RESEARCH REVIEW JOURNALS
Volume-03
Issue-12
December -2018

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Patterns from crime data using candidate generation approach

*[1]P.Prabakaran & [2]Dr. K. Rameshkumar

[1]Research Scholar, Computer Science department, Mass college of arts and science, Kumbakonam (India)
[2]Research Supervisor & Assistant Professor, Computer Science dept, Mass college of arts and science, Kumbakonam (India)

ARTICLE DETAILS

Article History
Published Online: 10 December 2018

Keywords
Data Mining, Rule Mining, Pattern mining algorithm, IRM, TPM

***Corresponding Author**
Email: spprabakaran@yahoo.com

ABSTRACT

The rule mining algorithm generates rules from frequent patterns which are mined from Theft Crime dataset. The various combinations of the rules which is produced by the rule mining algorithm may be efficient or inefficient. It is waste of time to run all the rules. So, in order to validate the most efficient rule, the proposed algorithm applied the existing support and confidence measures and additionally one more measure information gain to add more values to the rule generation and validation. The proposed Theft Pattern Mining algorithm is adapted with the Improved Rule Mining algorithm and it is applied to the Tamil Nadu Theft Crime dataset.

1. Introduction

Association rule mining is the process of generating association rules from those large itemsets with the constraints of minimal confidence. Suppose, one of the large itemsets is L_k, $L_k = \{I_1, I_2, \ldots, I_{k-1}, I_k\}$, association rules with this itemsets are generated in the following way: the first rule is $\{I_1, I_2, \ldots, I_{k-1}\} \Rightarrow \{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty.

The proposed rule mining algorithm is divided into two phases, Frequent Pattern generation and Rule construction. The first phase aimed to generate frequent patterns with the help of suitable measure. The second phase has extracted association rule from frequent patterns. The first phase aims to generate frequent itemset using frequent item list generation and frequent item list projection method. The proposed Theft Pattern Mining (TPM) algorithm is used to find the frequent itemset. The second phase consists of two processes. The first process generates association rules with the help of the proposed rule mining algorithm and the second process finds the suitable measures to validate the association rule. Support and Confidence measures are the suitable basic measures to validate the rules. These measures find the frequent item set which is above the support and confidence value which is fixed and validate it.

1.1. Problem Statement

The problem of mining association rules from data with uncertain value is to find all the rules that satisfy specified minimum support, confidence and information gain. The existing rule mining algorithm generates wide number of association rules which contains non interesting rules also. While generating rule mining algorithm, it considers all the discovered rules and hence the performance becomes low. It is also impossible for the end users to understand or check the validity of the large number of complex association rules and thereby restricts the usefulness of the data mining results. The generation of large number of rules also led to heavy computational cost and waste of time. Various methods have been formulated to reduce the number of association rules like generating only rules with no repetition, generating only interesting rules, generating rules that satisfy some higher level criteria etc.

Issues in Finding Association Rules

Multiple Scans across the Transactional Database

While finding frequent itemsets, it is necessary to scan the whole database many times. This multiple scans will lead to the following problems.

a) It leads to wastage of time, because searching the entire database for any type of item takes lot of time.

b) It leads to wastage of space, because the same dataset are scanning again and again many times and it is saved in memory. So, lot of memory space is needed.

A framing of an association rule like P, Q→ R, means that R tends to occur when P and Q occur. An itemset is just a collection of P,Q,R and it is frequent if its item tend to co-occur. To construct association rule, frequent itemsets are generated first and then post process them into rules.

The input of frequent itemset mining [150] is a transaction database and the minimum support threshold MinSupp. The output is the set of all itemsets appearing in at least MinSupp transactions. An itemset is just a set of items that is unordered.

1.2. Existing Work

A. Constraint Based Association Rule Mining algorithm (CBARM)

Constraints based association rule mining is to find all rules from a given data set meeting all the user specified constraints. Apriori and its variants only employ two basic constraints: minimal support and minimal confidence. There

are some other types of rules which can also add more strength to the basic constraints.

CBARM is more active in interactive mining environment, where it becomes a necessity to enable the user to express his interests through constraints on the discovered rules, and to change these interests interactively. The most famous constraints are item constraints that have restrictions on the presence or absence of items in a rule. These constraints can be in the form of conjunction or a disjunction. Such constraints have been introduced first and a new method for incorporating the constraints into the candidate generation phase of the Apriori algorithm was proposed.

B. Rapid Association Rule Mining algorithm (RARM)

Rapid Association Rule Mining (RARM) is proposed to further push the speed barrier so that association rule mining can be performed more efficiently in electronic commerce. To achieve large speed-ups even at low support thresholds, RARM constructs a new data structure called Support-Ordered Trie Itemset (SOTrieIT). This trie-like tree structure stores the support counts of all 1-itemsets and 2-itemsets in the database. All transactions that arrive are pre-processed; all 1-itemsets and 2-itemsets are extracted from each transaction. The extracted information is used to update the SOTrieIT. This structure is then sorted according to the support counts of each node in descending order.

C. Link Rule Miner algorithm (LRM)

LRM algorithm simply clutches high frequent behavior but low frequent behavior remains relied on the support threshold values. The LRM however cannot differentiate noise with the low frequent behavior. While applying this algorithm, the existence of highly low frequent behavior represents flexibility and lack of standardization.

1.3. Improved rule mining algorithm

The improved rule mining algorithm increases the efficiency through the process of reducing the computational time. It can be succeeded by reducing the number of passes over the database, by adding additional constraints on the pattern. In legal applications, some rules will have less weightage and inefficient and some rules will have more weightage.

The two important basic measures for association rules are support (Supp) and confidence (Conf). Support [166] is defined as the percentage or fraction of records that contain X Y to the total number of records in the database. The variable X and Y denotes the number of item set or dataset. For example, if the support of an item is 0.5% it means that only 0.5% of the transaction contain purchasing of that item.

Confidence of an association rule is defined as the percentage of the number of transactions that contain X Y to the total number of records that contain X. Confidence is the measure of strength of the association rule. For example, if the confidence of the association rule X=>Y is 75%, it means that 75% of the transactions that contain X also contain Y together.

A. Association Rule Mining Phase

The association rules are mined from frequent occurred database which is generated from Frequent Pattern Mining approach. Frequent Pattern Mining approach consists of two major things that are Frequent Item List Generation and Frequent Item List Projection which are described in chapter 3.4. Association Rule Mining phase is divided into two processes as follows: Frequent Itemset Generation and construction of association rule from frequent item set.

B. Frequent Item set Generation

The proposed Theft Pattern Mining algorithm is used to find frequent itemsets. The TPM algorithm used support measure and the proposed rule mining algorithm used confidence measures along with the added new measure information gain with TPM to find frequent itemset. The new procedure of IRM is given below.

Improved Rule Mining Procedure

Procedure find IRM (Transaction Database M, MinSupp S, MinGain G)

begin

I_1 = {I itemset}

J_1 = Transaction Database M (With all items not in $I_1$and $\forall N_i$ Nitems =1

removed)

for (i=2; $J_k - 1 \neq \varphi$;i++) do begin

J_i= rule construct($I_i - 1$)

$J_i=\varphi$

for all $j \in J_i$ do begin J=φ

Nj={n.TID | $n \in J_i - 1$,(j– j[i]) ∈n.set of itemsets ∧(j– j[i]) ∈ n.set of itemsets) }

if Support|N_j| ≥ S and Gain|Nj| ≥ G then begin $I_i=H_i\{j\}$

for all $r \in N_j$ do begin

if (Nitems >i) then begin

J= H_i<r,j>

end if

end for

if |J| ≠ 1 then

begin

end

if end if

end for

end for

ans = H_iI_i

end procedure

A. Association Rule Construction

Association rule is constructed with the help of rule construct procedure and the proposed Improved Rule Mining algorithm. The Improved Rule Mining algorithm generates association rule from frequently mined patterns. The ARM algorithms produced many rules which can be generated from patterns. The various rules (patterns) retrieve the effective combination of attributes and number of times it occurs.

The rule construct algorithm uses the following procedure to construct association rule.

Procedure IRM con (n-itemsets)

begin

for all large n-itemsets M_n, n ≥ 2 do begin
V_1 = {consequents of rules derived from M_n with one item in the
consequent};
call rule construct (M_n,V_1);
end procedure
Procedure rule construct (M_n : large n-itemset, V_i : set of i-item consequents)
begin
if (n > i+1) the begin
$V_{i+1=}$ IRM con (V_i)
for all w_{i+1} V_{i+1} do begin
conf=support(M_n)/support(M_n– w_{i+1})
if(conf≥ MinConf) ∧ (gain≥MinGain) then
output the rule (M_n– w_{i+1}→ w_{i+1} with confidence=Conf
informationgain = gain and
support = support (M_n)
else
delete w_{i+1} from V_{i+1}
end if
call rule construct (M_n, V_{i+1})
end for
end if
end procedure

Association rule mining is the efficient method which is used in finding the association rules. These rules describe the associations between the attribute values of the itemsets

Association Rule

An implication expression of the form P□Q, where P and Q are itemsets

Example:{lonely house, bureau pulling} □ {chain hooks}

Rule Evaluation Metrics

Support (s) : Fraction of transactions that contain both P and Q
Confidence (c) : Measures how often items in Q appear in transactions that contain P

{lonely house, bureau pulling}=>chain hooks
s = σ (lonely house, bureau pulling, chain hooks) / |T| = 2 / 5 = 0.4
c = σ (lonely house, bureau pulling, chain hooks) / σ (lonely house, bureau pulling) = 2 / 3 = 0.67

The Table 1.3.1 consist of list of item set which is the example for records of large number of transactions at a shopping center.

Table 1.3.1 Sample database

Transaction ID	*Itemset*
101	apple, orange, mango, banana
102	orange, mango, apple, banana
103	papaya, orange, apple, chickoo
104	apple, banana, orange, mango
105	chickoo,orange,banana, apple,mango

The following rules can be generated from the above table:

R1: apple, orange, mango ⇒ banana
[Supp=60%, Conf=65%, Gain=70%]
R2: apple, orange ⇒ banana, mango
[Supp=35%, Conf=50%, Gain=50%]
R3: apple ⇒ banana, mango, orange,
[Supp=40%, Conf=50%, Gain=55%]
R4: papaya, orange, apple ⇒ chickoo
[Supp=35%, Conf=40%, Gain=50%]
R5: papaya, orange ⇒ chickoo, apple
[Supp=50%, Conf=50%, Gain=60%]
R6: papaya ⇒ chickoo, apple, orange,
[Supp=45%, Conf=60%, Gain=60%]

A. Dataset Description

The Table 1.3.2 shows the number of transactions and attributes of real and synthetic datasets used in this evaluation. All datasets are taken from the UC Irvine Machine Learning Database Repository. Typically, these real datasets are very dense, i.e., they produce many long frequent itemsets even for very high values of support threshold. These datasets mimic the 94 transactions in a retailing environment. Usually the synthetic datasets are sparse when compared to the real sets.

Table 1.3.2 Synthetic and Real Dataset

Dataset	No. of Transactions	No. of Attributes
T40I10D100K	100000	942
Mushroom	8124	119
Gazella	59601	497
TTC Dataset	5000	82

T40I10D100K

This synthetic dataset is generated by IBM Quest Market basket synthetic data generator. An average of 10 items is available in each transaction. The parameters for generating a synthetic database are the number of transactions (in thousands), the average transaction size and the average length of so called maximal potentially large itemsets. The dataset contains 942 numbers of attributes and 1,00,000 transactions

Tamil Nadu Theft Crime Dataset

The crime dataset are constructed from the information collected from various investigating offices and also from the law journals. This dataset includes descriptions of theft cases corresponding to 5000 (transactions) cases with 82 attributes in each case. Each case is identified by its unique number. It consist of attributes such as name of the suspect, age, place of birth, modus operandi of committing the offence, jurisdiction of crime, weapons used etc.

Some suspects will repeatedly do the same style of criminal activities. Some suspects are first time offenders. The researcher aims to mine at a short span of time about the suspects who did the same criminal activity [108], who affects the same group of victims and using same modus operandi style of committing crime based on the repetition of activity or

based on their similarity in physical features. The data are collected from various sources like Madras Law Journals, Current Tamil Nadu cases, Law Reporters etc.

1.4. Experimental study and analysis

The proposed improved rule mining algorithm is implemented on Windows 7. The proposed algorithm is implemented in Python.

The IRM algorithm is tested with four datasets. The figure 1.4.1 show the comparative study of number of rules mined between the proposed IRM and CBARM algorithm varying minimum support using Tamil Nadu Theft Crime dataset respectively.

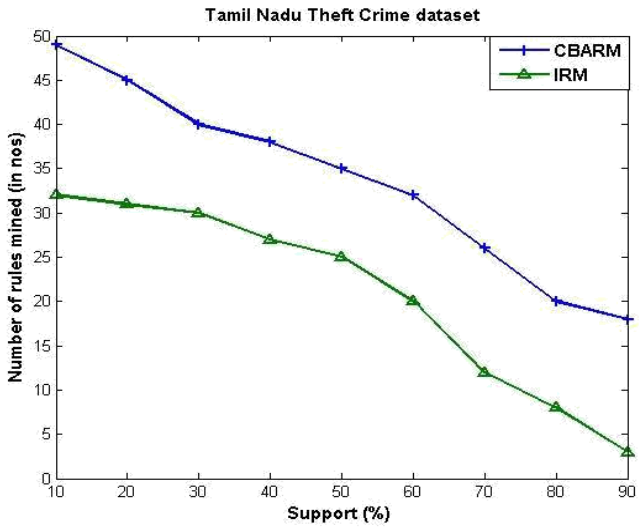

Figure 1.4.1 Comparing the performance of IRM and CBARM algorithm with varying support using Tamil Nadu Theft Crime dataset

From the above figure 1.4.1, the performance of the proposed IRM algorithm is better than small dataset into high density dataset. In TTC dataset, the number of rules mined in the proposed IRM algorithm is less than CBARM algorithm.

Figure 1.4.2 shows the comparative study of the number of rules mined between proposed IRM algorithm and CBARM algorithm varying information gain using TTC dataset respectively.

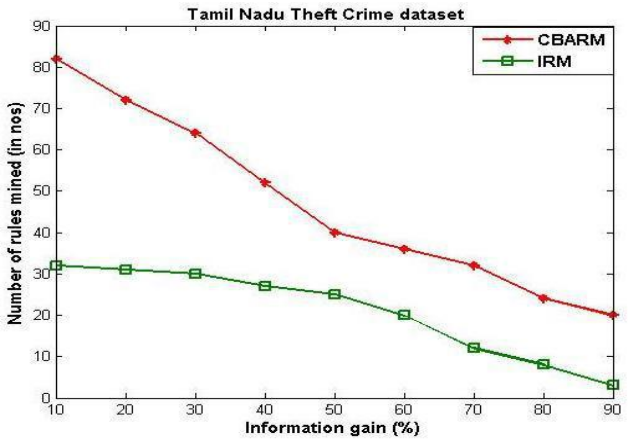

Figure 1.4.2 Comparing the performance of IRM and CBARM algorithm with varying information gain using Tamil Nadu Theft Crime Dataset.

2. Main title

- Problem statement
- Existing work
- Improve rule mining algorithm
- Experimental study and analysis

3. Conclusion

The proposed IRM algorithm has applied the measures support, confidence and information gain thresholds that are efficiently calculated in the dataset. The Improved Rule mining algorithm adapted the above TPM procedure for mining frequent itemsets. Experimental results have shown that IRM algorithm can always find most of the possible association rules with satisfying minimum thresholds. It performed well in high and small density datasets. The proposed TPM procedure is successfully integrated with IRM. The IRM algorithm needs to be extended to handle various crimes. Particularly crime against women is the burning issue and become a great challenge to the law enforcement agencies.

References

1. Aniruddha Kshirsagar, Lalit Dole. 2014. " A Review On Data Mining Methods For Identity Crime Detection", International Journal of Electrical, Electronics and Computer Systems, Vol.2, Issue.1, pp. 51-55
2. Applications of Data Mining [online] Available at: www.tutorialspoint.com/data_mining/dm_applications_trends .htm [Accessed on 19 December 2011]
3. Arun K. pujari, "Data Mining Techniques", 2st Edition. 2010. Publisher Orient Blackswan Pvt.Ltd. New Delhi, pp.1-340
4. Chae Chang Lee, Ji Won Yoon. 2013. "A data mining approach using transaction patterns for card fraud detection", Computing Research Repository, Seoul, Republic of Korea, pp.1-12
5. Crime and Everyday life [online] Available at :www.goodreads.com/book/show/4475249-crime-and-everyday-life [Accessed on 5 July 2012]
6. Crime Records [online] Available at: http://ncrb.nic.in/ciiprevious/Data/CD-CII2006/cii-2006/Chapters [Accessed on 12 April 2014].

