A New Soft Set Based Association Rule Mining Algorithm

1Sahiv Singh Kanojiya, 2Akhilesh Tiwari
Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P.), India

Abstract—In this paper a new algorithm has been proposed. Proposed algorithm is suitable for mining association rules containing only predetermined data items (PDIs). In case of databases having uncertain and redundant data, traditional algorithms are not able to offer desired accuracy and performance during mining process. Considering this fact soft set based algorithm has been developed which first performs the application of soft set for handling uncertainty issue and then with the help of PDI constraints performs the identification of transactions containing PDI items.

Keywords: - Association Rule, Predetermined Data Items (PDIs), Soft Set, Uncertain Data

I. INTRODUCTION

A. Association Rule Mining

Mining frequent item sets is a fundamental and essential operation in data mining applications. This is one essential step towards discovery of association rules, strong rules and correlations. Technological advancements contributed for the growing size of data and ultimately emphasizes on the use of automated tools and efficient algorithms for the analysis of huge data. Considering the amount of computations involved in data mining, high-performance computing is an essential component for any successful large-scale data mining application. From the time the AIS and Apriori [2,3] algorithms was proposed, researchers have developed several improved algorithms [4,5].

The problem of mining association rules is to generate all rules that have support and confidence greater than or equal to some user specified minimum support and minimum confidence threshold respectively. A formal statement of the association rule problem is given in [2].

Let I = \{i_1, i_2, i_3, \ldots, i_m\} be a set of m distinct literals called items, D is a set of transactions (variable length) over I. Each transaction contains a set of items i_1, i_2, i_3, \ldots, i_k, \subseteq I. Each transaction is associated with an identifier, called TID.

An association rule is an implication of the form X \Rightarrow Y, where X, Y \subseteq I and X \cap Y = \phi. Here X is called the antecedent and Y is called the consequent of the rule.

The rule X \Rightarrow Y holds in the transaction set D with confidence \alpha if among those transactions that contain X \alpha\% of them also contain Y.

The rule X \Rightarrow Y has support S in the transaction set D if S\% of transactions in D contains X \cup Y.

The selection of association rules is based on these two values (some additional constraints may also apply). These are two important measures of rule interestingness. They respectively reflect usefulness and certainty of a discovered rule. They can be described by the following equations:

\text{Support} (X \Rightarrow Y) = \frac{\text{Frequency}(X \cup Y)}{|D|} \quad (1)

\text{Confidence} (X \Rightarrow Y) = \frac{\text{Frequency}(X \cup Y)}{\text{Frequency}(X)} \quad (2)

where |D| represents the total number of transactions (tuples) in D.

A frequent itemset is an itemset whose number of occurrences is above a minimum support threshold. An itemset of length k is called k-itemset and a frequent itemset of length k as k-frequent itemset.

The association rules are considered interesting if it satisfies certain constraints, i.e. predefined minimum support (min_sup) and minimum confidence (min_conf) thresholds.

The main objective of association rule mining task is to discover all the rules having support and confidence values exceeding user specified or initially provided thresholds.

Traditional algorithms work fine if data inside the considered dataset is not uncertain but if data involves uncertainty then case specific algorithms are required.

Now-a-days, researchers are focusing on uncertainty issues by involving fuzzy sets and soft sets for the development of new algorithms [1, 20, 21]. Present work also proposes an improved soft set based algorithm for handling uncertain and redundant data.

Following Section gives the detailed description of soft set theory.

B. Soft Set Theory

Literature reveals that soft set theory can be applied for the solution of problems that contain uncertainties and application of soft sets can be useful for the development and use of algorithms for effective decision making. Molodtsov [7] suggested the concept of soft set for dealing with uncertainties.

In recent years, works on soft set theory and its applications have been progressing in rapid fashion.
Literature indicates that different types of operations of soft sets are defined and used in the works related to soft set theory and its applications. It has been observed that the main advantage of soft set theory is that it is free from the inadequacy of the parameterization tools; unlike in the theories of fuzzy set [8, 9, 10].

Soft Set: A soft set is a mapping of parameters set into the set of all subsets of the universe [7, 21, 22]. Mathematically Soft Sets can be defined as- Let U be an initial universe set and let E be a set of parameters. Then, to use an adequate parametrization, Molodtsov [7] gave the definition of soft sets as follows

A pair (F, E) is called a soft set (over U) if and only if F is a mapping of E into the set of all subsets of the set U. This mapping can be given by:

\[ F : E \rightarrow P(U) \]  

(3)

Where F is called approximate function and E is called parameters set of the soft set (F, E). Each parameter \( x \in E \) may be a fuzzy word, real number and so on. For \( x \in E \), the set \( F(x) \) is called \( x \)-approximation of the soft set \( (F, E) \) which may be arbitrary, some of them may be empty and some may have nonempty intersection. 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 is for mining association rules. However, not many researches are available on the application of soft sets for data mining. To illustrate the concept of soft sets, let us discuss following popular example.

Example: Suppose that there are six houses in the universe \( U = \{ h_1, h_2, h_3, h_4, h_5, h_6 \} \) and \( E = \{ e_1, e_2, e_3, e_4, e_5 \} \) is the set of parameters. Here, \( e_i \) (i = 1,2,3,4,5) stand for the parameters “expensive”, “beautiful”, “wooden”, “cheap” and “in green surroundings” respectively. In this case, to define a soft set means to point out expensive houses, beautiful houses and so on. Then, following soft sets describe the attractiveness of the houses which a couple is going to buy. Suppose that according to Mr. X’s, F\( (e_1) = \{ h_2, h_3, h_4 \} \), F\( (e_2) = U \), F\( (e_3) = U \), F\( (e_4) = \emptyset \) and F\( (e_5) = \{ h_1, h_2, h_3, h_4 \} \). If Mr. X considers a subset A = \{ e_1, e_4, e_5 \} of E, then soft set \( F_A = \{ \{ e_1, \{ h_2, h_3, h_4 \} \}, \{ e_4, \{ h_1, h_2, h_3, h_4 \} \} \)\). This soft set \( F_A \) can be represented with the help of boolean system and is shown in Fig. 2.

Therefore \( F (e_1) \) means “houses (expensive)”, whose functional value is the set \( \{ h_2, h_3, h_4 \} \). Thus, we can view the soft set \( (F, E) \) as a collection of approximations as below

\[
F, E=\begin{cases} 
\text{expensive houses} = \{h_2, h_3, h_4\} \\
\text{in green surroundings} = \{h_1, h_2, h_5, h_6\}
\end{cases}
\]

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_3, h_4\}”, we have the predicate “expensive houses” and the approximate value set or value set is \{h_2, h_3, h_4\}. Thus, a soft set \( (F, E) \) can be viewed as a collection of approximations. Following Fig. 2 shows the tabular representation of soft set \( F_A \).

II. RELATED WORK

Soft sets represent a powerful mechanism or tool for effective decision making in case of data mining. It has been observed that soft sets fulfills the expectation of drawing conclusions from data or analysing data, especially in situations where some uncertainty exists in the data.

It is efficient in dealing with uncertain situations because of parameterized concept. Various researches have been done in theory and practices. In [1] a soft set based approach for association rule mining has been discussed.

It has been observed that there is a direct applicability of soft set on the Boolean valued information system that contains large number of false frequent items and rare items whose support is less than the initial support. Due to the presence of such items in databases traditional approaches are slow.

Furthermore, there is negligible chance that the false frequent items and rare items become frequent and rare items and the approximate value set or value set is \{h_2, h_3, h_4\}. It has been observed that there is a direct applicability of soft set on the Boolean valued information system that contains large number of false frequent items and rare items whose support is less than the initial support. Due to the presence of such items in databases traditional approaches are slow.
Researchers have also used Rough Sets [6, 16, 17, 18, 19] for finding association rules based on decision table where first task is to find the conditional attributes and then on the basis of same construction of decision table. Further decision table is used for finding the association rules on the IF-THEN context.

With the help of rough sets association rules can be generated in less response time in comparison to the approaches proposed in [11, 12, 13, 14, 15].

It has been realized that the use of Rough Sets is effective but decision table maintenance and generation of association rules with the help of maintained decision table is a time consuming activity.

III. PROPOSED ALGORITHM

This paper proposes a new algorithm which is based on soft set theory. Proposed algorithm is capable in handling uncertainty and is useful for the generation of association rules containing pre-determined items.

Formal description of algorithm is given below.

A. Algorithm: PDISS

Input: Transactional dataset D, PDIs (Set of predetermined data items), minimum support, minimum confidence.

Output: Strong Association Rules (containing PDIs).

STEPS :-

1. Convert the input Transactional dataset in to Boolean valued information system (BVIS).
2. Apply the soft set (F, E) To BVIS.
3. Scan the soft set generated during step 2 and deletes those items in transaction dataset that does not belongs to PDI.
4. Apply the principle of parameter co-occurrence.
5. Calculate the support of item sets.
6. Compare the support of itemsets with min_sup, itemsets whose support is less than min_sup will not be considered as frequent itemsets.
7. Generation of association rule from frequent item sets.
8. Calculate the confidence of association rules and compare with the minimum confidence threshold.
9. Rules satisfying minimum confidence threshold will be considered as strong association rules containing PDIs.
10. END

Working Example:

This Section demonstrates the working of proposed algorithm considering sample dataset. Figure 3 shows the sample input transaction dataset which contains 10 transactions. Suppose predetermined data item set (PDI) contains elements: Canada, Iran, USA, crude jobs, cpi, sugar, earn, trade, cpi. Assume min_support is 30% and min_confidence is 50%.

TID Items
1 Canada, Iran, USA, crude, ship
2 Canada, Iran, USA, crude,
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

Firstly the input dataset is transformed into Boolean Valued Information System (BVIS) and then soft set is applied. The result of soft set application is as 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={8,9} acq={8,9} earn={3,6,10} crude={1,2} cpi={4,5,6} ship={1,2} jobs={8} sugar={7} tea={7} corn={6}}

Fig. 4: Soft Set

Suppose the predetermined data itemset contains elements: Canada, USA, Iran, trade, earn,ship, acq, crude.

Now application of step 3 will result in the following.
(F,E)={Canada={1,2,7,8,9,10} USA={1,2,3,4,5,6,8,9,10} Iran={1,2} trade={8,9} acq={8,9} earn={3,6,10} crude={1,2} cpi={4,5,6} ship={1,2}}

Fig. 5: Modified Soft Set

Due to this, initial dataset is reduced to the form shown below
1 Canada, Iran, USA, crude, ship
2 Canada, Iran, USA, crude, ship
3 USA, earn
4 USA, cpi
5 USA, cpi
6 USA, earn, cpi
7 Canada
8 Canada, USA, trade, acq
Fig. 6: Reduced Dataset

Now according to step 4 principle of parameter co-occurrence is applied to the dataset and result is shown below. Here coo represents co-occurrences of items in a transaction.

coo(u1) = Canada, Iran, USA, crude, ship

coo(u2) = Canada, Iran, USA, crude, ship

coo(u3)= USA, earn

coo(u4)= USA, jobs, cpi

coo(u5) = USA, jobs, cpi

coo(u6) = USA, earn, cpi

coo(u7)= Canada

coo(u8)= Canada, USA, trade, acq

coo(u9)= Canada, USA, trade, acq

coo(u10)= Canada, USA, earn.

Now the support of various itemsets is calculated

\[ \text{Sup\{canada\}} = \{u1,u2,u7,u8,u9,u10\} = 6 \]
\[ \text{Sup\{USA\}} = \{u1,u2,u3,u4,u5,u6,u9,u10\} = 9 \]
\[ \text{Sup\{Iran\}} = \{u1,u2\} = 2 \]
\[ \text{Sup\{canada,USA\}} = \{u1,u2,5,u9,u10\} = 5 \]
\[ \text{Sup\{canada,Iran\}} = \{u1,u2\} = 2 \]
\[ \text{Sup\{canada,Iran,USA\}} = \{u1,u2\} = 2 \]
\[ \text{Sup\{crude\}} = \{u1,u2\} = 2 \]
\[ \text{Sup\{ship\}} = \{u1,u2\} = 2 \]
\[ \text{Sup\{earn\}} = \{u3,u6,u10\} = 3 \]
\[ \text{Sup\{jobs\}} = \{u4,u5\} = 2 \]
\[ \text{Sup\{cpi\}} = \{u4,u5,u6\} = 3 \]
\[ \text{Sup\{trade\}} = \{u8,u9\} = 2 \]
\[ \text{Sup\{Canada,cpi\}} = 0 \]
\[ \text{Sup\{Canada, earn\}} = 1 \]
\[ \text{Sup\{USA,cpi\}} = \{u4,u5,u6\} = 3 \]
\[ \text{Sup\{canada,USA,acq\}} = 2 \]

Now the specified min_sup threshold is compared for the generation of frequent itemsets. Frequent itemsets generated are as follows.

\{canada,USA\}=5
\{earn\}=3
\{cpi\}=3
\{USA, cpi\}=3

As the next step, association rules are generated and shown below.

\[ \text{USA} \rightarrow \text{cpi} \]
\[ \text{cpi} \rightarrow \text{USA} \]
\[ \text{Canada} \rightarrow \text{USA} \]
\[ \text{USA} \rightarrow \text{Canada} \]

Finally, confidence of association rules is calculated as per the following.

\[ \text{Confidence of (USA \rightarrow cpi)} = \frac{3}{9} = 30\% \]
\[ \text{Confidence of (cpi \rightarrow USA)} = \frac{3}{3} = 100\% \]
\[ \text{Confidence of (Canada \rightarrow USA)} = \frac{5}{6} = 83\% \]
\[ \text{Confidence of (USA \rightarrow Canada)} = \frac{5}{9} = 55\% \]

It can be clearly noticed that only three rules are strong and are as follows.

\[ \text{cpi} \rightarrow \text{USA} \]
\[ \text{Canada} \rightarrow \text{USA} \]
\[ \text{USA} \rightarrow \text{Canada} \]

IV. EXPERIMENTAL ANALYSIS

In this Section, experimental analysis has been performed for assessing the overall performance of proposed algorithm. Comparison have been performed between proposed algorithm and the existing algorithm [1].

The proposed algorithm has been tested and executed on the dataset derived from [20]. The algorithm has been implemented using MATLAB.

Experimental dataset contains 30 transactions and 10 items. Execution time graph between proposed algorithm and algorithm in [1] is shown in Fig. 7, Fig. 8, Fig. 9 and Fig. 10. In the graph X-axis indicate the various steps of algorithms and Y-axis indicate time in seconds. Figure 11 and Fig. 12 shows the memory utilization corresponding to the proposed PDISS approach and existing algorithm [1].
Fig. 7: min_sup=1, product discard=1, min_conf=0.1

Fig. 8: min_sup=2, product discard=1, min_conf=0.1

Fig. 9: min_sup=2, product discard=2, min_conf=0.1

Fig. 10: min_sup=3, product discard=3, min_conf=0.2
TABLE I: Execution time taken by algorithms at different support threshold values

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Products Discarded</th>
<th>Minimum Support (min_sup)</th>
<th>Minimum Confidence (min_conf)</th>
<th>Step</th>
<th>PDISS (Newly Developed Algorithm)</th>
<th>SOFTSET (Existing Algorithm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.001249s 0.00292s</td>
<td>0.00120s 0.00284s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0.00365s 0.171s</td>
<td>0.00361s 0.170s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0.07705s 0.333s</td>
<td>0.07705s 0.333s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.09195s 0.478s</td>
<td>0.09195s 0.478s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.1052s 0.556s</td>
<td>0.1052s 0.556s</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>0.0002s 0.00292s</td>
<td>0.0002s 0.00284s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0.0004s 0.00414s</td>
<td>0.0004s 0.00414s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0.0790s 0.242s</td>
<td>0.0790s 0.242s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.08233s 0.339s</td>
<td>0.08233s 0.339s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.1008s 0.518s</td>
<td>0.1008s 0.518s</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>0.0002303s 0.0003308s</td>
<td>0.0002303s 0.0003308s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0.004102s 0.00413s</td>
<td>0.004102s 0.00413s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0.01048s 0.02458s</td>
<td>0.01048s 0.02458s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.01355s 0.03747s</td>
<td>0.01355s 0.03747s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.01818s 0.04952s</td>
<td>0.01818s 0.04952s</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0.2</td>
<td>1</td>
<td>0.00049s 0.00025s</td>
<td>0.00049s 0.00025s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0.00279s 0.00418s</td>
<td>0.00279s 0.00418s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>0.00537s 0.01999s</td>
<td>0.00537s 0.01999s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>0.007337s 0.02679s</td>
<td>0.007337s 0.02679s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.009231s 0.03287s</td>
<td>0.009231s 0.03287s</td>
</tr>
</tbody>
</table>

TABLE I shows execution time taken by developed PDISS Algorithm and the existing Algorithm (Soft Set Algorithm). It has been concluded that the developed PDISS is better and efficient in terms of execution time.

Now the bar graph shown below gives the clear understanding of how memory is used in case of both the approaches. First graph (Fig. 11) is for system memory in which red color indicates soft set approach i.e. existing algorithm and pink color indicates the memory utilization for newly developed PDISS algorithm.

Similarly, Fig. 12 shows virtual memory utilization. In the Figure brown color indicates soft set approach i.e. existing algorithm and yellow color shows the memory usage for newly developed PDISS algorithm.

TABLE II shows the memory usage and compares the memory requirements of both the algorithms. It can be clearly noticed that the newly developed algorithm requires less memory.

<table>
<thead>
<tr>
<th>S. NO.</th>
<th>ALGORITHM</th>
<th>SYSTEM MEMORY USED</th>
<th>VIRTUAL MEMORY USED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PDISS (newly developed algorithm)</td>
<td>3.00E+08</td>
<td>1.01E+09</td>
</tr>
<tr>
<td>2</td>
<td>SOFTSET (existing algorithm)</td>
<td>3.87E+08</td>
<td>1.59E+09</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This paper proposes a novel algorithm called PDISS for the generation of strong association rules containing PDIs. Proposed algorithm is suitable for situations where uncertainty is associated. In the initial stage uncertainty was present in the considered dataset which causes some difficulty in the result calculation. Therefore, firstly with the help of soft set uncertainty is handled and then use of additional constraints helps in the identification of items that are not present in PDI set. Due to this only, the item that is important or present in PDI remains in the dataset. Further, with the help of parameter co-occurrence, support of various elements is calculated and frequent itemsets are generated. After that strong association rules are generated with the help of minimum confidence threshold. It has been concluded that the proposed algorithm takes less amount of time and memory space than the existing one.
REFERENCES


