Vikram Rajpoot, Prof. Shailendra ku. Shrivastava, Prof. Abhishek Mathur/ International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
www.ijera.com Vol. 2, Issue4, July-August 2012, pp.2210-2215

An Efficient Constraint Based Soft Set Approach for Association Rule Mining

Vikram Rajpoot, Prof. Shailendra ku. Shrivastava, Prof. Abhishek Mathur
M.Tech,Department of IT ,SATI Vidisha
Head, Department of IT ,SATI
Asst. Prof. ,Department of IT,SATI

ABSTRACT

In this paper, we present an efficient approach for mining association rule which is based on soft set using an initial support as constraints. In this paper first of all initial support constraint is used which can filter out the false frequent item and rarely occurs items. Due to deletion of these items the structure of dataset is improved and result produced is faster, more accurate and take less memory than previous approach proposed in paper a soft set approach for association rules mining . After the deletion of these items the improved dataset is transformed in to Boolean-valued information system. Since the ‘‘standard” soft set deals with such information system, thus a transactional dataset can be represented as a soft set. Using the concept of parameters co-occurrence in a transaction, we define the notion of regular association rules between two sets of parameters, also their support, confidence and properly using soft set theory. The results show that our approach can produce strong association rules faster with same accuracy and less memory space.

Keywords:-**Association rules mining, Boolean-valued information systems, Soft set theory, Items co-occurrence, red_sup constraint**.

I. INTRODUCTION

1.1 Association rule

Association rule is one of the most popular data mining techniques and has received considerable attention, particularly since the publication of the AIS and Apriori algorithms [2,3]. They are particularly useful for discovering relationships among data in huge databases and applicable to many different domains including market basket and risk analysis in commercial environments, epidemiology, clinical medicine, fluid dynamics, astrophysics, and crime prevention.

The association rules are considered interesting if it satisfies certain constraints, i.e. predefined minimum support (min_sup) and minimum confidence (min_conf) thresholds.For Rule X→Y their support and confidence is calculated as:

$$\text{Support } (X \rightarrow Y) = \frac{(X \cup Y).\text{count}}{N} \quad (1)$$

N = Total number of transaction

$$\text{Confidence } (X \rightarrow Y) = \frac{(X \cup Y).\text{count}}{X.\text{count}} \quad (2)$$

In this X is antecedent and Y is consequent. The rule X→Y has support s% in the transaction set D if s% of transactions in D contain X∪Y. The rule has confidence c% if c% of transactions in D that contain X also contain Y. The goal of association rule mining is to find all the rules with support and confidence exceeding user specified thresholds. Many algorithms of association rules mining have been proposed. The association rules method was developed particularly for the analysis of transactional databases.

A huge number of association rules can be found from a transactional dataset.The rules that satisfy the minimum support threshold and minimum confidence threshold is called the strong association rules and rest of the rules is discrded.

1.2 Soft set

Soft set theory [7], proposed by Molodtsov in 1999, is a new general method for dealing with uncertain data. Soft sets are called (binary, basic, elementary) neighborhood systems. As for standard soft set, it may be redefined as the classification of objects in two distinct classes, thus confirming that soft set can deal with a Boolean-valued information system. Molodtsov [7] pointed out that one of the main advantages of soft set theory is that it is free from the inadequacy of the parameterization tools, unlike in the theories of fuzzy set [8]. Since the ‘‘standard” soft set (F,E) over the universe U can be represented by a Boolean-valued information system, thus a soft set can be used for representing a transactional dataset. Therefore, one of the applications of soft set theory for data mining is for mining association rules. However, not many researches have been done on this application.

Definition: - A pair (F,E) is called a soft set over U, where F is a mapping given by:

$$F : E \rightarrow P(U) \qquad (3)$$

In other words, a soft set over U is a parameterized family of subsets of the universe U. For e belongs E, F(e) may be considered as the set of e-elements of the soft set (F,E) or as the set of e-approximate elements of the soft set. Clearly, a soft set is not a (crisp) set.

To illustrate this idea, let we consider the following example.

Example . Let we consider a soft set (F, E) which describes the "attractiveness of houses" that Mr. X is considering to purchase.Suppose that there are six houses in the universe U under consideration,

$U = \{ h_1 , h_2 , h_3 , h_4 , h_5 , h_6 \}$

and

$E = \{ e_1 , e_2 , e_3 , e_4 , e_5 \}$

is a set of decision parameters, where e_1 stands for the parameters "expensive", e_2 stands for the parameters "beautiful", e_3 stands for the parameters "wooden", e_4 stands for the parameters "cheap", e_5 stands for the parameters "in the green surrounding".

Consider the mapping from equation (3)

$$F : E \rightarrow P(U),$$

given by "houses (.) ", where (.) is to be filled in by one of parameters e belongs to E.Suppose that

$F(e_1)=\{h_2, h_4\}$ $F(e_2)=\{h_1, h_3\}$ $F(e_3)=\{h_3, h_4, h_5\}$ $F(e_4)=\{h_1, h_3, h_5\}$ $F(e_5)=\{h_1\}$

Therefore $F(e_1)$ means "houses (expensive)", whose functional value is the set $\{ h_2 ,h_4\}$. Thus, we can view the soft set (F, E) as a collection of approximations as below

$$(F,E)=\left\{\begin{array}{c}\text{expensive houses}=\{h_2,h_4\},\\ \text{beautiful houses}=\{h_1,h_3\},\\ \text{wooden houses}=\{h_3,h_4,h_5\},\\ \text{cheap houses}=\{h_1,h_3,h_5\},\\ \text{in the green surrounding houses}=\{h_1\}\end{array}\right\}$$

Fig. 1 Soft set example

Each approximation has two parts, a predicate p and an approximate value set v.

For example, for the approximation "expensive houses = $\{ h_2 ,h_4 \}$ ", we have the predicate name of expensive houses and the approximate value set or value set $\{ h_2 ,h_4 \}$ .Thus, a soft set (F, E) can be viewed as a collection of approximations below:

$(F,E)= \{ p_1 = v_1, p_2 = v_2 ,p_3 = v_3,\ldots., p_n = v_n \}$

Tabular representation of soft set

U	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_3}$	$\mathbf{e_4}$	$\mathbf{e_5}$
$\mathbf{h_1}$	0	1	0	1	1
$\mathbf{h_2}$	1	0	0	0	0
$\mathbf{h_3}$	0	1	1	1	0
$\mathbf{h_4}$	1	0	1	0	0
$\mathbf{h_5}$	0	1	1	0	0
$\mathbf{h_6}$	0	0	0	0	0

Fig. 2 Soft set in Boolean system

Now here we summarize our paper section 2 describe the previous related works. Section 3 describe our proposed approach and section 4 describe our implementation and result of proposed CSS approach and section 5 conclude our paper.

II. RELATED WORK

In the previous paper A soft set approach for association rule mining [1] there are direct applicability of soft set on the Boolean valued information system that contains large number of false frequent item and also contains rare items whose support is less than initial support. Due to the presence of such items in database the previous approach is slow in result generation. These false frequent item and rare item is neither be frequent and no interesting rule is generated with the help of these items. These items is removed when we generated the frequent pattern latter in the process with the help of min_sup. If these item not deleted from input transaction then time complexity and space complexity of the approach is increased. Therefore previous approach has high time and space complexity.

In the previous papers methods proposed to found out association rule from the transaction dataset. These method is based on Rough set [16,18] to find association rule. In these method rough set is used to find the association rule on the basis of decision table .In these methods first of all find the conditional attribute and on the basis of which we construct the decision table. This decision table is used to find the association rules

in the IF-THEN context. With the help of Rough set for association rule we find rule with less response time than traditional techniques [14,15] of association rule mining. But in the rough set based approach the decision table is maintain and then association rule is derived from that decision table is also time consuming in rule generation.

III. PROPOSED WORK

In our proposed approach we reduce the dataset with the help of initial red_sup.Due to this the false frequent items and rare items is eliminated or deleted from the input transaction dataset and the response time of rule generation is increased .The algorithm of our proposed work is described below.

3.1 Proposed CSS algorithm

Input :- transaction dataset D (N is the total number of transactions, n is the total number of items present), initial red_sup (initial reduced support) ,min_sup (minimum support threshold), min_conf (minimum confidence threshold) .

Output :- Strong Association rule.

Algorithm

Step 1 :- Scan the dataset D for all transactions 1 to N.

Step 2 :- Calculate the support of all items present in the transaction dataset.

Step 3 :- for all items in dataset

If initial red_sup is greater than item support than delete that item from transaction dataset.

Step 4:- Convert the reduced dataset obtained in step 3 into Boolean valued information system $S=(U,A,V_{\{0,1\}},F)$.

Step 5:- Apply the soft set (F,E) on the Boolean valued information system S .

Step 6:- Apply the principle of parameter co-occurrence and calculate the count of various itemsets.

Step 7:- Generate the association rule from the frequent patterns and check with min_conf threshold to find out the rule is strong or not.

Step 8 **:-** End.

3.2 Proposed method Example

Fig. 3 shows the input transaction dataset that contain 10 transactions. Suppose initial red_sup is 2 ,min_sup is also 2 and confidence is 40%.The transaction dataset is used as an input for the proposed example is shown in Fig. 3 . We perform different steps of our CSS algorithm on it and also show the results of the step in the figure which is shown after the step is apply on the dataset.The figure give the clear view of the operation performed by the various step.

TID	Items
1	Canada , Iran , USA, crude, ship
2	Canada , Iran , USA, crude, Coffee,ship
3	USA, earn
4	USA, jobs , cpi
5	USA, jobs , cpi
6	USA, earn ,corn, cpi
7	Canada , sugar , tea
8	Canada , USA ,Africa, trade, acq
9	Canada , USA , trade, acq
10	Canada , USA , earn

Fig. 3 Transaction dataset

Now the first step of our proposed algorithm is apply means scan the transaction dataset and generate the support of various items present in the dataset.

The result of second step generate the support of various is shown below.

sup{canada}= 6 sup{USA}= 9 sup{Iran}= 2

sup{trade}= 2 sup{acq}= 2 sup{ sugar}=1

sup{tea}=1 sup{earn}= 3 sup{crude}=2

sup{corn}= 1 sup{Africa} =1 sup{coffee}=1

sup{cpi}=3 sup{ship}=2

Fig. 4 support of various items

Result of the second step is shown above i.e. the support of various items that present in transaction dataset.Now we apply step 3 of our approach delete those items from transaction dataset whose support is less than red_sup threshold. Since the minimum red_sup threshold is 2 then result of step 3 the reduced dataset is shown below in fig. 5.

TID	Items

1	Canada , Iran , USA, crude, ship
2	Canada , Iran , USA, crude,ship
3	USA, earn
4	USA, jobs , cpi
5	USA, jobs , cpi
6	USA, earn , cpi
7	Canada
8	Canada , USA , trade, acq
9	Canada , USA , trade, acq
10	Canada , USA , earn

Fig. 5 Reduced transaction dataset

The support of item Sugar ,Tea ,Africa, Corn, Coffee is 1 which is smaller than the predefined red_sup threshold therefore these items is deleted from the original transaction dataset and after deletion of these items we get the more accurate dataset that contains no false frequent items and no rare items.

Now we apply the step 4 of our algorithm convert the reduced dataset of step 3 into Boolean valued information system.In Step 5 soft set is apply to the Boolean valued information system obtained from the step 4.Result of step 5 is shown below.

(F,E)={canada={1,2,7,8,9,10}
USA={1,2,3,4,5,6,8,9,10}Iran={1,2}
trade={1,2} acq={8,9} earn={3,10}
crude={1,2} cpi={3,10} ship={1,2} jobs={4,5} }

Fig. 6 Soft set representation

After the sot set is apply in step 5 we apply the parameter co-occrance to generate the support of various combination of itemsets and deletet those items set whose support is less than min_sup.The result of step 6 shown below.

$coo(u_1)$ = Canada , Iran , USA, crude, ship

$coo(u_2)$ = Canada , Iran , USA, crude,ship

$coo(u_3)$= USA, earn

$coo(u_4)$= USA, jobs , cpi

$coo(u_5)$ = USA, jobs , cpi

$coo(u_6)$ = USA, earn , cpi

$coo(u_7)$= Canada

$coo(u_8)$= Canada , USA , trade, acq

$coo(u_9)$= Canada , USA , trade, acq

$coo(u_{10})$= Canada , USA , earn.

Fig. 7 Parameter co-occurance

Now with the help of parameter co-occurance we calculate the support of various itemsets .The support of various itemsets is shown below.

Sup{canada}={ $u_1,u_2,u_7,u_8,u_9,u_{10}$}=6

Sup{USA}={$u_1,u_2,u_3,u_4,u_5,u_6,u_8,u_9,u_{10}$}=9

Sup{Iran} = {u_1,u_2} = 2

Sup{canada,USA} = {u_1,u_2,u_5,u_9,u_{10}} = 5

Sup{canada,Iran} = {u_1,u_2} = 2

Sup{canada,Iran,USA} = {u_1,u_2} = 2

Sup{crude} = {u_1,u_2} = 2

Sup{ship} = {u_1,u_2} = 2

Sup{earn} = {u_3,u_6,u_{10}} = 3

Sup{jobs} = {u_4,u_5} = 2

Sup{cpi} = {u_4,u_5,u_6} = 3

Sup{trade} = {u_8,u_9} = 2

Sup{acq} = {u_8,u_9} = 2

Fig. 8 Support of itemsets

In the last step we generate association rule from the frequent patterns generate in the step 6 and check the rules satisfy the min_conf threshold.Rules that satisfies the min_conf threshold is strong association rules is accepted and rules that not satisfied the min_conf threshold is not strong association rules and rejcted.

Usa,Canada → ship

Conf(Usa,Canada → ship) =2 / 5

Conf(Usa,Canada → ship)= 40%

Therefore confidence of rule Usa,Canada→ ship is 40% which is equal to min_conf threshold.Thresfore this rule is strong association rule and accepted. In the same manner all other rules is generated and their confidence is calculated then on basis of min_conf thresholds we decide rule is strong or not.

IV. EXPERIMENT RESULT

In this section, we compare the proposed CSS method for association rules mining with the algorithm of [1]. The proposed approach CSS and Previous soft set[1] is executed on dataset derived from [20]. The algorithm of the proposed approach is implemented in MATLAB version 7.6.0.324 (R2008a).

A Dataset derived from the widely used Reuters-21578 [20].It contains 30 transactions with TIDs 1 to 30 and contains 10 items labelled P_1 to P_{10}.Now we show the execution time graph between CSS approach and Soft set approach.In execution graph the X-axis indicate the 6 function of approaches and Y-axis indicate time in second.After which we show the bar graph of memory used between CSS approach and Soft set approach and finaly we give the table that compare excution time differences as the Min_sup and Min_conf threshold is change.

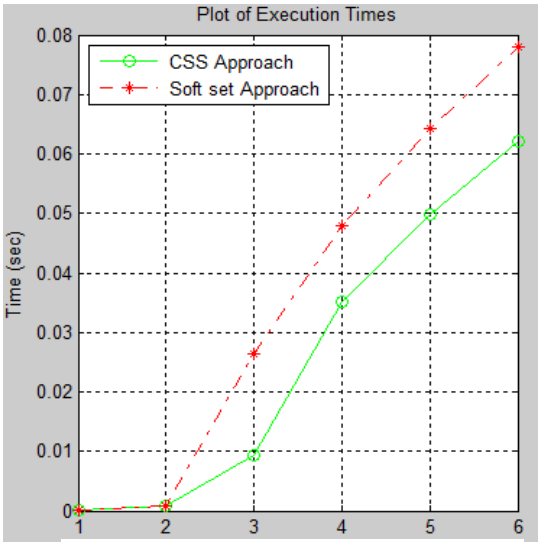

Fig.9 Execution time comparision

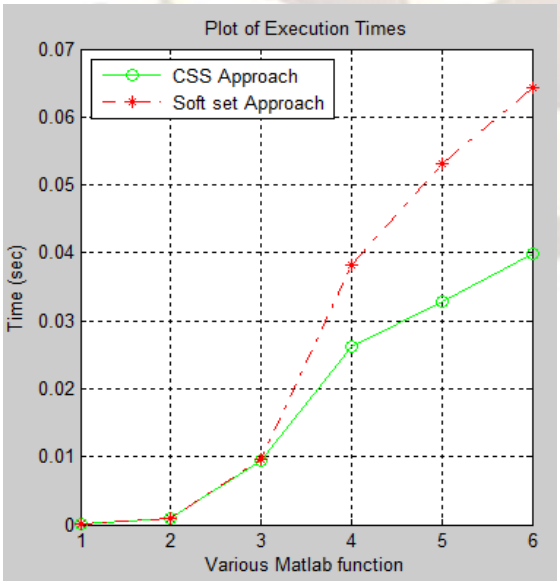

Fig.10 Execution time comparision min_sup=3

Now we show the Memory bar graph which repersent the memory used in the process (1) is soft set and (2) is CSS approach.

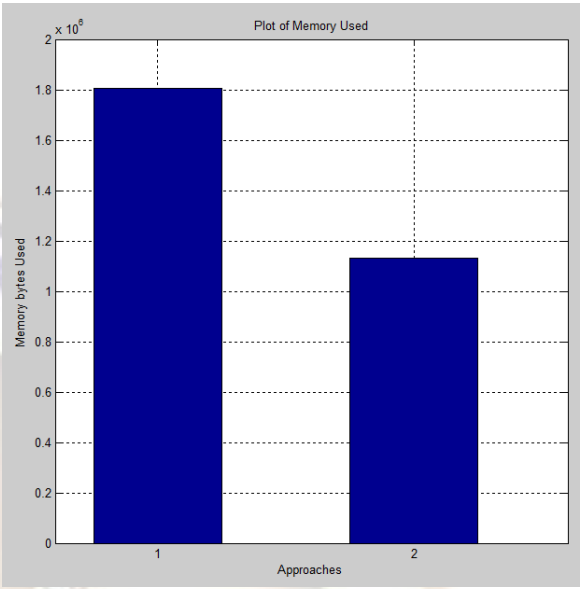

Fig. 11 Memory comparision

Here we give the tabular comparision of execution time between Soft set and CSS approach as the min_sup threshold is change.

Min sup	Min conf	Soft set approach (execution time in sec)	CSS approach (execution time in sec)
2	.6	.07794	.06225
3	.6	.06438	.03994
4	.6	.06323	.02759
5	.6	.05507	.0189

Table 1 Result Analysis

It is clear from the resullt shown above that our proposed CSS approach is faster and efficient than

Table 1 Result analysis oach

V. CONCLUSION

Soft set approach for association rule mining [1] is a new method for finding association rule .With the help of soft set we can handle the uncertainty present in the dataset. This approach has more time and space complexity and also has chances of some inaccurate result due to the presence of some false frequent items and rare items that never be frequent. In our proposed

approach firstly we reduce these items from input transaction dataset with the help of initial red_sup and then convert that reduced dataset in to Boolean valued information system.In the next step we apply soft set to handle uncertanity of information system.Now we apply the parameter co-occurance on the soft set to generate the count of various itemset and then generate the resulting strong association rule.In our approach due to deletion of false frequent items and rare items the space and time complexity is reduced and the generate the result with less time and take less memory space and same accurate than previous approach[1].

REFERENCES

[1] T. Herawan, M.M. Deris A soft set approach for association rule mining Knowledge-Based Systems 24 (2011) 186–195.

[2] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the ACM SIGMOD International Conference on the Management of Data, 1993, pp. 207–216.

[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 1994, pp. 487–499.

[4] M. Mat Deris, N.F. Nabila, D.J. Evans, M.Y. Saman, A. Mamat, Association rules on significant rare data using second support, International Journal of Computer Mathematics 83 (1) (2006) 69–80.

[5] A.H.L. Lim, C.S. Lee, Processing online analytics with classification and association rule mining, Knowledge-Based Systems 23 (3) (2010) 248–255.

[6] Y.L. Chen, C.H. Weng, Mining fuzzy association rules from questionnaire data, Knowledge-Based Systems 22 (1) (2009) 46–56.

[7] D. Molodtsov, Soft set theory-first results, Computers and Mathematics with Applications 37 (1999) 19–31.

[8] L.A. Zadeh, Fuzzy set, Information and Control 8 (1965) 338–353.

[9] P.K. Maji, A.R. Roy, R. Biswas, An application of soft sets in a decision making problem, Computers and Mathematics with Applications 44 (2002) 1077– 1083.

[10] D. Chen, E.C.C. Tsang, D.S. Yeung, X. Wang, The parameterization reduction of soft sets and its applications, Computers and Mathematics with Applications 49 (2005) 757–763.

[11] A.R. Roy, P.K. Maji, A fuzzy soft set theoretic approach to decision making problems, Journal of Computational and Applied Mathematics 203 (2007) 412–418.

[12] Y. Zou, Z. Xiao, Data analysis approaches of soft sets under incomplete information, Knowledge Based Systems 21 (2008) 941–945.

[13] Z. Kong, L. Gao, L. Wang, S. Li, The normal parameter reduction of soft sets and its algorithm, Computers and Mathematics with Applications 56 (2008) 3029–3037.

[14] R. Feldman, Y. Aumann, A. Amir, A. Zilberstein, W. Klosgen, Maximal association rules: a new tool for mining for keywords cooccurrences in document collections, in: The Proceedings of the KDD 1997, 1997, pp. 167–170.

[15] A. Amir, Y. Aumann, R. Feldman, M. Fresco, Maximal association rules: a tool for mining associations in text, Journal of Intelligent Information Systems 25 (3) (2005) 333–345.

[16] J.W. Guan, D.A. Bell, D.Y. Liu, The rough set approach to association rule mining, in: The Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), 2003, pp. 529–532.

[17] P.K. Maji, R. Biswas, A.R. Roy, Soft set theory, Computers and Mathematics with Applications 45 (2003) 555–562.

[18] Y. Bi, T. Anderson, S. McClean, A rough set model with ontologies for discovering maximal association rules in document collections, Knowledge-Based Systems 16 (2003) 243–251.