(2):12-23.

[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] Dan Zhang, Xiaoqing Zeng, Hongming Chen, Wei He, "Research on the Evaluation Models of Customer Value of Brokers in the Circumstances of Electronic Commerce with Intuitionistic Fuzzy Information", AISS, Vol. 3, No. 9, pp. 76 ~ 81, 2011.

[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] JIANG Fei, "Research on Association Rule Mining of Adaptive Genetic Simulated Annealing algorithm", JCIT, Vol. 8, No. 5, pp. 876 ~ 883, 2013.

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