Comparative study on the algorithm for mining association rules based on Data Mining

Jia Guo, Jing-yi Ren and Yu-jing Zhang

Department of Information Management and Engineering, Hebei Finance University,Baoding, 071000, China

Keywords: Data mining; Association rule; Algorithm; Frequent set

Abstract. This paper first introduces the classical algorithm -- Apriori algorithm of association rules in data mining. Then from several width, depth, partitioning, sampling, incremental updating and the angles of the association rules mining of classification discussion. Then using literature search and comparative analysis method to the common association rules mining algorithm are reviewed, mainly including FP-growth algorithm, DHP algorithm, Partition algorithm, FUP algorithm, CD algorithm. Development prospect of association rules mining is discussed.

Introduction

Data mining, also called knowledge discovery in database, is from the large, incomplete, noisy, fuzzy and random of large data extracting implicit in the process of which, people previously unknown and potentially valuable information and knowledge. Say simply, data mining is extracted from large amounts of data or "mining" people useful knowledge. In the face of the current "mass data, the status quo of trace information", an important research branch of data mining - association rule, as research data processing and analysis technology of a kind of advanced and intelligent is in the ascendant.

At present, more consistent argument is: Data Mining is from the large, incomplete, noise, fuzzy, random data in which the extraction of implicit, previously unknown to the people, but also is the information and knowledge process is potentially useful. There are many Data and Mining similar terms, such as knowledge discovery, data analysis, data fusion and decision support from the database. Raw data can be structured, such as the relationship between the data in the database can also be semi structured, such as text, graphics, image data, and even heterogeneous data distribution in the network. The method can be found knowledge of mathematics, can also be non mathematics; can be interpreted, can also be summed up. The discovery of knowledge that can be used for information management, query optimization, decision support, process control, but also can be used for self maintenance. Therefore, data mining is a generalized cross discipline, brought together researchers in various disciplines, especially database, artificial intelligence, mathematical statistics, visualization, parallel computation of scholars and engineering technical staff etc.Figure 1 shows the relationship between data mining and other disciplines.

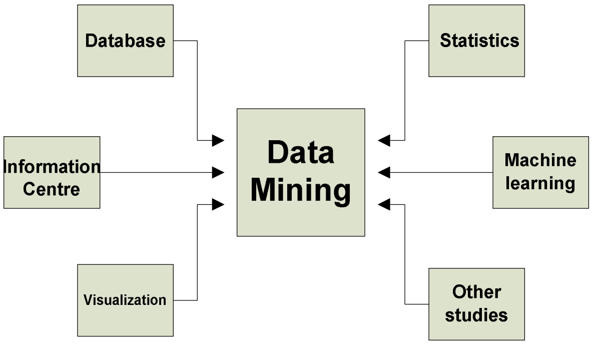

Figure 1.Relational data mining and other disciplines

The types of association rules

According to different standards, association rules can be divided into several types in many different ways, according to the mining model of association rules can be completely put into closed frequent itemsets and frequent itemsets mining completely, maximal frequent itemsets and constrained frequent itemsets. According to the rules of the data related to the layer and dimension can put the association rules mining is divided into single level association rules and multi-level association rules, single dimensional association rules and multidimensional association rules. Depending on the type of value rule processing can be divided into the association rules mining Boolean association rules and quantitative association rules. Based on the mining association rules put rule types can be divided into association rules and association rules. According to the mining model of association rules can be divided into the type of frequent itemsets mining, sequential pattern mining, structure pattern mining etc.. According to the types of constraints that mining association rules can be put into knowledge type constraints, data constraints, dimension constraints / layer, interest restriction, rule constraint.

Association Rule Mining Algorithm

Classic frequent set method

Agrawal was first proposed in 1993 equivalent to the problem of mining association rules between sets of items in the customer transaction database, its core is the frequency method of recursive method based on set theory. Many researchers optimized the original algorithm, such as the introduction of random sampling, the idea of parallel, to improve the efficiency of algorithm for mining rules; put forward various variants, such as association rule generalization and application of the association rules.

The core algorithm

Agrawal design is a basic algorithm in 1993, put forward an important method of association rules mining, which is a method based on the two stage frequency set of ideas, the association rules mining algorithm design is decomposed into two sub problems:

(1) to find all the support is greater than minimum support of itemsets, these itemsets called frequency set.

(2) using the first step to find the frequency set to produce the desired rule.

The second step here is relatively simple point. If a given frequency set $Y = I_1 I_2 \ldots I_k, k \geq 2, I_j \in I$,Produce only contains all the rules set in the $\{I_1, I_2, \ldots, \; I_k\}$ .Each rule is only a right,once these rules are generated, then only those minimum confidence greater than the user specified rules to be left. In order to generate all frequent sets, using a recursive method. Its core idea is:

Firstly generating frequent 1- itemsets L_1, then the frequent 2- itemsets L_2, until there is a r make L_r is empty, then the algorithm stops. Here in the article k cycles, process first to generate candidate itemsets C_k , C_k every itemset in only one out of two different belongs to L_{k-1} frequency set to do a $(k-2)-$ connection to produce. C_k set in the frequency set is used to generate candidate sets, a subset of the final frequency set L_k must be C_k sets. C_k each element required for verification in the transaction database to determine whether the join L_k , the verification process here is a bottleneck of the algorithm performance. This method requires multiple scans may be large transaction database, this will increase a lot of I/O load.

Other frequent set mining method

All of the above is a set of method of Apriori algorithm based on the frequency. Even if is optimized, but the Apriori method is still unable to overcome some of the inherent.

(1) may produce a large number of candidate sets: when the length of 1 frequency set when there is 10000, the length of 2 candidate set number will be more than 10M. When there is if you want to generate a long rule, intermediate elements to produce also huge amount of.

(2) Unable to carry on the analysis on the rare information: because of frequent set using the parameter minsup, so I can not for less than minsup events are analyzed; and if the minsup is set to a very low value, so, the efficiency of the algorithm is a very difficult problem.

Therefore, domestic and foreign researchers through the massive research, in-depth, and recently proposed some new algorithms. In the Apriori algorithm, play a decisive role is the support, but if the confidence placed in the first position, may mining has very high credibility rules. In the [4] literature describes a method to solve the second problems, that is, for each item in the database setting a minimum support MIS. Support so that we can set high to contain only the frequent item rules minimum support using the minimum project, to contain rare project rules setting lower minimum support degree. Method for determining a project support is based on the actual frequency data in the project to assign project MIS value. This kind of multi support mining algorithm has higher efficiency.

Algorithm for mining multilevel association rules

With the development of the data warehouse and OLAP technology, a large amount of data will be processed after integration, pre, and stored in a data warehouse. At present, most of the application of data warehouse are statistics, establish a multi-dimensional and OLAP analysis.

OLAP mining can provide mining in different data sets, based on different details, can be sliced, diced, expansion, filtering and other various rules of the operation, then add some visualization tools, can greatly enhance the ability of data mining and flexibility.

Multilevel association rules: for many applications, due to the scattered distribution of the data, it is difficult to find some strong association rules in the data the most detailed level. When we introduce the concept of level, can dig at a high level on. Although that the higher levels of the rule for a user may be more general information, but for another user is not necessarily the case. So, data mining should provide a mining at multiple levels of function. Multilevel association rules can be divided into the same level association rules and association rules between layers.

The same level association rules can use the following two support strategy:

(1) uniform minimum support: for different levels, all with the same minimum support, so for the user and the algorithm implementation are relatively easy, but the disadvantages are also obvious.

(2) decreasing the minimum support degree: each level has a different minimum support, the lower level of the minimum support degree is relatively small, also can use the upper some filtering by mining the information work.

Inter layer association rules into consideration when minimum support degree, should be based on the lower level of the minimum support to.

Research on the multilevel association rules mining in recent years very warm. In addition to the ML_T2L1 algorithm, such as literature [5] presented a distributed multilevel association rules mining algorithm. The algorithm at each layer using different support degree. Due to the characteristics of distributed system itself, multilayer existing association rule mining algorithms cannot be directly applied to the distributed system. The problem of communication between nodes distributed processing system of the DMARM algorithm using polling method, using a collection of "or" and "and" operation on each node, at the same time find the candidate frequent patterns obtained support mode, reduce the times of scanning database.

Method to measure the value of association rules

When we use the data mining algorithm and obtained some results, how the data mining system knows what the rule is useful for users? What rules are valuable? Need to consider two aspects of system and user.

System

Many algorithms use "support confidence framework", such a structure can sometimes produce some erroneous results. Sometimes a rule's support and confidence than another implication of positive association rules is low, but it may be more accurate. If we take the support and confidence set low enough, then we will get two contradictory rules. On the other hand, if we get those parameters are set high enough, we can only get the imprecise rules.

In short, no one to support and confidence combination can produce association completely correct.

Users

The above discussion is based on the consideration of aspects of the system, and a rule is useful or not should ultimately depends on the user's feeling. Only the user can decide the feasibility and effectiveness of the rules. We should be the actual needs of users and system combination. Can adopt a constraint based mining. Specific binding of the content can be a:

(1)The data constraint: the user can specify constraints on the data mining on which data, but not necessarily all data.

(2) Peacekeeping level mining specified: the user can specify what data mining on the dimension and the dimension of what levels of these.

(3) Rules: you can specify what type of rule is what we need. By introducing the concept of a template, the user and use it to determine what the rules are interesting, but which is otherwise: if a rule matching contains a template, it is interesting, however, if a rule matches a restrictive template, is considered to be a lack of interest.

Among them closely, some conditions can and improve the efficiency of the algorithm, and make more definite purpose of mining.

Conclusion

At present, the method to generate association rules in data mining and its application has been gradually mature, some research results have been integrated in some systems, such as IBM Quest project, Simon Farse DBMiner of the University of etc.. Association rules mining algorithm can be divided into two categories, one is to generate frequent itemsets from candidate algorithm, two is not to have the candidate algorithm. To sum up in the next few years, mining association rules can do further research in the following problems: the combination of OLAP and association rules problem; in the treatment of very large amounts of data, how to improve the efficiency of the algorithm; the results of mining visualization and generation of unstructured data.

Acknowledgements

This research is supported by the Youth Fund Project of Hebei Education Department (QN2015151), the Natural Science Key Project of Hebei Education Department (ZD20131083).

References

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [C]. Proceedings of the ACM SIGMOD conference on management of data, 1993. 207-216.

[2] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation [C]. Proc 2000 ACM-SIGMOD Int Conf Management of Data(SIGMOD’00), Dalas, TX, 2000.

[3]Srikant R,Agrawal R.Mining association rules with item constrains[A].Proc of the 3rd Int'l Conference on Knowledge Discovery in Data Bases and Data Mining[C].Newport Beach,California,August 1997.67～73

[4]Ng R,Lakshmanan L V S,Han J,et al.Exploratory mining and pruning optimizations of constrained associations rules[A].Proceedings of ACM SIGMOD International Conference on Management of Data[C].Seattle,Washington,June 1998.13～24

[5]Fu Y,Han J.Meta-rule-guided mining of association rules in relational databases[A].Proc 1995 Int'l Workshop on Knowledge Discovery and Deductive and Object-Oriented Databases(KDOOD'95) [C].Singapore,December 1995

[6]Park J S,Chen M S,Yu P S.An effective hash-based algorithm for mining association rules[A].Proceedings of ACM SIGMOD International Conference on Management of Data[C].San Jose,CA,May 1995.175～186

[7]J.G.Digalakis, K.G.Margaritis. An experimental study of benchmarking functions for genetic algorithms. Intern. J. Computer Math, vo1.79(4), 2002: 403-416.

[8]D.H.Ackley. A connectionist machine for genetic hillcimbing. Kluwer Academic Publishers, Boston, 1987.