International Journal of Database Theory and Application
Vol.6, No.6 (2013), pp.95-104
http://dx.doi.org/10.14257/ijdta.2013.6.6.09

An Algorithm of Association Rules Mining in Large Databases Based on Sampling

Zhi Liu, Tianhong Sun and Guoming Sang

Dalian Maritime University

lzsgmsc@126.com

Abstract

In recent years, the amount of data into a geometric growth puts forward higher requirements on data mining algorithm. In the process of frequent itemsets of traditional Apriori algorithm produced, frequent itemsets' generation and storage are quite a waste of time and space. In this paper, we put forward a new Hash table and use the technology to improve the algorithm and get SamplingHT algorithm, through a lot of contrast experiments showed that the new algorithm enhances performance when frequent itemset is generated, and effectively reduce the database scan times, In order to achieve more optima.

Key words*: *Association rules; SamplingHT; Hash table

1. Introduction

Many companies have accumulated large amounts of data in the day-to-day operations. For example, Communication companies collect lots of users' communication data every day, the large amount of data contains a lot of valuable information. So, Agrwaal first put forward association rules in 1993,and used it for mining valuable data information, after that， association rules mining is quickly gained attention by experts in the field of data mining [1]. With the rapid development of economy and the software and hardware technology, the data increases in a geometric pattern, using traditional Apriori algorithm can not meet modern need, In view of disadvantages of the algorithm. In 1996, Hannu Toivonen put forward Sampling algorithm, greatly improving the efficiency of the algorithm [2].

In 2002, Parthasarathy aimed at the problems from Sampling algorithm of Sampling size support-sensitive, interactive measure of accuracy and equivalence class method are put forward. And he also made a contrast experiment, the effect was very optimistic [3]. The traditional way will produce a lot of useless rules. In 2006, Hong Li and many other people put forward DMCASE algorithm, which combines traditional Sampling algorithm and Eclat algorithm based on constraints and uses Eclat algorithm to improve the disadvantages of Sampling [4].

In 2008, Junping Du and many other people put forward ideas that using the tree structure to store samples and variable minimum support, which decreases with the increase of the variable minimum support with layers in the tree structure, make the loss rate of long item set of the algorithm decreases significantly [5]. In 2012, Xiaoying Xie Ying Zhang and many other people put forward algorithm that combines confusion matrix and Sampling algorithm. According to the characteristics of confusion matrix, they also put forward the sample size calculation model and made a more reasonable sample size [6].

In this paper, in term of disadvantages of the algorithm, we improve making use of the combination Hash table technology [7, 8] and Sampling algorithm. We present a new Hash function and application in this improved algorithm, and also compare with some original algorithms. As a result, it shows that the improved algorithm is an effective algorithm of association rules.

The remainder of this study is organized as follows: Section 2 providing the relative concepts of association rules and the principle of Apriori algorithm and Sampling algorithm. Section 3 describes the SamplingHT algorithm that is Sampling improved algorithm. Section 4 describes the relevant experiments. Finally, summary and conclusion are presented in the last section.

2. Association Rule Data Mining Technology

2.1. Basic concept of Data Mining

Let $I = \{i_1, i_2, \cdots, i_m\}$ be the set of all items, and $T = (t_1, t_2, \cdots, t_n)$ be the set of all transactions, Every transaction t_i is a itemset, and meet $t_i \subseteq I$. An association rule is an implication expression of the form $X \to Y$, $X \subset I$ ， $Y \subset I$,where and X and Y disjoint itemsets, *i.e.*, $X \cap Y = \phi$ .The strength of an association rule can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set T, while confidence determines how frequently items in Y appear in transactions that contain X .The formal definitions of these metrics are[9]

$$\text{Support}, s(X \to Y) = \frac{\sigma(X \cup Y)}{N} \qquad (1)$$

$$\text{Confidence}, c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \qquad (2)$$

2.2. Association Rule Mining Algorithm

The most classic algorithm is Apriori algorithm, it is one of the most commonly used algorithm that generate association rules' frequent itemsets. It has a lot of improved algorithm, Sampling algorithm is one of them, it mainly using the sampling method to rule mining in database.

2.2.1. Apriori Algorithm

Apriori algorithm is the most influential method in the modern by study of association rules, Then according to this algorithm structure to improve the algorithm is emerge in endlessly, At present, the improvement algorithm have AprioriMend, AprioriTid, Sampling, DHP, *etc.*, Apriori algorithm using a method of iteration step by step to complete the work of mining frequent itemsets, and using the Apriori properties to prune, improve the efficiency of the generation of frequent itemsets.

Apriori property: If A is a frequent item set, all of the subset is frequent, as the downward closure property(DCP), If B is a infrequent itemset, all of the superset also is infrequent and vice versa, support of a itemset of below support a subset of it, as the anti-monotonicity of support.

A basic idea of Apriori algorithm is generate candidate itemsets, then scan the database for count to it, thus concluded that candidate itemsets is frequent itemset. The main steps are as follows [10]:

1. Connection step: Connect (k-1)-frequent itemsets generated k-frequent itemset. Allows the connection condition is that two former k-2 items of (k-1)-itemsets are equal and the k-1 item of first (k-1)-itemset is smaller than the k-1 item of second (k-1)-itemset. The purpose of which is to ensure that does not produce duplicate k-itemset.

2. Prune step: The use of Apriori property to k-itemset for pruning

3. Count step: Scan transaction database, the number of occurrences of computing k-items et in transaction database. For a transaction and a candidate, if the transactions include the candidate itemsets, the number of occurrences of the candidate itemsets will add one. According to the given minimum support threshold generate frequent k-itemsets.

2.2.2. Sampling Algorithm

With the development of economy, science and technology, gradually increases in the amount of data in the database, in traditional Apriori algorithm can not meet their needs, so,in 1996,Toivonen put forward Sampling algorithm. The improvement of Apriori algorithm is Sampling algorithm, the main idea is that it using the sampling method to sample from the original database. In order to directly stored in memory, according to the sample database mining frequent itemsets, reduce the mining time. Sampling algorithm is using random sampling method to proceed with sampling, random sampling method has the characteristics of simple and quick.

Sampling algorithm put forward an important technology, as Negative Border. Negative Border's main idea is that according to the relevant model it get a threshold value which is smaller than minimum support. At the time of scanning the database to get frequent itemset $F(i, \min\_fr)$ and support for itemset $NB(i, \text{Negative Border})$ between min_sup and Negative Border. Negative Border is applied in the extraction of sample can effectively avoid to frequent itemset loss [2].

However, because data mining itself is decided to face the huge amounts of data, and dimension is also growing, in many cases, therefore, will produce a large number of candidate itemsets, in particular, 2-itemset, efficiency of mining is seriously affected. so, in this paper, we put forward SamplingHT algorithm what using the Hash table method to solve a large number of candidate 2-itemsets.

3. SamplingHT Algorithm

SamplingHT algorithm use Hash table technology to the traditional Sampling algorithm in the sample. The first time candidate 1-itemset and 2-itemset what after scanning the sample database can be obtained directly. Not only time that scan sample database is reduced, but also effectively reduce a large amount of infrequent candidate 2-itemset's generation, space and time is greatly reduced. After k-itemset also is such.

3.1. The Main Steps of SamplingHT Algorithm

(1) According to the traditional Sampling put forward model what calculate sample size and calculate Negative Border model calculate sample size and Negative Border. Then according to its value sample the original database.

(2) For mining to extract the sample, at the time of frequent 1-itemset is generated by Hash table and 2-itemset also is generated directly.

(3) To the generated candidate 2-itemset, Negative Border pruned it to frequent 2-itemset according to the minimum support.

(4) Later the step that generate frequent k-itemset and above 2 - itemsets generation steps are the same, until unable to produce higher levels of frequent itemsets.

3.2. New Hash Function

The hash function is only and only used for processing 2-itemset in DHP algorithm. We presents a new Hash function what is not limited to 2-itemset and processing from 3-itemset to k-itemset. The following is the new Hash function:

$$HF = (\sum_{i=1}^{L-1} C_T^i + \sum_{i=0}^{L-1} (C_{T-X(i)}^{L-i} - C_{T-X(i+1)+1}^{L-i}) - n_k) \bmod m_k \qquad (3)$$

The terms used in the hash function are explained: T is the number of attributes in the database; L is the length of a calculating subset itemset; $X(i)$ is the location of the I'th item in calculating itemset in universal set what is made up of all attributes. m_k and n_k are the two parameters, m_k is the length of the k-itemset hash table, $m_k = prime(C_T^k)$, $prime(C_T^k)$ is the largest prime number what is less than C_T^k, because prime modulus can decrease the hash conflict, it improving the efficiency of the algorithm; n_k is the sum of number from 1-itemset to (k-1)-itemset.

3.3. SamplingHT Code:

Algorithm 1： SamplingHT algorithm

```
Input：Original database D；Minimum support min_sup；Error border $\varepsilon$；maximum possible values beyond $\varepsilon$ $\delta$ .
Output：Frequent itemset $F(l, \min\_\sup)$；
Method:
1.$|S| = compute(D, \varepsilon, \delta)$;// Calculate sample $S$ 's size;
2.$Negative\_border = compute(\min\_\sup, |S|, \delta)$;// Calculate $Negative\_border$
3.$S = random(D, |S|)$；// Sampling from original database D；
4.for(i = 2 ; $L_{i-1} \neq \phi$ ; i++){
//Scanning sample $S$ obtained $C_{i-1}$ and at the same time using Hash table obtained $C_i$；
5.$C_i =$ $ScanandHT(S)$
6.$L_i = Sampling\_con(C_{i-1}, \min\_\sup, Negative\_border)$；// By $C_{i-1}$'s own connection gets new frequent itemset;
7.    $F(l, \min\_\sup) = F(l, \min\_\sup) \cup L_i$；
}
8.return $F(l, \min\_\sup)$；
```

For example SamplingHT algorithm how generate candidate itemsets with the new Hash table.

As shown in Figure 1, transactional database D contains five trading records, TID include：001, 002, 003, 004, 005.Assume that min_sup = 40%, L_1 is produced in once scan database，L_1：{A}, {B}, {C}, {E}, Hash table that calculate number of occurrences of 2-itemset is built at the same time, according to the Hash table obtained L_2 again. And by this analogy, create a 3-itemset hash table, and according to L_2 and hash table obtained L_3.

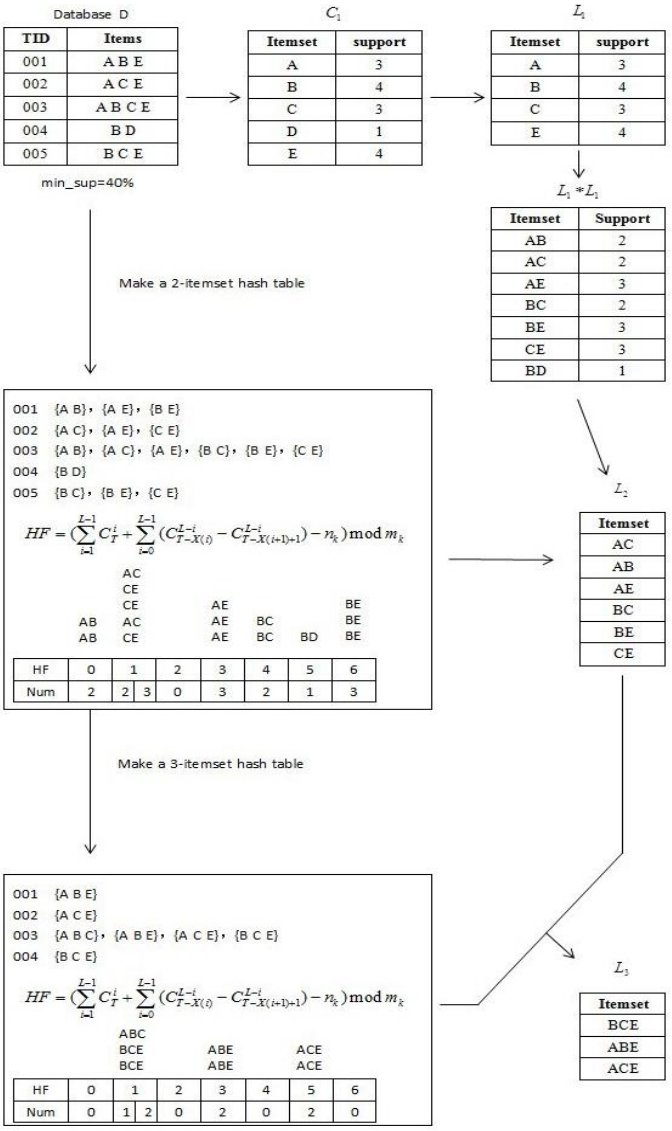
As shown in Table 1, Assume that $I = \{a, b, c\}$ is set. Minimum support min_sup = 5, Negative Border = 4, the following is candidate Itemset:

Table 1. Negative Border

Itemset	Support
a	5
b	6
c	4
a,b	5
a,c	3
b,c	4
a,b,c	3

We can get:

$$F(i, \min\_fr=) \quad \{a\{ \quad \}b, \{ \quad c\},$$
$$NB(i, \text{Negative Border}) = \{\{c\},\{b,c\}\}$$


Database D
min_sup=40%
C_1
L_1
$L_1 * L_1$
Make a 2-itemset hash table
L_2
Make a 3-itemset hash table
L_3
$HF = (\sum_{i=1}^{L-1} C_T^i + \sum_{i=0}^{L-1} (C_{T-X(i)}^{L-i} - C_{T-X(i+1)+1}^{L-i}) - n_k) \bmod m_k$

TID	Items
001	A B E
002	A C E
003	A B C E
004	B D
005	B C E

Itemset	support
A	3
B	4
C	3
D	1
E	4

Itemset	support
A	3
B	4
C	3
E	4

Itemset	Support
AB	2
AC	2
AE	3
BC	2
BE	3
CE	3
BD	1

001 {A B}, {A E}, {B E}
002 {A C}, {A E}, {C E}
003 {A B}, {A C}, {A E}, {B C}, {B E}, {C E}
004 {B D}
005 {B C}, {B E}, {C E}

	AB AB	AC CE CE AC CE		AE AE AE	BC BC	BD	BE BE BE
HF	0	1	2	3	4	5	6
Num	2	2 | 3	0	3	2	1	3

Itemset
AC
AB
AE
BC
BE
CE

001 {A B E}
002 {A C E}
003 {A B C}, {A B E}, {A C E}, {B C E}
004 {B C E}

		ABC BCE BCE		ABE ABE		ACE ACE	
HF	0	1	2	3	4	5	6
Num	0	1 | 2	0	2	0	2	0

Itemset
BCE
ABE
ACE

Figure 1. Frequent Itemset is Generated with Hash Able Technology

Through the above description can be seen that SamplingHT algorithm only need to scan once database can be L_1 , L_2 to L_k , the incidence of infrequent candidate k-itemset is reduced to a large extent. Technology what combine Sample with Hash table is applied in huge amounts of data.at present, data into a geometric growth, Combination of Sample and Hash table effective solving problem, there is reason to believe that this method is effective applied in the future.

4. Experiments and Analysis

This algorithm is realized in Windows 7 system, 4GB memory, 3.20GHz processor PC, written to JAVA language. The data is download from free UCI (data mining laboratory data). The original UCI dataset contains 67557 data records, up to 42 kinds of attributes, Candidate 2-itemset be generated that can't imagine with Sampling algorithm. the parameters of UCI dataset such as Table 2.

Table 2. Parameters of UCI Dataset

Sign	Meaning	Initial size
D	The number of transactions in a original database	67557
min_sp	Minimum support	40%
ε	Error border	0.1
δ	Maximum possible values beyond ε	0.01

4.1. Experiment 1

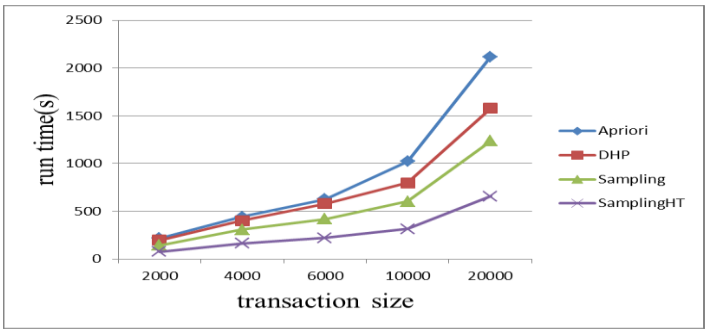


Figure 2. Contrast Test of SamplingHT algorithm's Running Time

Figure 2 is running time of SamplingHT algorithm and other algorithm that compared to experimental results, abscissa represent sample size. Sample size was calculated by different ε and δ , it is based on sample size model. The value of ε and δ as shown in Table 3 ordinate represent algorithm's run time, in seconds. As can be seen from the experiment results SamplingHT algorithm's run time much faster than other algorithm. We also can be inferred from the side frequent itemset be generated that cost a lot of time in multi-dimensional data.

Table 3. Sample Size

ε	δ	$\lvert s \rvert$
0.1	0.01	2000
0.1	0.005	4000
0.1	0.0005	6000
0.01	0.01	10000
0.01	0.005	20000

4.2. Experiment 2

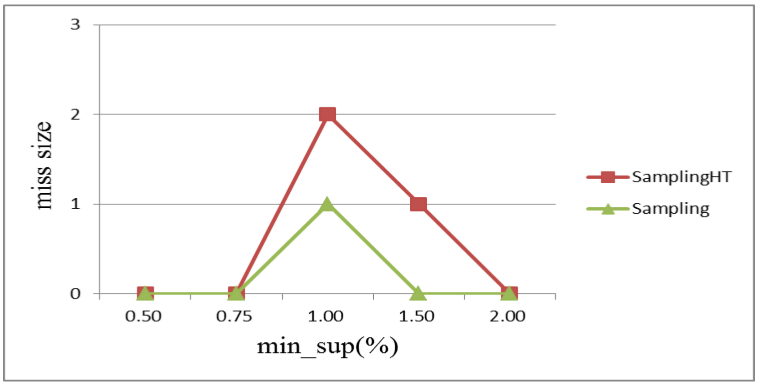


Figure 3. Frequent Itemset's Missing Number Test in Different Algorithm

Figure 3 is frequent itemset's missing number of SamplingHT algorithm and Sampling algorithm what compared to experimental results, abscissa represent minimum support min_sup. ordinate represent frequent itemset's missing number. As can be seen from the experiment results, although missing number of SamplingHT algorithm is better than Sampling algorithm in some minimum support, the result is within the acceptable range.so, the missing number of SamplingHT algorithm's frequent itemset meet the requirements.

4.3. Experiment 3

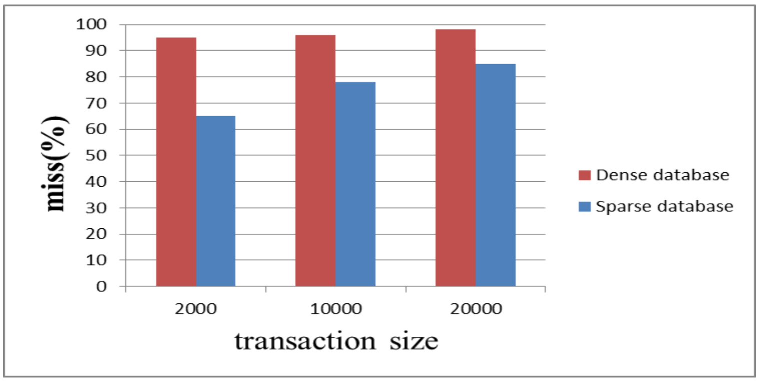


Figure 4. Frequent Itemset's Missing Rate Test in Different Types of Databases

Figure 4 is frequent itemset's missing rate of SamplingHT algorithm what compared to experimental results in different types of databases. Abscissa represents the number of transaction in the database ordinate represent frequent itemset's missing rate. As can be seen from the experiment results, SamplingHT algorithm suitable for use in dense database, and do not suit in sparse database because when high level item set be generated in hash function sparse data generated by the hash function of high level itemset number is often lower than the minimum support. But dense data generated a large number of high level itemset.

5. Conclusion

In this paper, through to the frequent itemsets generated process analysis of SamplingHT, we can see that Hash table technology can be effectively reduced frequent itemset's size especially the frequent 2-itemset's, and the algorithm's running time is reduced. Although algorithms require additional space to store Hash table, the running time can be greatly reduced. With the continuous development of the hardware device, method what using space for time will be more and more widely. Finally, although SamplingHT algorithm is only suitable for dense database, through step by step method to reduce minimum support, we do save missing rate in sparse database.

Acknowledgements

This work was supported by Science and Technology Planning Project of Dalian City, China (No. 2011E15SF100, 2011A17GX073) and the Fundamental Research Funds for the Central Universities (No. 3132013337).

References

[1] G. -J. Mao, L. -J. Duan and S. Wang, "Principle and algorithm of data mining", Tsinghua University Press, Beijing, **(2005)**.

[2] H. Toivonen, "Samplin, Large Databases for Association Rules", VLDB, India, **(1996)** September 3-6.

[3] Srinivasan, "Efficient Progressive Sampling for Association Rules", IEEE International Conference, Calgary, **(2002)** August 19-20.

[4] H. Li, S. -Q. Chen and J. -F. Du, "An Algorithm Research for Distributed Association Rules Mining with Constraints Based on Sampling", ICCI, Beijing, **(2006)** July 17-19.

[5] D. Junping, Z. Min and T. Xuyana, "The realization of distributed sampling association rule mining algorithm in tourism", Intelligent Control and Automation, Chongqing, **(2008)** June 25-27.

[6] X. Xie, Y. Zhang and Y. Xu, "Sampling learning based association rules mining algorithm", ICACI, Nanjing, **(2012)** October 18-20.

[7] J. Soo Park, M.-S. Chen and P. S. Yu, "An effective hash-based algorithm for mining association rules", ACM SIGMOD international conference, San Jose, **(1995)** May 22-25.

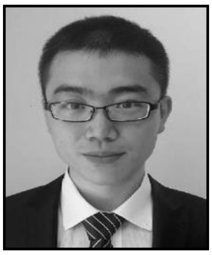
[8] C. -M. Wu and Y. -F. Huang, "Generalized association rule mining using an efficient data structure", Expert Systems with Applications, vol. 38, no. 6, **(2010)**.

[9] P. -N. Tan and M. Steinbach, "Introduction to Data Mining, Posts&Telecom Press", Beijing, **(2011)**.

[10] J. Huang and Z. B. Yin, "Improvement of Apriori Algorithm for Mining Association Rules", Journal of UEST of China, vol. 1, no. 32, **(2003)**.

Authors

Zhi Liu received the M. Sc. degree from Dalian University of Technology, PRC in 1999. She received the Ph.D. from Dalian Maritime University in 2006. She is now an associate professor of Dalian Maritime University, PRC. Her research interests include data mining and artificial intelligence.
E-mail: lzsgmsc@126.com



TianHong Sun received the Bachelor of Science degree in Computer science and technology from Heihe College, Heihe, PRC in 2011.He is now a master student at the Dalian Maritime University. His research includes Association rule in data mining.
E-mail: yanzhiren888@163.com

Guoming Sang received the M.Sc. degree in the major of Computer Application from Dalian University of Technology, PRC in 1999. He is now an associate professor at the school of Information Science and Technology of Dalian Maritime University. His research interests include wireless sensor networks, artificial intelligent and data mining.
E-mail: sangguoming@dlmu.edu.cn