2009 International Forum on Computer Science-Technology and Applications

Mining Association Rules Based on Apriori Algorithm and Application

WANG Pei-ji[1],SHI Lin[1],BAI Jin-niu[2], ZHAO Yu-lin[3]
[1] School of mathematics ,physics and biological engineering, Inner Mongolia University of Science and Technology, Baotou014010, China
[2] School of medicine, Inner Mongolia University of Science and Technology, Baotou014010, China
[3] Branch 2,Inner Mongdia No.1 Machinery,Baotou014010, China
wpjbt@126.com

***ABSTRACT**: In the data mining research,mining association rules is an important topic.Apriori algorithm submitted by Agrawal and R.Srikant in 1994 is the most effective algorithm. Aimed at two problems of discovering frequent itemsets in a large database and mining association rules from frequent itemsets , I make some research on mining frequent itemsets algorithm based on apriori algorithm and mining association rules algorithm based on improved measure system.Mining association rules algorithm based on support,confidence and interestingness is improved,aiming at creating interestingness useless rules and losing useful rules. Useless rules are cancelled,creating more reasonable association rules including negative items.The above method is used to mine association rules to the 2002 student score list of computer specialized field in Inner Mongolia university of science and technology.*

***KEYWORDS:**apriori algorithm;recognizable matrix; association rules mining; application*

I. INTRODUCTION

Apriori algorithm[1] [2]submitted by Agrawal and R.Srikant in 1994 is the most effective algorithm of mining association rules.Two problems of discovering frequent itemsets in a large database and mining association rules from frequent itemsets need to be solved. Technique based hash[3],partition[4],sampling[5] is putted forward by J.S.Park, A.Savasere,H.Toivonen,Jiawei Han[6].They mainly research on problem of decreasing I/O operation and reducing amount of candidate sets[7]. In this paper, I make some research on improved algorithm based on recognizable matrix and improved algorithm for mining association rules. The above schema is used to mine the students' data table.

II. SCHEMA OF MINING ASSOCIATION RULES

Information System (S) is a structure {U,I,F} where U is a object set { $X_1,X_2, \ldots X_P$ },I is a property set {$I_1,I_2,\ldots I_m$} ,f_i is a injection on U and I_i,V_i is Value field of I_i,written as f_i:U × I_i → V_i,i=1,2,…,m,F is a injection on U and I,V=$\bigcup_{i=1}^{m} V_i$ is Value field of I,written as F:U×I → V. If for a= I_i ∈ I,we have F(X_i,a)=f_i (X_i,I_i)∈ V.

In the Information System S={U,I,F},if $\varphi_i : V_i \to \{0,1\}$ $\varphi_i(x)=1$ $x \in v_i^1$ $\varphi_i(x)=0$ $x \in v_i^2$, $v_i^1 \cup v_i^2 = V_i$

$\varphi : V \to \{0,1\}$, $x \in V_i$ $\varphi(x)=\varphi_i(x)$,matrix D_{pxm} is obtained from φ and V.Matrix D_{pxm} is called recognizable matrix of I.

In the Information System S={U,I,F},if U={ $T_1,T_2, \ldots T_P$ } and I= {$I_1,I_2,\ldots I_m$} ,$D_{p\times m}$=（D_1、D_2、⋯D_m）.If a=I_j , $V_a=V_i \subset V$ and $D_j=\varphi_j(V_j)$,j=1,2, …,m.

$$D_j=\begin{pmatrix} d_{1j} \\ d_{2j} \\ \cdots \\ d_{pj} \end{pmatrix}$$

is called recognizable vector of I_j.

$$\text{support-count}（I_j）=\sum_{i=1}^{p} d_{ij}$$

$$D_{p\times m}=(D_1\ D_2 \cdots D_m)=\begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots \\ d_{p1} & d_{p2} & \cdots & d_{pm} \end{pmatrix}$$

2- itemsets {I_i, I_j} written as R_{ij},

$$D_{ij}=D_i \wedge D_j=\begin{pmatrix} d_{1j} \\ d_{2j} \\ \cdots \\ d_{pj} \end{pmatrix} \wedge \begin{pmatrix} d_{1j} \\ d_{2j} \\ \cdots \\ d_{pj} \end{pmatrix} = \begin{pmatrix} d_{1j} \\ d_{2j} \\ \cdots \\ d_{pj} \end{pmatrix}$$

is called recognizable vector of R_{ij} ,$d_{kij}=d_{ki} \wedge d_{kj}$

$$\text{support-count}（R_{ij}）=\sum_{k=1}^{p} d_{kij}$$

k- itemsets {$I_1,I_2,I_3,\ldots I_k$} written as $R_{12\ldots K}$

$D_{12\ldots K}=D_1 \wedge D_2 \wedge \cdots \wedge D_k=(D_1 \wedge D_2 \wedge \cdots \wedge D_{k-1}) \wedge D_k$ is called recognizable vector of $R_{12\ldots K}$.

The set consisting of suffix of each item in itemset is called suffix set of itemsets. The set consisting of suffix of each item in infrequent itemset is called infrequent itemset suffix set L' .L'（w）= {y | y∈L'且 y⊂w} is called lower approximate set of suffix w.

In the transactional databases D, itemset I= {$i_1,i_2, \ldots i_n$} is called positive word set. Inverse of element in positive word set is called negative word.If A(⊂I) is a positive word set and B is a mixed set of positive word and negative word, implication pattern A ⇒ B is called negative association rules, also called association rules of D.[7]

A ⇒ B is association rules of D. Probability of including A and B in transaction of D is called support of A ⇒ B. Probability of including B in transaction of including A is

DOI 10.1109/IFCSTA.2009.41

called confidence of $A \Rightarrow B$. Probability of including B in transaction of D is called expectedconfidence of $A \Rightarrow B$.

$$i = \frac{\text{confidence } (A \Rightarrow B)}{\text{expectedco nfidence}(A \Rightarrow B)} = \frac{\sup(\{A,B\})}{\sup(A)\times\sup(B)}$$

is called interestingness of $A \Rightarrow B$.

If I= $\{ i_1, i_2, \ldots i_n \}$ is positive word set and S($\subset$I) is a itemset, $S_N = \{ i_j'; i_j' \in S$ or $\bar{i}_j' \in S, 1 \leqslant j \leqslant m \}$ is called a inverse itemset of S.[8]

III. DISCOVERING FREQUENT ITEMSETS L IN INFORMATION SYSTEM S

Improved algorithm based on recognizable matrix:

Step1 Input information system S and minsupport

Step2 Recognizable matrix D is obtained from S

Step3 L_1 is obtained from D and let k be 2

Step4 Set R of k- itemsets is obtained from k-1- itemsets and suffix set W of R is also obtained.

Step5 Computing lower approximate set of each suffix w_i in W. If there exist suffix $w_i \in W$ such that $L'(w_i) \neq \Phi$,let W be W-$\{w_i\}$ and L' be $L' \cup \{w_i\}$.

Step6 Obtaining recognizable vector of itemsets in W and computing Support-count (R_{wi}).If Support-count (R_{wi}) < minsupport$\times \left|D\right|$, let W be W-$\{w_i\}$ and L' be $L' \cup \{w_i\}$.

Step7 There are two cases:

Case1 If $W \neq \Phi$,we obtain frequent k-itemsets from W and let k be k+1.Goto step4.

Case2 If $W = \Phi$,we go to step 8

Step8 Output frequent itemsets $L = \bigcup_k L_k$

IV. MINING NEGATIVE ASSOCIATION RULES FROM FREQUENT ITEMSETS L

There exist two cases to mining association rules from frequent itemsets. Case1: we obtain association rules such that support<minsupport and confidence< minconfidence. Case2: If interestingness<1,we obtain negative association rules such that support<minsupport, confidence< minconfidence and interestingness< mininterestingness.

Algorithm based on improved measure system:

For each frequent k-itemset $l_k \in L_k, k \geqslant 2$

$\{H_1=\{i|i\in l_k\}$;genrules（l_k,H_1）; }

$R = \bigcup_k R_k$

genrules（l_k,H_m）

{m<=k-2

{ H_{m+1}=Apriori-gen（H_m）;$h_{m+1} \in H_{m+1}$

{Conf=sup（l_k）/sup （l_k- h_{m+1}）;Ec= sup（h_{m+1}）;Int=Conf/Ec;

If Int>min-int conf⩾min-conf;

$R_K = R_K \cup \{l_k - h_{m+1} \Rightarrow h_{m+1}$ With sup,conf,int}

If Int<1;

For each negative itemset h_{m+1}' of h_{m+1}

{ S=sup（l_k- $h_{m+1} \cup h_{m+1}'$）;C=sup（l_k- $h_{m+1} \cup h_{m+1}'$）/ sup（l_k- h_{m+1}）;E=sup（h_{m+1}'）;I=C/E

If S⩾min-sup C⩾min-conf I⩾min-int;

$R_K = R_K \cup \{l_k - h_{m+1} \Rightarrow h_{m+1}'$ With S,C,I}

}

}

genrules（l_k,H_{m+1}）;

}

}

If B only include one item, we have simple algorithm because 1 is between int(l_k- i $\Rightarrow$ i) and int(l_k- i $\Rightarrow$ i').

For each frequent k-itemset $l_k \in L_k, k \geqslant 2$

{ $H_1=\{i|i\in l_k\}$;$i \in H_1$

{Conf=sup（l_k）/sup（l_k-i）;Ec= sup（i）;Int=Conf/Ec;

If Int>=min-int conf⩾min-conf;

$R_K = R_K \cup \{l_k - i \Rightarrow i$ With sup、conf、int}

else

S=sup（l_k- $i \cup i'$）;C=sup（l_k- $i \cup i'$）/ sup（l_k- i）;E=sup（i'）;I=C/E

If S⩾min-sup C⩾min-conf I⩾min-int;

$R_K = R_K \cup \{ l_k - i \Rightarrow i'$ With S,C,I}

}

}

V. MINING ASSOCIATION RULES FROM THE STUDENTS' DATA TABLE

A. Interrelated data

A transactional database is 2002-student score list of computer specialized field in NeiMongol university of science and technology where a student's score is a record, score of a course is student's property. The sum of record is sum of students (98), the sum of property is sum of courses (9).

B. Data preprocess

Data transform:1) Course coding: Let names of courses be i_1, i_2, …, i_9 ;2) Score separate: Let score be rate.

Input: Student score list

Output: Recognizable matrix

C. Discovering frequent 1-itemsets L_1

Input: Recognizable matrix and minsupport 0.25

Output: L_1={ $i_1,i_3,i_4,i_5,i_6,i_7,i_8$}

D. Discovering frequent item sets L

Input:L_1

Output: $L = \bigcup_{k=1}^{4} L_k$ L_1={ $i_1,i_3,i_4,i_5,i_6,i_7,i_8$},$L_2$= {$i_{14},i_{16},i_{17},i_{34},i_{45},i_{46},i_{47},i_{48},i_{56},i_{67}$},$L_3$={ $i_{146},i_{147},i_{167},i_{456},i_{467}$},$L_4$={ i_{1467}}

E. Mining association rules set R

Input:L,minconfidence 0.75 and mininterestingness 1.6

Output:

R_1: $i_1 \Rightarrow i_{46}$ (s=0.3061,c=0.7895,i=1.6118)

R_2: $i_7 \Rightarrow i_{14}$ (s=0.2959,c=0.8788,i=2.6097)

R_3: $i_1 \Rightarrow i_{47}$ (s=0.2959,c=0.7632,i=2.3372)

R_4: $i_7 \Rightarrow i_{16}$ (s=0.2653,c=0.7879,i=2.4129)

R_5: $i_7 \Rightarrow i_{46}$ (s=0.2959,c=0.8788,i=1.7942)
R_6: $i_{67} \Rightarrow i_{14}$ (s=0.2653,c=0.8966,i=2.6625)
R_7: $i_{47} \Rightarrow i_{16}$ (s=0.2653,c=0.8125,i=2.4883)
R_8: $i_{17} \Rightarrow i_{46}$ (s=0.2653,c=0.8966,i=1.8305)
R_9: $i_{16} \Rightarrow i_{47}$ (s=0.2653,c=0.8125,i=2.4883)
R_{10}: $i_{14} \Rightarrow i_{67}$ (s=0.2653,c=0.7879,i=2.6625)
R_{11}: $i_7 \Rightarrow i_{146}$ (s=0.2653,c=0.7879,i=2.5737)

F. Explaining rules

R_1: score of Analog Electronics Technique is more than 75 $\Rightarrow$ score of Assembly Language is more than 75,score of operating system is more than 75 (minsupport =0.3061,minconfidence=0.7895,mininterestingness=1.618).

G. Application

R_1:A.Minsupport that scores of Analog Electronics Technique, Assembly Language, operating system are fine is 25%,it illustrate that strong influence exist between the three courses and the order is rationale. B.In practice teaching, the teachers can predict scores of other courses from scores of some courses and speak or act with a well-defined objective in mind. To be good for students' mature, the teachers carry out different train pattern to inhomogenous students.

REFERENCES

[1] R.Agrawal and R.Srikant.Fast algolithms for mining association rules in large databases.In Research Report RJ 9839,IBM Almaden Research Center,San Jose,CA,June 1994.

[2] *Lenarcik A. and Piasta Z.Probabilistic rough classifiers with mixture of discrete and continuous variables. In Lin T.Y. and Cercone N., editors, Rough Sets and Data Mining: Analysis for Imprecise Data, pages 373-383. Boston: Kluwer Academic Publishers,1997.*

[3] *J.S.Park,M.S.Chen,and P.S.Yu.An effective hash-based algorithm for mining association rules. In Proc.1995 ACM-SIGMOD Int.Conf.Management of Data (SIGMOND 95),pages 175-186,San Jose,CA,May1995.*

[4] A.Savasere,E.Omiecinski,and S.Navathe.An efficient algorithm for mining association rules in large databases. In Proc.1995Int.conf.Very Large Data Bases(VLDB'95),pages 432-443.Zurich,Switzerland, Sept 1995

[5] H.Toivonen.Sampling large databases for association rules. In Proc.1996Int.conf.Very Large Data Bases(VLDB'96),pages 134-145,Bombay India,Sept.1996

[6] Jiawei Han,Micheline Kamber.Data mining concepts and techniques. Peking：China Machine Press, 2002

[7] Zhu Shaowen. Technique and development of association rule mining.Computer Engineering, 2000(9),4-7

[8] Zhu Yangyong,Zhou xin,Shi Baile.ARMiner a data mining tools based on association rules. High Technology Letters， 2000， Vol.10,No.3:19-22