A Soft Set Model for Interesting Rules: A Case Study on Post Operative Patient Data

Satya Ranjan Dash
School of Computer Application
KIIT University
Bhubaneswar, India

Susil Rayaguru
School of Computer Application
KIIT University
Bhubaneswar, India

Satchidananda Dehuri
Department of Systems Engineering
Ajou University,South Korea

ABSTRACT

The proposed model is mining the interesting association rules based on soft set theory. We have introduced a threshold function in the aforesaid model to eliminate the user defined threshold value for minimum support and confidence to discover interesting rules. In addition, a preprocessing step has been carried out for transforming the quantitative data into Boolean-valued data i.e., all entries of the dataset is holding either a value 0 or 1. This method is validated through a case study on postoperative patient data retrieved from UCI machine learning repository.

Keywords

Soft Set, Association Rule Mining, Interesting Rule.

1. INTRODUCTION

Soft set theory [1], is to deal with uncertain data. It is otherwise known as (binary, basic, elementary) neighborhood system. There are other methods like fuzzy sets that can also be applied with uncertain data but it introduces the inadequacy of parameterization tool [7], where as no such parameterization issues present with soft set approach.

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. The soft set is a parameterized family of subsets of the set U. Every set F(e), e $\in$ E, from this family 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. Soft set can be redefined as the classification of objects in two distinct classes; hence a soft set (F, E) over the universe U can be represented as a Boolean value information system. Soft set can be applied for finding interesting association rules from a market basket like databases [6].

Association rule mining is one of the popular and useful tools in data mining, used to extract interesting correlations, frequent patterns and/or associations among set of items in the transaction database. Due to its high degree of implementation/ applications in areas such as telecom networks, risk management, inventory control, etc., it has been realized as one of the well researched techniques of data mining. ARM uses two user defined threshold value as minimum support and confidence for finding interesting association rule [10].

Here, we have proposed a soft set model which can mine interesting association rules. We have introduced a threshold function to eliminate the user specified threshold value such as minimum support and confidence which is inevitable with ARM.

The rest of the paper is organized as follows. Section 2 discusses the preliminaries of this work like fundamental concept of association rule mining and soft set theory. Section 3 describes about transformation of a post operative data set to a Boolean value information system that soft set can consume directly. Section 4 describes about soft set model with threshold function for mining interesting association rules. Section 5 includes conclusion and future work.

2. PRELIMINARIES

Association rule mining has been the research area for many practitioners for many years [10, 13, 14, 15]. The approach which widely used is the support-confidence framework [13], where an association rule is an implication between two sets of items, and the interestingness of a rule is measured by two factors: its support and confidence. An association rule is about relationships between two disjoint item sets X and Y. The statement $X \Longrightarrow Y$ implies the pattern when X occurs, Y occurs.

Support and confidence are two measures of how interesting a rule is. A support of 2% says that out of all transactions, 2% show that X and Y are bought together. Whereas, a confidence of 70% says that 70% of customers who purchased X also bought Y. Let I = $\{I_{1,}\ I_{2,\dots,}I_m\}$ be a set of items. Let D be a database where each transaction T is a set of items such that $T \subseteq I$. The rule $X \Longrightarrow Y$ has a support *s* and confidence *c*, where *s* is the probability of transactions in D containing X $\cap$ Y and *c* is the probability of transactions in D containing X that also contain Y. A detailed analysis can be found in [2]**.**

When we apply ARM over a large dataset it can bring forth a large number of associations but all the association are not interesting. Hence for finding interesting association ARM uses a user specified threshold value as minimum support and confidence [2]. However in the real world scenario while mining real world application, mining different data base required different values of minimum support.

Therefore it becomes a major challenge for the user to specify the parameter as minimum support for their data base to find out the interesting rules and that is why we have proposed a threshold function to guess the interesting item set.

Soft set theory [1], proposed by Molodtsov in 1999, is a method for dealing with uncertain data. It can classify the objects into two distinct classes, thus confirming the subset can deal with Boolean valued information system. Molodtsov [1] pointed out that the main advantage of soft set theory from the previous theories is that it is free from inadequacy of parameterization tools, unlike in the theories of fuzzy sets.

A soft set over U is a parameterized family of subsets of the universe U. For e ϵ 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 U = $\{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$ be a set of patients under treatment for the following set of diseases

E = $\{e_1,e_2,e_3,e_4,e_5,e_6\}$

e_1 stands for low blood pressure patients,

e_2 stands for diabetes patients,

e_3 stands for psychiatric patients,

e_4 stands for back pain patients,

e_5 stands for cancer patients,

e_6 stands for common cold patients.

Now we can bring out the subsets of U who are satisfying above parameter $e_{i..n}$ as

$F(e_1) = \{p_1,p_5\}$, $F(e_2) = \{p_3,p_4\}$, $F(e_3) = \{p_2,p_6\}$,

$F(e_4) = \{p_7,p_5\}$, $F(e_5) = \{p_2,p_1\}$, $F(e_6) = \{p_3,p_7\}$, $F(e_7) = \{p_2,p_7\}$

In the aforesaid mapping function $F(e_1)$ maps to those patient's record who has got the symptom of low blood pressure. In this way we can view the soft set as a collection of approximation as listed in Figure 1.

$$(F, E) = \begin{cases} \text{Low blood pressure patients} = \{p_1, p_5\} \\ \text{Diabetes patients} = \{p_3, p_4\} \\ \text{Psychiatric patients} = \{p_2, p_6\} \\ \text{Back pain patients} = \{p_2, p_6\} \\ \text{Cancer patients} = \{p_2, p_1\} \\ \text{Common cold patients} = \{p_2, p_7\} \end{cases}$$

Fig. 1 Soft set representation of data set

The above soft set can be viewed as a collection of approximation as

$$(F, E) = \{p_1=v_1, p_2=v_{2,\ldots,}p_n=v_n\},$$

where p_i, i=1, 2, .., n, stand for set of predicate like diseases and v_i, i=1, 2, .., n, for approximation value set. Now we can see the soft set is the mapping from parameter to the crisp subset of the universe, hence a soft set can classify the objects into two classes (yes/1, no/0). Thus we can make one to one correspondence between a soft set and Boolean value information system as shown in Table 2.

Recall that the theory of soft set is introduced by Molodtsov [1]. This theory is a relatively new approach to discuss vagueness. It is getting popularity among the researchers and a good number of papers are being published every year. In [4] Maji and Roy discussed theoretical aspect of soft sets and they introduced several operations for soft sets. In soft set theory membership is decided by adequate parameters, rough set theory employs equivalence classes, whereas fuzzy set theory depends upon grade of membership.

Let U be an initial universe set and let E be a set of parameters. 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(i.e. F: E→P (U)).

In other words, the soft set is a parameterized family of subsets of the set *U*. Every set $F(\mathcal{E})$, $\mathcal{E} \in$ E, from this family may be considered as the set of $\mathcal{E}$-elements of the soft set (*F*,*E*), or as the set of $\mathcal{E}$-approximate elements of the soft set.

Assume that we have a binary operation, denoted by $*$, for subsets of the set *U*. Let (*F*, *A*) and (*G*, *B*) be soft sets over *U*. Then, the operation $*$ for soft sets is defined in the following way:

$(F, A) * (G, B) = (H, A \times B)$,

where $H(\alpha, \beta) = F(\alpha) * G(\beta)$, $\alpha \in A$ and $\beta \in B$.

This definition takes into account the individual nature of any soft set. If we produce a lot of operations with soft sets, the result will be a soft set with a very wide set of parameters. Sometimes such expansion of the set of parameters may be useful.

3. TRANSFORMATION OF A DATASET INTO BOOLEAN INFORMATION SYSTEM

A dataset can be represented as a relation in a relational database. Entities in relational databases are represented by tuples of attribute values, similarly data set can have tuples as S = {U,A,V, *f* }, where U=$\{u_1, u_{2,} u_{3,} u_{4,\ldots,} u_{|u|}\}$ is non empty set of objects, A=$\{a_{1,} a_{2,} a_{3,\ldots,} a_n\}$ is a non empty finite set of attributes, V=U_a ϵ A ⟶ V_a, V_a is the domain value of attribute a, *f*: U×A ⟶V is the dataset function such that (u,a) ∈V_a, for every (u,a)∈ U×A. Here each Object u_i can be represented as

$u_i = f(u_i,a_1), f(u_i,a_2), f(u_i,a_3),\ldots,f(u_i,a_{|A|})$

where i=1,2,3,…,|U|.

In a data set if S = {U,A,V, *f* }, if V_a ={0,1} for every a∈ A then S is called a Boolean-valued information system.

In this paper, we have considered post operative data set [16].

3.1 Transforming post operative patient data into soft set acceptable data

Let I = $\{i_1, i_{2,} i_3\ i_4\ i_{5\ldots\ldots,}\ i_n\}$ be a set of attributes (patient symptoms) and D = $\{t_1, t_{2,}\ldots.t_{|u|}\}$ be a set of patient record in post operative data set. For the Boolean - valued information system we have the following transformation S = {U, A, $V_{(1,0)}$,f} we have the following transformation.

For every a∈ A and u∈U, we have the mapping function *f*: U ×A→ {0,1} such that *f*(u,a)=1 if a appears in t, otherwise *f*(u,a)=0. In figure 2 we have the representation of post operative data set to soft set.

$i_1 \longrightarrow a_1$

$i_2 \longrightarrow a_2$

$i_3 \longrightarrow a_3 \Leftrightarrow \{i_1, i2_{,}\ldots,i_{|A|}\} \longrightarrow A = \{a_1, a_{2,} a_{3,\ldots,} a_{|A|}\}$

……

. .. . . .

$i_{|A|} \longrightarrow a_{|A|}$

$t_1 \longrightarrow u_1$

$t_2 \longrightarrow u_2$

$t_3 \longrightarrow u_3 \Leftrightarrow D= \{t_1,t_2,\ldots,t_{|u|}\} \longrightarrow U= \{u_1, u_2, u_{3,\ldots,} u_{|U|}\}$

. . . . ..

. .. . . .

$t_{|U|} \longrightarrow u_{|U|}$

Fig. 2 Transformation of post operative data set to Boolean valued information system.

3.2 Atrribute Information of Post Operative dataset

1. L-CORE (patient's internal temperature in Celsius): internal temperature high: ihigh (> 37), internal temperature mid: imid

(>= 36 and <= 37), internal temperature low: ilow (< 36)

2. L-SURF (patient's surface temperature in Celsius): surface temperature high: shigh (> 36.5), surface temperature mid: smid

(>= 36.5 and <= 35), surface temperature low: slow (< 35)

3. L-O2 (oxygen saturation in %): excellent (>= 98):oe , good

(>= 90 and < 98):og, fair (>= 80 and < 90):of, poor (< 80):op

4. L-BP (last measurement of blood pressure): blood pressure high

(> 130/90):bph, blood pressure mid (<= 130/90 and >= 90/70):bpm, blood pressure low (< 90/70):bpl

5. SURF-STBL (stability of patient's surface temperature): surface temperature stable: sstbl, med surface temperature: smstbl, surface temperature unstable: sustbl

6. CORE-STBL (stability of patient's core temperature) core temperature stable: cstbl, core temperature mod-stable: cmstbl, core temperature unstable: custbl

7. BP-STBL (stability of patient's blood pressure) blood pressure stable: bpstbl, blood pressure mod-stable: bpmstbl, blood pressure unstable: bpustbl.

Table 1: Instances of post operative data set.

Sl	L-CORE	L-SURF	L-02	L-BP	SURF-STBL	CORE-STBL	BP-STBL	Decision
1	imid	Slow	oe	bmp	sstbl	cstbl	bpstbl	A
2	imid	Shigh	oe	bph	sstbl	cstbl	bpstbl	S
3	imid	Slow	og	bpm	sstbl	cstbl	bpustbl	I
4	imid	smid	oe	bpm	sustbl	custbl	bpstbl	S
5	imid	smid	oe	bpm	sstbl	cstbl	bpmstbl	S
6	ilow	smid	oe	bph	sustbl	cstbl	bpustbl	S
7	imid	smid	oe	bph	sstbl	cstbl	bpstbl	A
8	imid	smid	og	bpm	sstbl	cstbl	bpstbl	A
9	imid	slow	og	bph	sustbl	custbl	bpstbl	S
10	ihigh	shigh	oe	bph	sustbl	cstbl	bpustbl	A

Decision to infer considering the above post operative patient record as below

I (patient sent to Intensive Care Unit),
S (patient prepared to go Home),
A (patient sent to general Hospital floor)

As explained earlier, soft set is a mapping from parameter to the crisp subset of universe. From such case we may see the structure of a soft set can classify the objects into two classes (yes/1, no/0). Thus we can construct a Boolean valued information system where attributes can take value as 0 or 1. The Boolean representation of Table-1 is depicted in Table-2.

Table 2: Conversion of post operative data set to Boolean information system

S L	L-CORE			L-SURF			L-02				L-BP			SURF-STBL			CORE-STBL			BP-STBL			DEC		
	L	M	H	L	M	H	E	G	F	P	L	M	H	S	M	U	S	M	U	S	M	U	I	S	A
1	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1
2	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1	0
3	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	1	0
5	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0
6	1	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0
7	0	1	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1
8	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1
9	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0
10	0	0	1	0	0	1	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	1

3.3 Subsequent Soft Set Theory for Association Rule Mining

For Soft set theory can be applied for mining interesting association rules in a data set. In this process data set is first converted into a soft set then it is transferred to a Boolean information system. Figure 3 is the schematic representation of the soft set approach for mining interesting rules [12].

Soft set (F, E) over the universe U representing a Boolean valued information system $S=(U,A,V_{\{0,1\}},f)$ from a data set $D = \{t_1, t_{2,} t_{3\ldots,}t_{|U|}\}$, an items co-occurrence set in a transaction u can be defined as

$\text{Coo}(u) = \{ e \in E : f(u,e) = 1\}$.

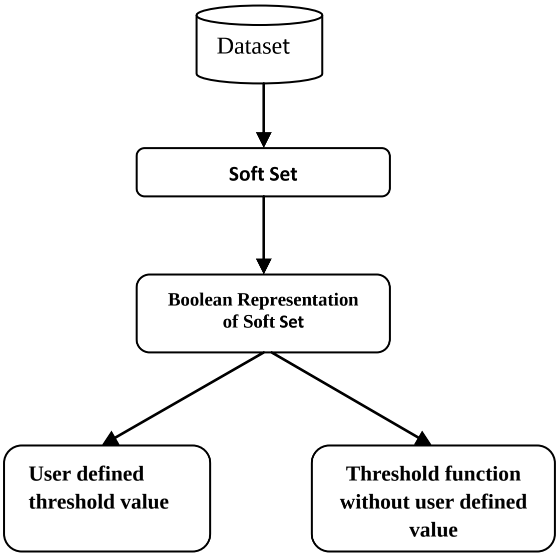
Dataset
Soft Set
Boolean Representation of Soft Set
User defined threshold value
Threshold function without user defined value

Fig 3. Mining interesting rules soft set.

The soft set representation of Table 1 is given below with the co-occurrence of each data set instances.

$$(F,E) = \left\{ \begin{array}{l} \text{imid} = \{ 1,2,3,4,5,8,7,9\};\ \text{ilow}= \{6\}; \\ \text{ihigh}=\{10\};\ \text{smid}=\{4,5,6,7,8,10\}; \\ \text{slow}=\{1,3\};\ \text{shigh}=\{2,10\}; \\ \text{oe} = \{1,2,4,5,6,7,10\};\ \text{og} =\{3,8,9\}; \\ \text{bmp}=\{1,3,4,5,8\};\ \text{bph}=\{2,6,7,,910\}; \\ \text{sstbl} =\{1,2,3,5,7,8\};\ \text{sustbl}=\{4,6,9,10\}; \\ \text{bpstbl}=\{1,2,4,7,8,9\};\ \text{bpustbl} =\{3,6,10\}; \\ \text{bpmstbl}=\{5\} \end{array} \right.$$

Fig. 4: Soft Set representation of Table 1.

Coo(u_1) = { imid,slow,oe,bmp,sstbl,cstbl,bpstbl},
Coo(u_2)={ imid,shigh,oe,bph,sstbl,cstbl,bpstbl},
Coo(u_3)={ imid,slow,og,bpm,sstbl,cstbl,bpustbl },
Coo(u_4)={ imid,smid,oe,bpm,sustbl,custbl,bpstbl }
Coo(u_5)={ imid,smid,oe,bpm,sstbl,cstbl,bpmstbl },
Coo(u_6)={ ilow,smid,oe,bph,sustbl,cstbl,bpustbl }
Coo(u_7)={ imid,smid,oe,bph,sstbl,cstbl,bpstbl },
Coo(u_8)={ imid,smid,og,bpm,sstbl,cstbl,bpstbl },
Coo(u_9)={ imid,slow,og,bph,sustbl,custbl,bpstbl},
Coo(u_{10})={ ihigh,shigh,oe,bph,sustbl,cstbl,bpustbl }

Fig.5 The co-occurrence of disease symptom in the data set.

Let *(F, E)* be the soft set over the universe U and X, Y $\subseteq$ E, where X $\cap$ Y= ϕ . An association rule between X and Y is an implication of the form X $\longrightarrow$ Y. The item set X is called antecedent and Y is called consequent. The support of association X $\longrightarrow$ Y denoted by

$$\sup(X \longrightarrow Y) = \sup(X \cup Y) = |\{ u: X \cup Y \subseteq \text{Coo}(u)\}|.$$

The confidence of an association rule X $\longrightarrow$ Y denoted respectively by conf (X $\longrightarrow$ Y) and it can be defined as:

$$\text{conf}(X \longrightarrow Y) = \frac{|\{u : X \cup Y \subseteq Coo(u)\}|}{|\{u : X \subseteq Coo(u)\}|}\ .$$

Here the model requires the user specified threshold value, minimum support and confidence to mine interesting rules. As stated earlier it becomes a major challenge for the user to specify the parameters precisely and that is why we may lose some interesting association rules. Hence to circumvent this problem we have specified a threshold function which can computer these threshold value without user intervention.

3.4 Threshold Function to Compute Minimum Support and Confidence

In this model we have considered positive association rules where *conf*(X $\longrightarrow$ Y) should be larger than, or equal to, *supp(Y).*

Positive confidence rules can be defined as

$$pconf(X \longrightarrow Y) = \frac{\sup(X \cup Y) - \sup(X) - \sup(Y)}{\sup(X)(1 - \sup(Y))}$$

When *1- sup(Y)* = 0 or *sup(X)* =0, this definition is of no use. It can be easily observed that *sup(X $\cup$ Y) = sup (X)* for an item set X if *1- sup(Y)* = 0. If *sup(x)* =0, then *sup(X $\cup$ Y)* = 0 for any item set Y. We can also define that *pconf*(X $\longrightarrow$ Y) = 0 when *sup(x)* = 0. In both cases, rule X $\longrightarrow$ Y is called a trivial rule.

If the above equation earns a positive value for the association under consideration, then is presumed to be interesting.

We have tested the aforesaid soft set model with the post operative patient data set using the threshold function and user specified threshold parameter. Below in "Table 4" is the comparative study for finding interesting association rules with respect to both the methodology.

Table 4: Interesting Rules

Applied Method	Minimum support	Interesting Rules	Decision Infer
Threshold Function	**NA**	**13**	**A**
User specified	**0.8**	**5**	**A**
Threshold Function	**NA**	**13**	**A**
User specified	**0.4**	**9**	**S**

Item set having same support can be put in the one class, so we have 3 clusters, and we can cluster the whole dataset to show the patients belong to which clusters.(i.e., I,A or S).

4. CONCLUSION

In this paper, we have presented a soft approach for association rules from a transactional data set. This approach is started by a transformation of a post operative data set into a date set that can directly consume soft set. Using the concept of parameters co-occurrence in the transaction by user defined threshold value, it fails to discover interesting association rule. To circumvent this problem, we have introduced a threshold function to eliminate the user defined threshold value for minimum support and confidence to find out the interesting rules. The experimental result confirms the superiority of our model.

5. REFERENCES

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

[2] Han, J., Kamber, M. 2010. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, ISBN 978-81-312-0535-8

[3] Herawan, T., Mat Deris, M. 2011. A soft set approach of association rule mining, Knowledge-Based System, vol 24, 1, pp. 186-195.

[4] Maji, P.K., Roy, A. R. 2002. An application of soft sets in a decision making problem, volume 44, Issues 8–9, , Pages 1077–1083.

[5] Noda, E., Freitas, A. A. and Lopes, H. S. 1999 "Discovering interesting prediction rules with a genetic algorithm", proc.IEEE Congr. Evolutionary Comput. CEC '99, pp. 1322-1329.

[6] Cheung, Y. and Fu, A. 2004. “Mining frequent itemsets without support threshold: With and without item constraints,” IEEE Transactions on Knowledge and Data Engineering.

[7] Zhang, S., Lu, J. and Zhang, C.2004. “A fuzzy-logic-based method to acquire user threshold of minimum-support for mining association rules”, Information Sciences.

[8] Feldman, R., Aumann, Y., Amir, A., Zilberstein, A. and Klosgen, W.1999. “Maximal association rules: a new tool for mining for keywords co-occurrences in document collections,” in: The Proceedings of the KDD pp. 167–170.

[9] Guan, J.W., Bell, D.A. and Liu, D.Y.2005. The rough set approach to association rule mining, in: The Proceedings of the Third IEEE International Conference on Data Mining (ICDM’03), pp. 529–532.

[10] Han, J., Pei, J. and Yin, Y. 2000. Mining frequent patterns without candidate generation. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, pp 1–12.

[11] Park, J., Chen, M. and Yu, P.1997. Using a hash–based method with transaction trimming for mining association rules. IEEE Trans., Knowledge and Data Eng. 9(5): pp. 813–824.

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

[13] Agrawal, R., Imielinski, T. and Swami, A..1993. “Mining association rules between sets of items in large databases”, In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207–216.

[14] Au, W. and Chan, C.2002. “An evolutionary approach for discovering changing patterns in historical data” In Proceedings , SPIE, pp. 398–409.

[15] Silverstein, C., Brin, S. and Motwani, R. 1998.“Beyond market baskets: Generalizing association rules to dependence rules”, Data Mining and Knowledge Discovery, pp. 39–68.

[16] Post Operative Dataset http://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient