A New Improved Aprior Algorithm in Big Data Environment

Zhou Bei

Sichuan Agricultural University, Ya'an 625014, China

Keywords: Apriori algorithm; genetic algorithm; partitioning technology; fitness function

Abstract: This is an improvement in the deficiency of the Apriori algorithm, and the genetic algorithm and the technique of partitioning, to establish a pattern of association rules, and to increase the efficiency and accuracy of the algorithm, and in the process of improvement, the proposed solutions to the coding and the fitness function are proposed, and by simulation, the improved Aprior algorithm has proven to be better than the performance of the Aprior algorithm on the big data mining.

1. Introduction

With the development of the big data industry, data mining has become a hot topic in the current society [1]. In data mining, correlation algorithm is the key research direction of data mining [2]. The mining of association rules was proposed by scholars such as Agrawal in 1993, and subsequently led to extensive research in the field of database knowledge discovery. In the database, such as X>=, X and both are collections of related items, X represents the previous part of the association rules, and Y represents the association rules. In general, for a study of association rules is a transactional database as the research object, in the actual application, to research and analyze the customer's shopping cart, is a typical example of data mining. In our study of the items in the customer's shopping basket, we identified the purchase behavior of each customer as a transaction, representing the products purchased by the customer in the shopping. In the process, the link between the products in the shopping basket was found to get the user's living habits at shopping. In practical application, data mining is carried out by using association rules, as well as the sequence detection of the catalogue design of goods and so on.

At present, China has made some achievements in the data mining of association rules [3]. The ISS-DM algorithm proposed by MAO guojun and others reduces the pressure of database access. When mining, the algorithm only needs to scan the information once, and can complete [4]. Li xiufei et al. decomposed the project set, and used the support of multiple segments to calculate the support degree of the project set to eliminate the project set with fewer times earlier [5]. Zhu yuquan and zhu shawen et al. analyzed the relationship between association rules and concept lattice [6]. Cao zhang et al. proposed a Separate-M algorithm for incremental mining under the constraint conditions [7].

There is a breakthrough in the research on the mining subject of association rules. As early as 1994, Agrawal et al. improved a algorithm named Apriori on the basis of predecessors. The improved algorithm is still used as a classical algorithm.

2. Overview of Apriori Algorithm

2.1Apriori algorithm

The classical association rule data mining algorithm Apriori algorithm is widely used in various fields, through the data are analyzed and the correlation of mining, excavated the information in the decision-making process is of important reference value.

The Apriori algorithm is widely used in business and is applied in the analysis of consumer market price. It can quickly find out the price relationship between various products and the influence between them. Through the data mining, marketers can target customers, using individual stock prices, the latest information, special campaign or some other special means of information, thus greatly reduce the advertising budget and increase their income. Department stores, supermarkets and some old - sized retail stores are also doing data mining to gauge consumer spending habits over the years.

Apriori algorithm is applied to network security, such as network intrusion detection technology. Early in large computer systems are collecting audit information to establish the tracking file, the purpose of the audit trail is charged for performance testing or more, so less useful information provided by the attack detection. It can discover the abnormal behavior patterns of network users through the learning and training of patterns. The effect degree of Apriori algorithm has weakened the Apriori algorithm of mining results rules, is a network intrusion detection system can quickly find the user's behavior patterns, able to quickly lock the attacker, improve the detection of intrusion detection system based on association rules.

Apriori algorithm is applied in university management. With the increase of the number of impoverished students in colleges and universities, it is more difficult for the school administration to subsidize the work. Aiming at this phenomenon, a method based on data mining algorithm is proposed. Apply the Apriori algorithm of association rules to poor educational system, and the light of the defects in the classical Apriori mining algorithm was improved, the first transaction database mapping to a Boolean matrix, with a kind of layered incremental ideas to dynamically allocate memory for storage, reuse vector "and" operation, finding frequent itemsets. The experimental results show that the improved Apriori algorithm has great improvement in the operating efficiency, excavated rules can effectively assist targeted to carry out the poor student school management department.

Apriori algorithm is widely used in mobile communication. Mobile value-added business has gradually become the most dynamic, most promising and most high-profile business in the mobile communications market. With the recovery of the industry, more and more value-added services have shown strong momentum of development, presenting the characteristics of diversified application, branding, management centralization and deepening cooperation. In view of this trend, Apriori algorithm widely used in association rule data mining is used by many companies. Relying on a telecom operators are value-added services for the construction of Web platform, data warehouse for survey data from mobile value-added business related mining processing, thus obtained about user behavior and indirectly reflect market demand dynamic useful information, the information in the guidance on the business operations of operators and auxiliary business provider's decision-making has very important reference value.

The Apriodri algorithm was first proposed by Rakesh Agrawal and Ramakrishnan Skrikant, with a very wide influence. The algorithm used to search way, step by step a is the main idea is as follows: first, the frequent item 1 - collection, and marked as L1, and then use L1 to find frequent item 2 - a collection of L2 similarly L2 to find L3, cycle to find frequent itemsets. In order to improve the efficiency of the algorithm, the researchers added Apriori properties to the compression of the search space.

The basic idea of the algorithm is that all frequency sets are identified first, and the frequency of these items appears at least as much as the predefined minimum support. A strong association rule is then generated by the frequency set, which must satisfy minimum support and minimum confidence. Then use the step 1 to find the frequency set to produce the desired rule, produce only contains all of the items of the set rules, each rule is only a right, here is the rule in the definition. Once these rules are generated, only the rules that are larger than the user's given minimum credibility are left behind. In order to generate all the frequency sets, a recursive method is used.

```
(1) L1 = find_frequent_1-itemsets(D);
(2) for (k=2; Lk - 1 indicates Φ; K++) {
(3) Ck = apriori_gen(Lk-1,min_sup);
(4) for each transaction t.
(5) Ct = subset (Ck, t); //get the subsets of t that are candidates.
(6) for each candidate c.
(7) Arthur c. ount++;
(8)}
(9) Lk ={c hours Ck|c.count or min_sup}
(10)}
(11) return L= union k Lk;
```

2.2 Description of association rules

At present, the research of associationrule has become an important research direction in data mining. The association rule pattern belongs to the description mode, and the algorithm of discovering association rules is the method of unsupervised learning. Since 1993, when famous scholars such as Agrawal first proposed the mining of association rules, the association rules have become a hot topic in the field of Knowledge Discovery in Database (KDD). In a database, the definition of association rules is an x. ->Y expression (implied) is a form of knowledge representation. Among them, x and Y are the collections of Item (Item), which are called the front and back parts of the rule respectively. The study of association rules starts with transactional databases, and a typical application of association rule mining is basket analysis. In the supermarket shopping blue analysis, each transaction is considered to represent a customer's purchase behavior, while the transaction corresponds to the item that the customer purchases at one time. The process analyzes customers' buying habits by finding a link between the different items in their shopping basket. Association rule mining is also widely used in commodity catalogue design, fire sale analysis, network intrusion detection, biological sequence detection, etc.

Association rules can be understood as A proposition, that is, if A transaction support project set A, it also has the certain possibility support project set B, the credibility of this possibility call this rules, notes for the conf (R) or c (R). a

The credibility of rule R is that the transaction T that supports the project set A also supports the conditional probability of project set B. The degree of support of association rules reflects the frequency of the rule, and its credibility indicates the correctness of the whole rule. An association rule must have sufficient support and credibility. For a given minimum confidence minconf and minimum support minsup. If conf [R] is greater than or equal to minconf, supp(R) is greater than minsupp, then the association rule R refers to the establishment of the database. Rules are called strong rules, and the foregoing and subsequent items of rules must be frequent and necessary conditions for the establishment of an association rule.

An association rule must satisfy both minimum support and minimum confidence conditions. If the credibility of a rule is 100%, it does not describe a common schema in the database, which should also be deleted because it does not meet the constraint of minimum support. Suppose a rule has enough support, but the credibility is low, for example, 2% of customers who buy soap also buy fruit. This fact is supported by most data in the database, but because it does not represent a strong correlation between projects, it cannot exist as an association rule.

2.3 Research status of association rules.

Some progress has been made in mining association rules. MAO guojun et al. proposed the ISS - DM algorithm to further reduce the access to the database[8]. It only needs a scan to complete the mining of association rules. Li xiufei et al. proposed a multi - segment support algorithm, which was used to calculate the support degree of the project set[9], and to eliminate the non-frequent item sets early. Lu jianjiang et al. discussed the domain of the number attribute by means of the normal fuzzy number model, thus generating a series of problems of language value association rules[10]. Yu-quan zhu, shao-wen zhu, chih-peng hsieh has analyzed the relation between concept lattice and association rule extraction m delete. Koyujing et al. studied the increase of association rules in a constrained environment.

Volume mining, the proposed Separate-M algorithm. Foreign countries have made great achievements in mining association rules. Such as 1994. On the basis of previous work, Agrawal has perfected an association rule mining algorithm called Apriori. This algorithm one

As a classic association rule mining algorithm is referenced. But the Apriori algorithm has two fatal performance bottlenecks:

(1) multiple scan transaction database requires a large 1 / o load.

(2) there may be a large selection.

Whether it is to find a complete set of frequent itemsets, or to simplify the set, or the maximum frequent itemset collection, there is an efficiency problem of the algorithm. Many methods have been proposed to speed up the mining of association rules. This includes from reduce the database scan number of Partion algorithm, DIC algorithm, such as speeding the support count of dar chain and vertical format, etc., from DHP were studied.the optimum processing algorithm of pruning of the set of candidate OSSM algorithm and some other methods such as sampling, incremental mining, parallel, etc. These methods greatly improve the efficiency of mining association rules algorithm. A lot of tests have been done on the performance of existing algorithms with real data rather than artificial data, and it finds that the performance of existing algorithms can be improved. The study of association rules include association rule mining under distributed environment and tense related association rule mining, association rule mining based on rough set theory, mining association rules in multiple relationship between soil better results expression (such as visual, protect data hidden in association rules mining in order to protect the privacy of users, provide users with more interaction mechanism and so on

3. Mining Model of Association Rules Based on Genetic Algorithm

Genetic algorithm has been widely used in many areas, it is in the performance of global optimization search makes it have excellent performance in data mining, thus cause the attention of the researchers, related research achievements are constantly emerging. With the increasing of genetic algorithm research, the application in data mining is also increasing. Is chaotic, nonlinear and stochastic genetic algorithm for the solution of the problem showed the characteristics of the emergence of the algorithm for the more complex data mining provides a new solution for the problem of, in front of the huge amount of data, genetic algorithms (ga) is a very good choice. Compared with other algorithms, genetic algorithm (ga) to solve the problem of local optimal solution, it will be natural evolution theory into the global search algorithm, this article will use the algorithm for the discovery of rules. The algorithm of Apriori also has its own limitations. This paper will improve the algorithm and combine the genetic algorithm to establish a mining model of association rules.

3.1 Improved Apriori algorithm.

In the actual application, because of the Apriori algorithm may produce a large number of candidate itemsets, for example in a frequent 1-104 concentration when using this algorithm, the number of candidate 2 - itemsets can reach 107, this is a very hostile place, on the other hand, when executed, Apriori algorithm to repeatedly database operation, when large amount of data, too much to read and write operations will greatly reduce the computing efficiency.

In order to reduce the computer consumption and improve the performance of the algorithm, this paper designs an improved algorithm. The improvement algorithm of this paper is based on partition technology, and the core part of the algorithm is divided into two steps. First, the database D used to extract association rules is divided into non-intersecting databases: D1, D2, D3,... , Dn, Di (I = 1, 2, 3,... , n). To control its size, to make sure every things can into the memory database, database for Di points things, then USES the Apriori algorithm to find the strength set in the database Di Li, then the set of all strengths into a potential strength set in the database D:

$$lp = \cup_{i=1}^{n} l^{i} \quad (1)$$

Secondly, after the initialization of the potential item, the genetic algorithm is used to obtain the strength set of potential itemset lp. In this process, the database only needs to be accessed again, which greatly reduces the pressure of the database and improves the performance of the algorithm.

3.2 The ap-gen model design.

In this section, we introduce how to apply the genetic algorithm in the model:

3.2.1 Encoding strategy.

In order to solve the problem of complex parameter, the countermeasures adopted in this paper are multi-parameter coding technology. That is, encoding each parameter, obtaining a string group, and then combining the string into a whole chromosome. Because each chromosome presents an association rule, this is the goal of data mining. Binary coding is also used.

3.2.2 The design of fitness function.

Support and credibility are two main aspects of the evaluation of association rules. When mining association rules, the biggest problem is how to define the association rules of min sup and min conf. At this point, the fitness function is defined as: $Fit(x) = a * Sup(x) + b * Con(x)$,in this function, the variable x is the rule, a and b are constants, is the support and credibility in the evaluation of the share, and satisfy the 0 < a, b < 1, Sup (x) as the support, Con (x) as the credibility. When the individual can't use a reasonable rules to explain it, we consider the individual is a useless, you can set it Fit = 0, namely fitness is zero, and in the variables a and b values need to be adjusted according to the actual situation of the users themselves.

3.2.3 Design of genetic manipulation.

(1) Operator selection.

In the simple model of genetic algorithm, the mating group selection problem can take the method of gambling. This method can easily lead to the optimal solution appearing in the search process, even if it is improved or maximized by individual competitiveness, the principle of the algorithm cannot be completed. Therefore, by sorting algorithm, the sorting algorithm of individual adaptive value in population is realized by sorting algorithm.

(2) Crossover operator

In the design of crossover operator, the following two problems need to be taken into account. The exchange of genes. Single-point cross crossover operator is one of the most simple pattern, because is a form of binary code, this way of cross feeling for the whole algorithm can don't have much impact, on the performance, superior to multi-point crossover, so this article is using a single point of intersect.

(3) Mutation operator

In this paper, we need to determine the location of the variation point and the replacement of the gene. In this paper, a uniform variation method is adopted, which USES a random number in a range to calculate a small probability that randomly changes the value of the loci. This way doesn't guarantee complete genetic properties, this paper will use a dynamic random probability Pm for gene mutation, determine the variation of the individual, its all loci has, in turn, the value of the change, guarantee after completion of the mutation, gene on the attribute value will be there.

(4) Adaptive Pc,Pm.

In many literatures, the algorithm with fixed when using Pc, Pm, when the their values through the hours, group is affected by the variation factor and crossover factor is very small, lead to a new gene into too difficult; When the value is too large, the original good genes in the population may be destroyed, so the value of Pc and Pm should be dynamically balanced, and the value of small value should be improved, and the value of the large value should be reduced. The following is the formula of Pc,Pm:

$$P_c = \begin{cases} P_c * \dfrac{F_{\max} - X_{\max}}{F_{\max} - F_{avg}}, \text{when} X_{\max} \geq F_{avg} \\ P_C, \text{when } \ X_{\max} \geq F_{avg} \end{cases} \tag{2}$$

$$P_{\mathrm{m}} = \begin{cases} P_m * \dfrac{F_{\max} - Y_{\max}}{F_{\max} - F_{avg}}, \mathrm{when} Y_{\max} \geq F_{avg} \\ P_m, \mathrm{when} Y_{\max} \geq F_{avg} \end{cases} \tag{3}$$

In the above formula, Xmax represents one of the larger values of the two individuals, and Y is the fitness value calculated by the individual after the mutation.

The improved Apriori algorithm first scan transaction libraries build 1 - a transaction set bit string and frequent itemsets bit string, the bit string of 1 - a logic "and" operation, through the statistical results compared with the given support threshold to generate frequent itemsets, bit string of frequent itemsets of the logical "or" operation, the statistical results are produced in the transaction libraries repeat the number of occurrences of candidate itemsets.

The improved Apriori algorithm can be divided into the following steps:

Step 1: Set the support degree and credibility threshold required for mining rules.

Step2 : Scan the transaction libraries, in turn to the transaction in the library items take statistics appear in each transaction, generated the corresponding "bit string," item in the transaction is recorded as "1", does not appear to as "0". Then the bit string of each item statistics can get each candidate 1 - itemsets support count, according to the given support threshold selected is greater than or equal to support threshold candidate 1 - itemsets as frequent 1 - itemsets collection of L1.

Step 3: Sequence S according to L1.

Step 4:The first step is to generate the candidate 2-item set from all the items in L1, and the set is C2.

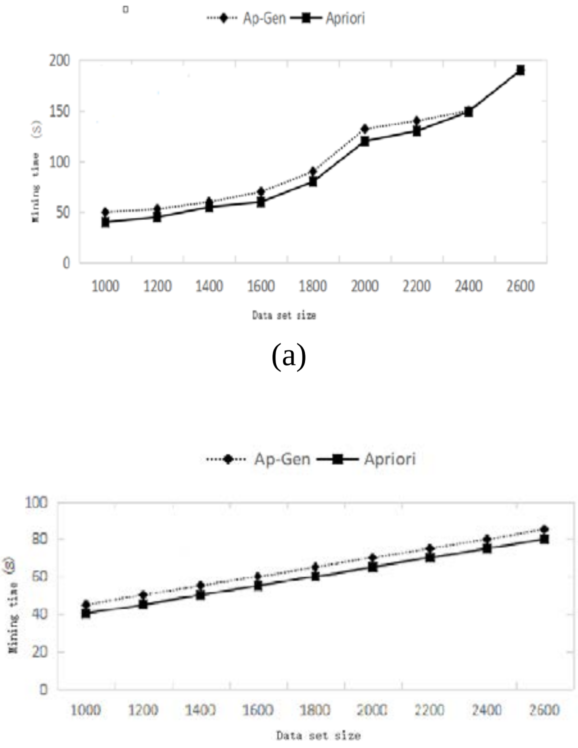
Step 5: Will Ck (k 2) all the items in the bit string for logic "and" operation, to generate the new bit string of the number of "1" in the statistics, statistical result is the new generated after connecting the support counts of candidate itemsets and meet the minimum support threshold of frequent itemsets of candidate item sets, to generate the set of frequent itemsets of Lk. According to the binary code of S to Lk, each item in Lk generates a bit string, forming a bit set containing | Lk | single bit string. A bitstring in a set is performed with two or two logical "or" operations, and the operation results are counted. In the operation results, the number of "1" is k+1 and the number of repeated occurrences is C, and the candidate (k+1) - item set is generated, and the candidate (k+1) is generated according to the sequence S.

Step6:Loop execution, end loop when $C_{k+1} = \varphi$ or $|C_{k+1}| \leq k+1$, end of algorithm.

4. Simulation Test and Result Analysis.

The data source we used was a school's performance database, and we set the computer's memory as 64M to simulate the actual effect.

The model of the experimental simulation platform is Matlab2006a, and the memory space is 64MB to 512MB respectively, and the data set of the test is 1000,1200,1400. , 2600. When the minimum support was 0.1, the ap-gen algorithm was compared with the Apriori algorithm, as shown in figure 1 below. On the other hand, the simulation results of the number of strokes in the data set are shown in FIG. 2 below:



(a)

(b)

Figure 1. Performance comparison between Ap-Gen and Apriori

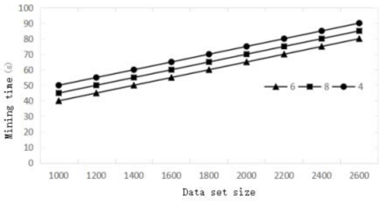


Figure 2. The influence of the number of partitions on the performance of the algorithm

Can be seen from the figure 1, if the operating system's memory to 512 MB, this paper improved the Ap - Gen is higher than that of Apriori algorithm of time, so, in this condition, Ap - Gen to the performance of the Apriori algorithm; When the memory of the running system is 256MB, the ap-gen algorithm is close to the Apriori algorithm in performance, only slightly higher than it, and the larger the data set is, the closer the performance is. When the memory space is 128MB, the larger the data set, the performance of ap-gen will gradually catch up and exceed the Apriori algorithm. When the inner abundant space is 64MB, it can be seen in the figure that the time used by ap-gen is significantly less than the Apriori algorithm.

Can see from the above 2, dataset partition, the arithmetic is consumed less time, in this experiment, after 8 times division, in time on consumption is much less than 4, so when the number of data sets the performance impact is bigger, and the algorithm the time required to did not present a positive correlation with dividing the number of relations. This indicates that APriori requires frequent access to the database while performing the calculation, and the number of candidate sets produced is also huge. Therefore, the efficiency of the APriori algorithm is greatly reduced. Ap-gen reduces the number of accesses to the database and performs the global optimal solution in the set of association rules. As the data continues to grow, the ap-gen algorithm gradually exceeds the Apriori algorithm in performance. Based on the above experiments, the ap-gen algorithm is of great value in large database data mining.

5. Conclusion

In this paper, the shortage of the Apriori algorithm is improved, combined genetic algorithm and classification technology, association rules mining model is set up, improve the efficiency and accuracy of the algorithm, when carries on the improvement, the coding scheme and the fitness function put forward the corresponding solutions, and by way of simulation test to verify the Ap - Gen algorithm in data mining on practicality.

References

[1] Lin K C, Zhang K Y, Huang Y H, et al. Feature selection based on an improved cat swarm optimization algorithm for big data classification[J]. Journal of Supercomputing, 2016, 72(8):3210-3221.

[2] Sun D, Zhang G, Yang S, et al. Re-Stream: Real-time and energy-efficient resource scheduling in big data stream computing environments[J]. Information Sciences, 2015, 319:92-112.

[3] Werdell P J, Bailey S W. An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation[J]. Remote Sensing of Environment, 2005, 98(1):122-140.

[4] Bao C, Hao H, Li Z X, et al. Time-varying system identification using a newly improved HHT algorithm[J]. Computers & Structures, 2009, 87(23-24):1611-1623.

[5] Lin Y, O'Malley D, Vesselinov V V. A Computationally Efficient Parallel Levenberg-Marquardt Algorithm for Large-Scale Big-Data Inversion[J]. Water Resources Research, 2015, 52(9).

[6] Madan N. Improved scheduling algorithm in cloud environment[J]. Annals of the New York Academy of Sciences, 2014, 452:407–408.

[7] Chen X, Huang J. Evolutionarily Stable Spectrum Access[J]. IEEE Transactions on Mobile Computing, 2013, 12(7):1281-1293.

[8] Yang P, Wu W. Efficient Particle Filter Localization Algorithm in Dense Passive RFID Tag Environment[J]. IEEE Transactions on Industrial Electronics, 2014, 61(10):5641-5651.

[9] Sheikh S A, Fana P. New blind equalization techniques based on improved square contour algorithm ☆[J]. Digital Signal Processing, 2008, 18(5):680-693.

[10] Urbanski S P, Salmon J M, Nordgren B L, et al. A MODIS direct broadcast algorithm for mapping wildfire burned area in the western United States.[J]. Remote Sensing of Environment, 2009, 113(11):2511-2526.