Optimization of Association Rules for Students' Data: A Soft Set Approach

Sasanko Sekhar Gantayat

Abstract*: Association rule (AR) mining is a common method to find the associations between objects in a transaction or in a data set, which is simple logical rules, determine the critical relationship among the objects. The rule generation is difficult if it satisfies a predefined threshold value. To trace the relationship is very important for domains to reveal crucial and hidden information from the data set. Some critical relationship of the objects may appear rarely and to trace them using classical methods are difficult one. The apriori algorithm (AA) with hashing technique along with the PDI algorithm is used to find the appropriate ARs from the students' dataset. Since soft set is a tool for handling imprecise parameterized data, the soft set concept is used in optimization process to get optimized ARs. In this paper, the ARs are optimized using soft set for further analysis.

***Index Terms*: Association Rule(AR), Data Mining, Hash based Apriori Algorithm(AA), , PDI, Soft Sets**

I. INTRODUCTION

Now a day, a huge amount of data is generated on different application domains in real-time. The traditional statistical techniques and data base management tools are not suitable for analysing this huge amount of data. These data can be analysed and summarised by data mining techniques to acquire the useful information for the decision making process. One of the evolving interdisciplinary research area in educational context is educational data mining. Educational data mining is a process to apply data mining techniques to student data to find the performance of the students in an institution. The primary concern on the educational data mining is to develop methods that can discover the interesting information extracted from educational settings, and work on these methods for the better understanding of the students, and the backgrounds in which they learn.

Association Rule (AR) Mining is a common and popular technique used in data mining application to find the association between the various objects or attributes. The concept of AR mining was first introduced by Agrawal et al. in1993 as known as market-basket analysis [1, 2, 3, 6]. There are two main steps involved during the AR generation. First one is to find all frequent items (FI) from transaction data and in the second step to generate the common ARs for the frequent items. A FI set occurs more than a predefined minimum threshold value. The supports of the item set are defined as a probability of item set that occurs in the transaction. Besides this, confidence is another alternative measurement used for pair the item sets in ARs. This confidence is defined as the probability of generating strong rules if it meets some minimum threshold value. Association is to identify the association relationships or correlations in a set of items of transactions. Association analysis is widely used in any type of data which is available in real environment for decision making processes. Now days, many educational institutions are implementing this technique, directly or indirectly, to give a quality education for the students. Due to the lack of knowledge and useful information on the part of the management about students', it is difficult to identify the students who are not achieving the standard and qualitative objective. Data mining methods includes a great deal of assignments that can be utilized proficiently to examine the exhibition and accomplishments of the understudies. Because of expanding rivalry the understudies are required to be evaluated on various parameters for which data mining methods can be convenient to actualize. Instructive Data Mining is a rising field which identifies with growing new techniques to break down data on understudies and staff data. Association analysis can be used to study this types of concepts. In this present paper, the students' data onto higher education environment is studied and analysed. Educational information mining, which concentrates valuable, already obscure examples of educational databases to improve and upgrade the learner's presentation and improve educational quality.

II. LITERATURE REVIEW

The point of this area is to give some survey of the past research endeavors completed to arrangement of graduates' exhibition. This segment indicates different existing proposed works in the territory of ieducational information mining. M. Ramaswami et.al.(2010) tested the exhibition of graduates, which relies upon various variables like scholastic, statistic, mental, socio-financial aspects and different components [12]. In light of these variables they developed their logical procedure to distinguish the distinctive execution components of graduates utilizing ARs. In 2011, J. Mamcenko et.al. built up a database on inquiries and answers from the graduates [9]. At that point they connected affiliation mining systems to find or recognize diverse learning examples of each graduate.

2.1. Soft Set

The soft sets concept is given by Molodtsov in 1999 as a new concept for taking care of issues handling with uncertainties [10]. He defined that a soft set is a attribute family of subsets of a universal set, in which each object is considered as a set of approximate objects of the set.

Later, the soft set theory have been studied by various researchers in different ways. Maji et al. in 2003, presented

Revised Manuscript Received on July 10, 2019.

Sasanko Sekhar Gantayat, Department of Computer Science and Engineering GMR Institute of Technology, Rajam, Andhra Pradesh, India.

a detailed theoretical study of soft sets which includes subset and superset of a soft set, equality of soft sets, operations on soft sets such as union, intersection, AND and OR operations etc. and discussed the basic properties on these operations [8].

Definition. Let U is an universal set and A is a set of attributes with respect to U. Let the power set of U is P(U) and $C \subseteq A$. A pair (α, C) is a soft set over U with a mapping given by:

$$\alpha: C \rightarrow P(U) \quad (1)$$

i.e., a soft set (α, C) over U is an attribute family of subsets of U. For, $x \in C$, $\alpha(x)$ is the set of a-elements of the soft sets (α, B) and (α, C) is defined as

$$(\alpha, C) = \{\alpha(x) \in P(U): x \in A,\ \alpha(x) = \phi \text{ if } x \notin C\} \quad (2)$$

2.2. AR Mining

AR mining is to find frequent patterns, associations, correlations, among the sets of objects in databases. The rules are generated to predict the occurrence of an item based on the occurrences of other items in the transaction [1, 6].

Let $A=\{a_1, a_2, a_3. . .\}$ is a set of items and $B=\{b_1, b_2, b_3. . .\}$ is a set of transactions, each of which contains items of the itemset *A*. Thus, each transaction b_i is a set of items such that b_i forms a subset *A*.

A rule R takes the form $\alpha \rightarrow \beta$

i. $\alpha, \beta \subseteq A$
ii. $\alpha \cap \beta = \emptyset$
iii. α and β are called itemsets.
iv. α is the rule's antecedent (left-hand side)
v. β is the rule's consequent (right-hand side)

Support: The support of the rule R: $\alpha \rightarrow \beta$ is defines as:

$$sup(\alpha \rightarrow \beta) = |\alpha \cap \beta| / |B|$$

Confidence: The confidence of a rule R: $\alpha \rightarrow \beta$ is defined as:

$$conf(\alpha \rightarrow \beta) = |\alpha \cap \beta| / |\alpha|$$

2.3. AA and PCY Algorithm

Numerous algorithms are available in the literature to improve the proficiency of incessant examples and AR revelation. A large portion of the algorithms, for example, Apriori, FP-tree, partitioning and sampling calculation are created and discovered some great yields. In example mining calculation center around either FI generation or finding the ARs from the itemset. Apriori gives answers for the two issues. A few calculations are utilized for parallel ARs, yet they are similarly work with quantitative ARs.

2.4. Apriori Algorithm (AA)

The apriori calculation [1, 2, 3, 6] was the main endeavor to mine ARs from a huge dataset. The calculation can be utilized for discovering continuous examples and further more getting ARs from these examples. The basic principle of Apriori is that the subset of a FI is also frequent. Generally FIs are used to find ARs. There are two steps used in the apriori, these are Join and Prune steps.

2.5. PCY Algorithm ([14])

In this present paper, hash-based itemset counting algorithm is implemented which is developed by J. Park, M. Chen, and P. Yu, 1995 [11]. PCY algorithm improves the performance of the apriori algorithm. In Pass_1 of AA, some of the memory are idle. The memory is used to keep bucket_counts with hashed pairs of items with counts where candidate pairs must satisfy on Pass_2.

Pass_1: It uses the following steps.

For each basket
    For each item, update the count;
    For each pair:
        Use hashing pairwise to a bucket and increment the bucket_count.

Intermediate_Passes: It uses the following steps.

i) The bucket is substituted by a bit-vector

ii) bit-vector = 1 if the bucket_count $\geq$ the support; 0 if it did not exceeded the support.

iii) Integers are substituted by bits, so the bit-vector uses a small space, and identify the frequent items for second pass.

Pass_2: The Pass_2 uses the following steps.

Count for all pairs {i, j} of the frequent Itemsets with the following conditions:

1. The pair {i, j}, generates a hash number as a bucket number with the bit-vector 1.

Multistage Algorithm: It uses the following steps.

i) After Pass_1, only the qualify pairs for Pass_2 are rehashed.

ii) In intermediate_pass, the left out pairs in the buckets, are the false positives, i.e. frequent buckets with no frequent pair.

In Pass_3: It uses the following steps.

1. Only those pairs {i, j} satisfied are counted where i and j are Itemsets.
2. The first hash function is used for the pair hashes in a bucket whose first bit-vector is 1.
3. The second hash function is used for the pair hashes to a bucket whose second bit-vector is 1.

III. METHODOLOGY AND EXPERIMENTAL WORK

In this paper, the analysis of the student data is carried out using the most popular data mining tool Weka and the Microsoft Excel. Weka is selected as it is easy to analyse data and its visual interpretation. The architecture of the proposed model is given in figure 1.

The experimental setup consists of:

(a) Data Collection (b) Pre-processing (c) Visualization of Data (d) AR Mining by hash based AA (e) Frequent Itemsets Generation (f) Classifying the Rules According to Condition (AND/OR/AND-OR) (g) Optimizing the Decision Rules Using Soft Set.

3.1, Data Collection

The data is collected from the students with the interaction with them. From which the training set consists of 50 students details with their attributes. The following describes the attributes of the collected data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Attribute Name	Description	Values
PSM	Previous Sem Marks	{First >=60%, Second >45 % and <60%, Third >36 % and <45%, Fail <36%}
CTG	Class-Test Grade	{Poor, Average, Good}
SEM	Seminar Presentation	{Poor, Average, Good}
ASS	Assignment	{Yes, No}
GP	General Proficiency	{Yes, No}
ATT	Attendance	{Poor, Average, Good}
LW	Lab Work	{Yes, No}
ESM	End Semester Marks	{First >=60%, Second >45 % and <60%, Third >36 % and <45%, Fail <36%}

Table 1. Description of the data

As a component of the data preparation and pre-processing of the data set and to show signs of improvement input data for data mining procedures, some pre-processing steps have accomplished for the gathered data before stacking the data set to the data mining programming, unimportant qualities are expelled. Utilizing Weka the data is standardization.

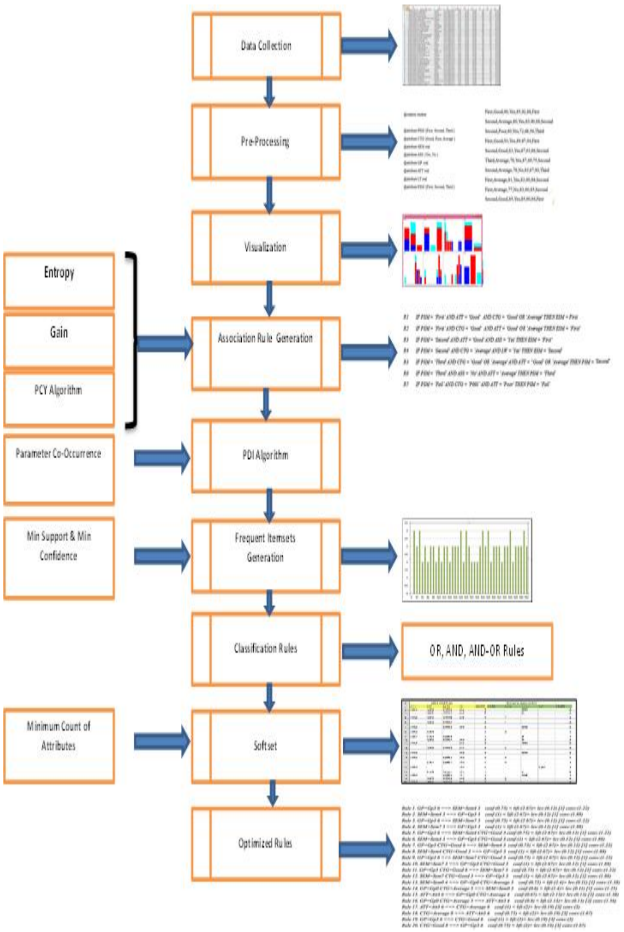

Fig. 1 Architecture of the model

Now we get the minimum, maximum, mean and the standard deviation of each numeric attribute. The data which is pre-processed is taken into another file and that can be visualized in fig 2.

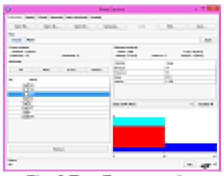

Fig. 2 Data Pre-processing

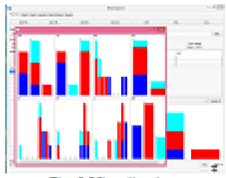

Fig. 3 Visualization

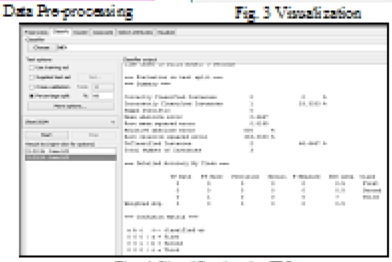

Fig. 4 Classification by ID3

3.2. Pre-Processing of the Data

The minimum support is considered is 0.15 for 8 attributes and the minimum confidence is 0.9 with the number of cycles 17. Generated sets of large Itemsets are: L(1) is 16, L(2) is 29, L(3) is 12 and L(4) is 2.

@relation student

@attribute PSM {First, Second, Third }
@attribute CTG {Good, Poor, Average }
@attribute SEM real
@attribute ASS {Yes, No }
@attribute GP real
@attribute ATT real
@attribute LT real
@attribute ESM {First, Second, Third }

First,Good,90,Yes,85,92,88,First
Second,Average,80,Yes,83,90,88,Second
Second,Poor,80,Yes,72,68,56,Third
First,Good,93,Yes,89,87,94,First
Second,Good,83,Yes,87,93,86,Second
Third,Average,78,Yes,87,69,75,Second
Second,Average,78,No,83,87,80,Third
First,Average,81,Yes,83,90,88,Second
First,Average,77,No,83,90,85,Second
Second,Good,85,Yes,85,90,94,First

Fig. 5 Data Table

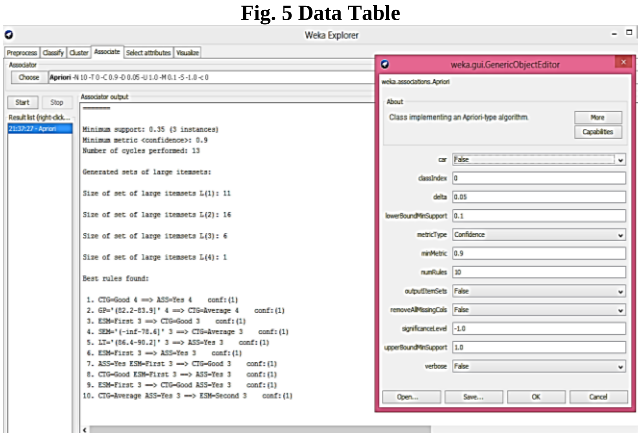

Fig. 6 Rules Generated by Weka

The interpretation of the generated rules for different confidence values shows the students' performance in various parameters or attributes. From the outcomes, it very well may be deciphered that to get the great execution of understudy they ought to have great participation and class test marks. Again it is seen that the last execution of an understudy is influenced by the understudy's class test marks.

3.3. Rule Generation

The following rules are generated from the AA.

R1 IF PSM = 'First' AND ATT = 'Good' AND CTG = 'Good' OR 'Average' THEN ESM = First

R2 IF PSM = 'First' AND CTG = 'Good' AND ATT = 'Good' OR 'Average' THEN ESM = 'First'

R3 IF PSM = 'Second' AND ATT = 'Good' AND ASS = 'Yes' THEN ESM = 'First'

R4 IF PSM = 'Second' AND CTG = 'Average' AND LW = 'Yes' THEN ESM = 'Second'

R5 IF PSM = 'Third' AND CTG = 'Good' OR 'Average' AND ATT = "Good' OR 'Average' THEN PSM = 'Second'

R6 IF PSM = 'Third' AND ASS = 'No' AND ATT = 'Average' THEN PSM = 'Third'

R7 IF PSM = 'Fail' AND CTG = 'P001' AND ATT = 'Poor' THEN PSM = 'Fail'

Entropy (E)

E(S)

$= -p_{First}\log_2(p_{First}) - p_{Second}\log_2(p_{Second}) - p_{third}\log_2(p_{third}) - p_{Fail}\log_2(p_{Fail})$

$= -(14/50)\log_2(14/50) - (15/50)\log_2(15/50) - (13/50)\log_2(13/50) - (8/50)\log_2(8/50) = 1.964$

The Gain (G) value of the Each Attribute

$G(S, PSM) = E(S) - |S_{First}|\,E(S_{First})/|S| - |S_{Second}|\,E(S_{Second})/|S| - |S_{third}|\,E(S_{third})/|S| - |S_{Fail}|\,E(S_{Fail})/|S|$

Sl. No.	Rule Generated by Weka	Conf	Lift	Lev	Conv
1	if PSM =First then ESM=First	1	3.57	0.17	8.64
2	if PAR= second & ESM=second then PSM=second	1	2.5	0.12	6.01
3	if PAR= first & PSM=first then ESM=first	1	3.57	0.13	6.48
4	if MSM=good & ESM=first then PSM=first	1	4.17	0.12	6.08
5	if PSM= first & MSM=good then ESM=first	1	3.57	0.12	5.76
6	if ATT= poor & ESM=fail then PAR=third	1	3.57	0.1	5.04
7	if PSM=first & ATT=good then MSM=good	1	3.57	0.1	5.04
8	if PSM=first & ATT=good then ESM=first	1	3.57	0.1	5.04
9	if PAR= second & ESM=average & ESM=second then PSM=second	1	2.5	0.08	4.2
10	if ATT=good & MSM=good & ESM= first then PSM=first	1	4.17	0.11	5.32

Table 2 Rule Generation

IV. RESULTS AND DISCUSSIONS

After calculating those value the Weka tool generate the best rules which are useful with high confidence value is 1. The ars generated by Weka by using AA for finding the frequent item set are listed as below.

4.1. 'AND' Rule Generation by the Previous ARs

```
1.   ASS = Yes and ESM = Second ==> GP = '(85.6-87.3]' or LT = '(86.4-90.2]'
2.   CTG = Average and ESM = Second ==> ATT = '(88-90.5]' or LT = '(71.2-75]'
3.   CTG = Average and ESM = Second ==> ATT = '(88-90.5]' or PSM = Third
4.     CTG = Average and ATT = '(88-90.5]' ==> GP = '(82.2-83.9]'
5.    ASS = Yes and PSM = First ==> GP = '(87.3-inf)' or LT = '(86.4-90.2]'
6.   ASS = Yes and PSM = First ==> SEM = '(80.2-81.8]' or ESM = First
7.    ASS = Yes and PSM = First ==> SEM = '(91.4-inf)' or LT = '(86.4-90.2]'
8.     ASS = Yes and ATT = '(88-90.5]' ==> SEM = '(83.4-85]' or LT = '(86.4-90.2]'
9.    ATT = '(88-90.5]' and ESM = Second ==> GP = '(82.2-83.9]'
10.  GP = '(82.2-83.9]' and ESM = Second ==> ATT = '(88-90.5]'
11.   ASS = Yes and ATT = '(88-90.5]' ==> GP = '(83.9-85.6]' or LT = '(86.4-90.2]'
12.   ASS = Yes and ESM = Second ==> SEM = '(80.2-81.8]' or GP = '(85.6-87.3]'
13.    ASS = Yes and PSM = Second ==> CTG = Good or SEM = '(78.6-80.2]'
14.   CTG = Good and PSM = First ==> SEM = '(89.8-91.4]' or GP = '(87.3-inf)'
15.   ESM = First and PSM = First ==> SEM = '(89.8-91.4]' or GP = '(87.3-inf)'
16.   SEM = '(-inf-78.6]' and GP = '(82.2-83.9]' ==> ASS = No
17.   ASS = Yes and PSM = First ==> CTG = Good or SEM = '(80.2-81.8]'
18.   ASS = Yes and PSM = First ==> LT = '(86.4-90.2]' or ESM = First
19.   CTG = Average and ATT = '(88-90.5]' ==> ESM = Second
20.   ASS = Yes and ATT = '(88-90.5]' ==> LT = '(86.4-90.2]' or ESM = First
21.   CTG = Average and ATT = '(88-90.5]' ==> LT = '(86.4-90.2]' or PSM = First
22.   GP = '(82.2-83.9]' and ATT = '(88-90.5]' ==> LT = '(86.4-90.2]' or PSM = First
23.   GP = '(82.2-83.9]' and ESM = Second ==> LT = '(86.4-90.2]' or PSM = First
24.   ATT = '(88-90.5]' and ESM = Second ==> LT = '(86.4-90.2]' or PSM = First
25.  CTG = Average and ATT = '(88-90.5]' ==> ASS = No or LT = '(86.4-90.2]'
26.   CTG = Average and ASS = Yes ==> GP = '(85.6-87.3]' or LT = '(86.4-90.2]'
27.  CTG = Average and ASS = Yes ==> ATT = '(-inf-70.5]' or LT = '(86.4-90.2]'
28.  ATT = '(88-90.5]' and ESM = Second ==> CTG = Average
29.  GP = '(82.2-83.9]' and ATT = '(88-90.5]' ==> ESM = Second
30.   ATT = '(88-90.5]' and ESM = Second ==> ASS = No or LT = '(86.4-90.2]'
```

Fig. 7 'AND' Rules (Sample)

4.2. 'OR' Rules Generation by the Previous AR

The following rules are generated by the AA are not optimized. In-order to optimize these rules we have implemented the soft set technique for ARs. Before to minimize these ARs, we have implemented the Pre-Determined itemset (PDI) Algorithm [14].

```
1.    CTG = Average ==> GP = '(82.2-83.9]' or LT = '(71.2-75]'
2.    CTG = Average ==> GP = '(82.2-83.9]' or PSM = Third
3.    CTG = Average ==> SEM = '(-inf-78.6]' or GP = '(82.2-83.9]'
4.     CTG = Good ==> ATT = '(90.5-inf)' or LT = '(90.2-inf)'
5.     CTG = Good ==> ATT = '(90.5-inf)' or ESM = First
6.     CTG = Good ==> SEM = '(81.8-83.4]' or ESM = First
7.     CTG = Good ==> SEM = '(81.8-83.4]' or GP = '(83.9-85.6]' or LT = '(90.2-inf)'
8.     CTG = Good ==> SEM = '(83.4-85]' or GP = '(87.3-inf)' or ATT = '(90.5-inf)'
9.     CTG = Good ==> SEM = '(91.4-inf)' or GP = '(83.9-85.6]' or ATT = '(90.5-inf)'
10.    ESM = First ==> GP = '(83.9-85.6]' or LT = '(90.2-inf)'
11.    ESM = First ==> SEM = '(89.8-91.4]' or LT = '(90.2-inf)'
12.    ESM = First ==> SEM = '(91.4-inf)' or GP = '(83.9-85.6]'
13.    SEM = '(-inf-78.6]' ==> ASS = No or LT = '(71.2-75]'
14.    SEM = '(-inf-78.6]' ==> ASS = No or PSM = Third
15.     CTG = Average ==> GP = '(82.2-83.9]' or ATT = '(-inf-70.5]'
16.     CTG = Average ==> GP = '(82.2-83.9]' or ESM = Second
17.     CTG = Average ==> SEM = '(-inf-78.6]' or ATT = '(88-90.5]'
18.     CTG = Average ==> SEM = '(-inf-78.6]' or LT = '(86.4-90.2]'
19.     CTG = Average ==> SEM = '(-inf-78.6]' or ESM = Second
20.     CTG = Average ==> ASS = No or ESM = Second
```

Fig. 8 'OR' Rules (Sample)

4.3. 'AND' Rules with their Decision parameter

1	And condition			IF								Decision						
2	ASS	ESM	CTG	ATT	PSM	GP	SEM		GP	LT	ATT	PSM	SEM	ESM	CTG	ASS		
3	yes	s							85.6-87.3	86.4-90.2								
4		s	avg							71.2-75	88-90.5							
5		s	avg								88-90.5	t						f=first
6			avg	88-90.5					82.2-83.9									s=second
7	yes				f				>87.3	86.4-90.2								t=third
8	yes				f								80.2-81.8	f				avg=average
9	yes				f					86.4-90.2			>91.4					g=good
10	yes			88-90.5						86.4-90.2			83.4-85					n=no
11		s		88-90.5					85.6-87.3				80.2-81.8					yes=yes
12	yes				s								78.6-80.2		g			
13			g		f				>87.3				89.8-91.4					
14		f			f				>87.3				89.8-91.4					
15						82.2-83.9	<78.6									n		
16		s				82.2-83.9					88-90.5							
17	yes			88-90.5					83.9-85.6	86.4-90.2								
18	yes				s								78.6-80.2		g			
19	yes	s							85.6-87.3				80.2-83.9					
20	yes	f											80.2-81.8		g			
21	yes	f								86.4-90.2				f				
22			avg	88.0-90.5						86.4-90.2				f				
23	y			88.0-90.5						86.4-90.2				f				
24			avg	88.0-90.5						86.4-90.2		f						
25				88.0-90.5		82.2-83.9				86.4-90.2		f						
26		s				82.2-83.9				86.4-90.2		f						
27		s		88.0-90.5						86.4-90.2		f						
28			avg	88.0-90.5						86.4-90.2						n		
29	yes		avg						85.6-87.3	86.4-90.2								
30		s		88.0-90.5											avg			

Fig 9 'AND' Rule with Decision Parameters

4.4. 'OR' Rules with their Decision Parameters

1	OR condition									Decision parameters					
2	ASS	ESM	CTG	ATT	PSM	GP	SEM	LT		ATT	PSM	SEM	ESM	CTG	ASS
3						82.2-83.9		71.2-75						avg	
4					t	82.2-83.9								avg	
5						82.2-83.9	<78.6							avg	
6				>90.5				90.2						g	
7		f		>90.5										g	
8		f					81.8-83.4							g	
9						83.9-85.6	81.8-83.4	>90.2						g	
10						>87.3	83.4-85	>90.5						g	
11						83.9-85.6	>91.4	>90.5						g	
12						83.9-85.6		>90.2					f		
13							89.8-91.4	>90.2					f		
14						83.9-85.6	>91.4						f		
15	n							71.2-75				<78.6			
16	n				t							<78.6			
159		t					78.6-80.2				s				
160		t					78.6-80.2				s				
161						85.6-87.3		82.6-86.4					s		
162					t			82.6-86.4					s		
163			p				81.8-83.4	78.8-82.6			s				
164			p				83.4-85	78.8-82.6			s				
165						85.6-87.3	80.2-81.8						s		
166							80.2-81.8	82.6-86.4					s		
167		t					81.8-83.4				s				
168						<73.7	81.8-83.4	78.8-82.6			s				
169		t					83.4-85				s				
170						<73.7	83.4-85	78.8-82.6			s				

Fig 10 'OR' Rule with Decision Parameters

1	OR-Condition					Decision parameter				
2	ATT	GP	SEM	LT	count	PSM	ESM	CTG	GP	count
3	Att1	Gp2	Sem1	Lt1	4			avg		1
4		Gp3	Sem4	Lt5	3			g		1
5	Att2	Gp3	Sem6	Lt5	4		f			1
6		Gp3	Sem7		2		f			1
7	Att3		Sem2	Lt4	3			avg		1
8	Att3	Gp4			2		s			1
9	Att2	Gp3	Sem4		3			g		1
10		Gp3	Sem7	Lt3	3			g		1
11	Att3			Lt1	2			avg		1
12		Gp2	Sem4	Lt1	3		s			1
13	Att3			Lt2	2			avg		1
14	Att3		Sem1	Lt3	3		s			1
15		Gp2	Sem1	Lt3	3		s			1
16	Att3			Lt2	2				Gp2	1
17		Gp4	Sem6	Lt5	3			g		1
18	Att3		Sem3	Lt1	3			avg		1
19		Gp4	Sem3	Lt3	3		s			1
20	Att4	Gp5	Sem5	Lt4	4		f			1

Fig 11 Ordinal Valued Table

4.5 PDI-Algorithm

The PDI algorithm has the following steps.

1. The transaction data is converted into Ordinal Valued Information System (OVIS).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

2. Soft set (α, E) is applied to above OVIS.
3. Delete those items generated and does not belongs to PDI. (with either minimum count or empty elements) are deleted.

1	IF			Decision			
2	ESM	ATT	GP	GP	LT	SEM	
3	s			85.6-87.3	86.4-90.2		
4	s				71.2-75		
5	s						f=first=1
6		88-90.5		82.2-83.9			s=second=2
7				>87.3	86.4-90.2		t=third=3
8						80.2-81.8	avg=average=a
9					86.4-90.2	>91.4	g=good=b
10		88-90.5			86.4-90.2	83.4-85	n=no=0
11	s	88-90.5		85.6-87.3		80.2-81.8	yes=yes=1
12						78.6-80.2	SEM->(<78.6=0)(78.6-80.2=1)(80.2-81.8=2)(81.8-83.9=3)(89.8-91.4=4)
13				>87.3		89.8-91.4	LT=(<71.2=0)(71.2-75=1)(86.4-90.2=2)
14	f			>87.3		89.8-91.4	
15			82.2-83.9				
16	s		82.2-83.9				
17		88-90.5		83.9-85.6	86.4-90.2		
18						78.6-80.2	
19	s			85.6-87.3		80.2-83.9	
20	f					80.2-81.8	
21	f				86.4-90.2		
22		88.0-90.5			86.4-90.2		
23		88.0-90.5			86.4-90.2		
24		88.0-90.5			86.4-90.2		
25		88.0-90.5	82.2-83.9		86.4-90.2		
26	s		82.2-83.9		86.4-90.2		
27	s	88.0-90.5			86.4-90.2		
28		88.0-90.5			86.4-90.2		
29				85.6-87.3	86.4-90.2		
30	s	88.0-90.5					
31							
32		88.0-90.5	82.2-83.9				
33	s	88.0-90.5			86.4-90.2		

	A	B	C	D	E	F	G	H	I	J	K
1	Rules	OR-Condition					Decision parameter				
2	R	ATT	GP	SEM	LT	count	PSM	ESM	CTG	GP	count
3	R1	Att1	Gp2	Sem1	Lt1	4			avg		1
4	R2		Gp3	Sem4	Lt5	3			g		1
5	R3	Att2	Gp3	Sem6	Lt5	4		f			1
6	R4		Gp3	Sem7		2		f			1
7	R5	Att3		Sem2	Lt4	3			avg		1
8	R6	Att3	Gp4			2		s			1
9	R7	Att2	Gp3	Sem4		3			g		1
10	R8		Gp3	Sem7	Lt3	3			g		1
11	R9	Att3			Lt1	2			avg		1
12	R10		Gp2	Sem4	Lt1	3		s			1
13	R11	Att3			Lt2	2			avg		1
14	R12	Att3		Sem1	Lt3	3		s			1
15	R13		Gp2	Sem1	Lt3	3		s			1
16	R14	Att3			Lt2	2				Gp2	1
17	R15		Gp4	Sem6	Lt5	3			g		1
18	R16	Att3		Sem3	Lt1	3	f		avg		2
19	R17		Gp4	Sem3	Lt3	3		s			1
20	R18	Att4	Gp5	Sem5	Lt4	4		f			1
21	R19	Att4			Lt5	2		f			1
22	R20	Att2	Gp3	Sem7	Lt3	4			g		1
23	R21	Att3		Sem1	Lt4	3				Gp2	1
24	R22			Sem5	Lt3	2			g		1
25	R23			Sem4	Lt1	2				Gp4	1
26	R24		Gp1		Lt2	2		t			1

Sheet4 Sheet1 Sheet2 Sheet3

Fig 13 Soft Set Optimized Rules

4. Find the coo parameter.
5. Evaluate the support of item sets and compare with the min_sup.
6. Produce the optimized ARs from frequent itemsets.
7. Evaluate the confidence and compare with the min_ conf of the rules.

The strong ARs that containing in PDIs and satisfying minimum confidence threshold. In the step 2 the student data set is converted into Ordinal variable values as follows. By applying this algorithm to the ARs at step 4, we have deleted the parameters with minimum count and most empty spaces which are not required for the optimized analysis.

Using the softest concept, the counts of the attributes in each rule, the maximum count of the attributes are used from the left hand and right hand part of the rules.

4.6 Principle of Parameter Co-Occurrence (coo)

coo(s1)=Att1,Gp2,Sem1,LT1,0,0,Avg,0
coo(s2)=0,Gp1,Sem4,Lt5,0,0,G,0
coo(s3)=Att2,Gp3,Sem6,Lt5,0,F,0,0
coo(s4)=0,GP3,Sem7,0,0,F,0,0
coo(s5)=Att3,0,Sem2,Lt4,0,0,Avg,0

$$M - factor(X \overset{max}{\Longrightarrow} Y) = \frac{C_D^{max}(X \overset{max}{\Longrightarrow} Y)}{C_v(X => Y)} \quad (4)$$

The support of each rule is calculated which is shown below.

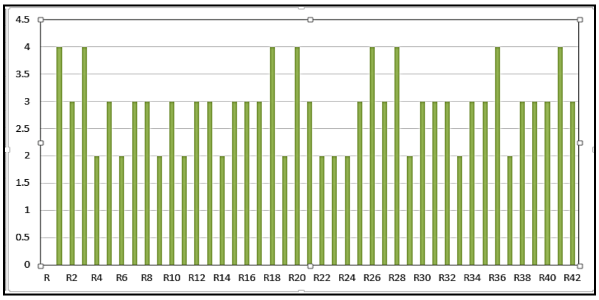

Fig 12 Support Count

Then the rules are optimized for 'AND' and 'OR' rules based on the soft set principle of eliminating minimum count attributes and most empty attributes. The real valued table is converted into an ordinal valued table which defines each rule with max_sup. The 'AND-OR' rules are ignored in this case as only few 'AND-OR' rules are generated and it is not required to optimize.

The attributed are named with

Att: *Att0=null, Att1<=70.5, Att2=85.5-88, Att3=88-90.5, Att4>=90.5*

Gp:*Gp0=null,Gp1<=73.7, Gp2=82.2-83.9, Gp3=83.9-85.6, Gp4=85.6-87.3, Gp5>=87.3*

Sem:*Sem0=null,Sem1<=78.6,Sem2>=78.6,Sem3=78.6-80.2, Sem4=80.2-81.8, Sem5=81.8-83.4, Sem6=83.4-85, Sem7=89.8-91.4, Sem8>=91.4*

Lt: *Lt0=null, Lt1<=40, Lt2=71.2-75, Lt3=78.8-82.6, Lt4=82.6-86.4, Lt5=86.4-90.2, Lt6>=90.2, Lt7=90.2, Lt8>=90.5*

PSM: *First, Second, Fail.* **ESM:** *First, Second, Third.*

CTG: *Good, Average, poor.*

The optimized rules that are generated using the soft set analysis have more support and confidence values.

Rule 1. GP=Gp3 6 ==> SEM=Sem4 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 2. SEM=Sem4 3 ==> GP=Gp3 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 3. GP=Gp3 6 ==> SEM=Sem7 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 4. SEM=Sem7 3 ==> GP=Gp3 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 5. GP=Gp3 6 ==> SEM=Sem4 CTG=Good 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 6. SEM=Sem4 3 ==> GP=Gp3 CTG=Good 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 7. GP=Gp3 CTG=Good 6 ==> SEM=Sem4 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 8. SEM=Sem4 CTG=Good 3 ==> GP=Gp3 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 9. GP=Gp3 6 ==> SEM=Sem7 CTG=Good 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 10. SEM=Sem7 3 ==> GP=Gp3 CTG=Good 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 11. GP=Gp3 CTG=Good 6 ==> SEM=Sem7 3 conf:(0.75) < lift:(2.67)> lev:(0.12) [1] conv:(1.22)
Rule 12. SEM=Sem7 CTG=Good 3 ==> GP=Gp3 3 conf:(1) < lift:(2.67)> lev:(0.12) [1] conv:(1.88)
Rule 13. SEM=Sem0 4 ==> GP=Gp0 CTG=Average 3 conf:(0.75) < lift:(2.4)> lev:(0.11) [1] conv:(1.38)
Rule 14. GP=Gp0 CTG=Average 5 ==> SEM=Sem0 3 conf:(0.6) < lift:(2.4)> lev:(0.11) [1] conv:(1.25)
Rule 15. ATT=Att3 6 ==> GP=Gp0 CTG=Average 4 conf:(0.67) < lift:(2.13)> lev:(0.13) [2] conv:(1.38)
Rule 16. GP=Gp0 CTG=Average 5 ==> ATT=Att3 4 conf:(0.8) < lift:(2.13)> lev:(0.13) [2] conv:(1.56)
Rule 17. ATT=Att3 6 ==> CTG=Average 6 conf:(1) < lift:(2)> lev:(0.19) [3] conv:(3)
Rule 18. CTG=Average 8 ==> ATT=Att3 6 conf:(0.75) < lift:(2)> lev:(0.19) [3] conv:(1.67)
Rule 19. GP=Gp3 6 ==> CTG=Good 6 conf:(1) < lift:(2)> lev:(0.19) [3] conv:(3)
Rule 20. CTG=Good 8 ==> GP=Gp3 6 conf:(0.75) < lift:(2)> lev:(0.19) [3] conv:(1.67)

Fig 14. Final Optimized Set of Rules

V. CONCLUSION

The prominence of an institute is incredibly included by the consequences of the graduates. In this paper, a basic procedure dependent on the affiliation investigation calculation utilizing PCY hash based apriori calculation. It is utilized to decide set of feeble students in the present semester by contrasting the exhibition of understudies of past semesters based on imprints got at the alumni level, marks got in past semester, current semester participation and

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

current semester class test marks. The guidelines are improved by utilizing delicate set procedure to acquire the solid connection between the qualities that influence the students' presentation. This philosophy will help the scholastic foundations to make the strides at the underlying dimension in the ensuing years with the end goal that the general aftereffect of the organization can improve. This will result in great positions and thusly expanding the quality admission of understudies. Thus this model will assume significant job for scholarly organizers to decide the arrangement of understudies that requires additional exertion from organization side.

Again, it is shown how soft set concept can be applied in multi-criteria decision making and optimizing the ARs. In this context, we have implemented real "students` data set to generate optimized rules to extract information for further analysis. This technique can be implemented for various types of real time data sets like employee data, customer data, business transaction data etc, with uncertainty where information is correctly unknown. Complete software can be developed to automatically generate rules and optimize the rules as per user requirement. Optimizing the rules can be implemented using fuzzy soft sets and interval valued soft sets, which will give more accurate result compare to the current one.

REFERENCES

1. "Fast Algorithms for Mining Association Rules", Proc. of 1994 Int. Conf. Very Large Data Bases (VLDB'94), Agrawal R. and Srikant R., pp. 487-499, Santiago, Chile, Sept. 1994.
2. "Fast Algorithm for Mining Association Rules in Large Databases", Agrawal R. and Srikant R., Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA, June 1994.
3. "Mining Sequential Patterns", Proc. 1995 Int. Conf. Data Engineering (ICDE'95), Agrawal R. and Srikant R., pp. 3-14, Taipei, Taiwan, Mar. 1995.
4. "Softset Relations and Functions", Journal of Computational and Applied Mathematics, Babitha K.V., Sunil J.J., Vol. 60, pp.1840-1849, Oct 2010.
5. "Association Mining', Ceglar, Aaron, Roddick, John F., ACM Computing Surveys, Vol. 38, Issue 2, 2006.
6. "Data Mining: Concepts and Techniques", Han, J. and Kamber, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.
7. "A Soft Set Approach for Association Rules Mining", Herawan Tutut, Mustafa Mat Deris, Knowledge-Based Systems, Vol. 24, pp. 186–195, 2011.
8. "Soft Set Theory", Maji P. K., Biswas R., Roy A.R., Journal of Computational and Applied Mathematics, Vol. 45, pp. 555–562, 2003.[12]
9. "Analysis of E-Exam Data using Data Mining Techniques", Mamcenko J., Jelena, Irna S., Jurgita L. and Regina K., Proceeding of 17th International Conference on Information and Software Technologies, pp. 215-219, 2011.
10. "Soft Set Theory-First Results", Molodtsov D., Journal of Computational and Applied Mathematics, Vol. 37, pp. 19–31, 1999.
11. "An Effective Hash-based Algorithm for Mining Association Rules", Park J., Chen M., and Yu P., Proceedings of the 1995 ACM SIGMOD (SIGMOD '95) International Conference on Management of Data, San Jose, California, USA, pp. 175-186, May 22 - 25, 1995.
12. "A CHAID based Performance Prediction Model in Educational Data Mining", Ramaswami M. and Bhaskaran R, International Journal of Computer Science Issues (IJCSI), Vol. 7, Issue 1, No. 1, January 2010.
13. "A New Soft Set Based Association Rule Mining Algorithm",Kanojiya S. Singh, Tiwari A., International Journal of Computing Science and Communication Technologies, Vol.6. No. 2, January. 2014.
14. "Mining Association Rules from Dataset Containing Predetermined Decision Itemset and Rare Transactions", Sha Zongyao, Chen Jiangping, *Seventh International Conference on Natural Computation*, Pp.167-170, 2011.
15. "Pattern Mining Approach to Categorization of Students' Performance using Apriori Algorithm", Verma S. K., Thakur R. S., Jaloree S., *International Journal of Computer Applications* (0975 – 8887) Vol. 121, No.5, July 2014.
16. "The Normal Parameter Reduction of Soft sets and its Algorithm", Z. Kong, L. Gao, L. Wang and S. Li, Computers and Mathematics with Applications, Vol. 56, pp.3029-3037 (2008)

AUTHOR PROFILE

Sasanko Sekhar Gantayat is working as an Associate Professor in the Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India. He has completed his Ph.D. in Computer Science from Berhampur University, Berhampur (Odisha). He has published many research papers in national, international journals and conferences. He is a reviewer of many referred international journals, and his name appears in the editorial board of *American Journal of Database Theory and Application* (USA). He is a member of IE (I), ISTE, CSI, ACM, ACM-CSTA, IAENG, IOSRD, AIRCC, IRSS, IACSIT, ISIAM, IMS, WSEAS, and SSI. He is a member of IEEE and IEEE Computational Intelligence Society. His research interests include Fuzzy Sets and Systems, Rough Set Theory and Knowledge Discovery, List Theory and Applications, Soft Computing, Cryptography, Graph Theory and Machine Learning.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication