Mining Frequent Pattern with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP)

Spits Warnars
Database, Datawarehouse & Data Mining Research Center
Human Computer Interaction department
Surya University
Tangerang, Indonesia
Spits.warnars@surya.ac.id

***Abstract*—This paper is extended version from previous paper which proposed AOI-HEP as novel data mining technique. This paper will explain how AOI-HEP mining technique can be used to mine frequent pattern. AOI-HEP is influenced by Attribute Oriented Induction (AOI) and Emerging Pattern (EP) mining techniques by applying AOI characteristic rule algorithm and improvement EP growth rate. The experiment used adult dataset from UCI machine learning repository with 48842 instances, run in 3 seconds and the instances were discriminated between government and non government concepts based on learning on workclass attribute. AOI-HEP mining interest for frequent pattern will be influenced by learning on their chosen attribute. The experiments showed that adult dataset which learn on workclass attribute had AOI-HEP mining interest for frequent pattern and there are four frequent patterns which have strong discrimination rule. Meanwhile, extended experiments upon adult dataset which learn on marital-status attribute showed there is no AOI-HEP mining interest for frequent pattern.**

Keywords—Data Mining; Attribute Oriented Induction; Emerging pattern ;AOI-HEP, High Level Emerging Pattern

I. INTRODUCTION

Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) was proposed as a novel data mining technique [1,2] has been success to mine:

- Total Subsumption HEP (TSHEP) patterns as HEP[1] which are frequent in one rule but less frequent in another rule [3,4].
- Subsumption Overlapping HEP (SOHEP) patterns as HEP which are numerous unlike the rare TSHEP patterns [1].

This paper will discuss how AOI-HEP is extended to mine frequent pattern and next are improvements from previous paper[1]:

- The term of TSHEP or SOHEP will be revised to mine frequent pattern.
- AOI-HEP framework will be revised only to search frequent pattern.
- High Level Emerging Pattern (HEP) algorithm will be revised only to search frequent pattern.

II. AOI-HEP FRAMEWORK

Figure 1 shows the proposed AOI-HEP framework where traditional AOI characteristic rule algorithm is run twice with two datasets D1 and D2 (horizontal partitions of the dataset). AOI uses concept hierarchy as background knowledge for data generalization. AOI eliminates distinct attributes and tuples until they are less or equal than attribute and rules thresholds respectively [5]. AOI's outputs are rulesets R_i^1 and R_j^2 from datasets D1 and D2 respectively. Rulesets R_i^1 and R_j^2 are inputs for HEP algorithm which include two functions i.e. similarity function $C\{R_i^1, R_j^2\}$ and growth rate function $GR\{R_i^1, R_j^2\}$. The $C\{R_i^1, R_j^2\}$ function is a metric similarity function which applies cartesian product between rulesets R_i^1 and R_j^2, and eliminate the cartesian product with non frequent pattern. The $GR\{R_i^1, R_j^2\}$ function is ratio of the supports between rulesets R_i^1 and R_j^2, can eliminates frequent pattern as the outputs from $C\{R_i^1, R_j^2\}$ function with growth rate less equal than GrowthRate threshold, as shown in line number 10 and 11 HEP algorithm in figure 2.

A. HEP algorithm

Figure 2 shows the HEP algorithm as part of AOI-HEP framework in figure 1, where The HEP algorithm has inputs such as rulesets R_i^1 and R_j^2, GR_threshold, num_attr, |D2| and |D1|.

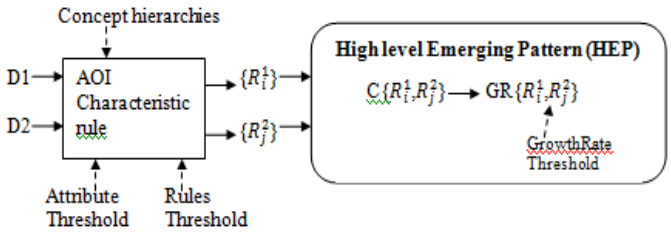

Figure 1. AOI-HEP framework

Input : $\{R_i^1\}, \{R_j^2\}$, GR_threshold, num_attr, |D2|, |D1|

```
1.  While( noAllANY(R_i^1))
2.    While ( noAllANY(R_j^2))
3.      SLV=0,F=0
4.      for x=1 to num_attr
5.        If R_i^1[x]== R_j^2 [x] and R_i^1[x]=="ANY"          SLV=SLV+2.1
6.        If R_i^1[x]== R_j^2[x] and R_i^1[x]!="ANY"           SLV=SLV+2
7.        If R_i^1[x]!= R_j^2[x] and R_j^2[x] subsump by R_i^1[x]  SLV=SLV+0.4
8.        If R_i^1[x]!= R_j^2[x] and R_i^1[x] subsump by R_j^2[x]  SLV=SLV+0.5,F++
9.      if(SLV>=(num_attr-1)*0.5+0.4 and SLV<=(num_attr-1)*0.5+2.1) and F>=x-1
10.       growth rate=(R_j^2[x+1]/|D2|) / (R_i^1[x+1]/|D1|)
11.       If growth rate > GR_threshold
12.         Print  (R_j^2[x+1]/|D2|), (R_i^1[x+1]/|D1|), growth rate, SLV,
```

Figure 2. HEP algorithm

The HEP algorithm inputs are in accordance with inputs for HEP in AOI-HEP framework figure 1 where for HEP in figure 1 there are rulesets R_i^1 and R_j^2 inputs, GR_threshold for GR$\{ R_i^1 , R_j^2 \}$ function. The GR_threshold threshold has default value 0 and maximum value 100. Moreover, num_attr input is the number attributes in rulesets R_i^1 and R_j^2 as m in equation 1. |D2| and |D1| are total number of instances in dataset D2 and D1 respectively as shown in equation 3. Meanwhile, F variable is frequent pattern indicator to eliminate non frequent pattern.

The outputs from HEP algorithm are (R_j^2 [x+1]/|D2|) as support target dataset , (R_i^1 [x+1]/|D1|) as support contrasting dataset GrowthRate and SLV value. In the HEP algorithm, line number 1 and 2 are used to exclude rule with ANY values in all attributes in rulesets R_i^1 and R_j^2 respectively. Rules with ANY values are less meaningful and do not offer meaningful interpretation. C$\{ R_i^1 , R_j^2 \}$ and GR$\{ R_i^1 , R_j^2 \}$ functions in figure 1 are shown between line number 4 and 8 and line number 10 and 11 in HEP algorithm respectively. Moreover, statement in line number 9, to eliminate non frequent pattern.

B. HEP definition

For HEP, let D1 and D2 be horizontal partitions of some dataset $D^x = \{A_1,...,A_p\}$ with p attributes $1 \le i \le p$ and $1 \le x \le 2$. Rulesets $\{ R_i^1 \}$ and $\{ R_j^2 \}$ from datasets D1 and D2 are represented as $R^x = \{r_1^x, r_2^x, ..., r_n^x\}$ in figure 3. In figure 3 each ruleset Rx consists of n rules where n $\le$ R.Thr, a rules threshold. Each rule in a ruleset Rx is represented by attributes $r_n^x = \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\}$, where $|r_n^x|$ is the number of tuples forming the rule and m is is the number of attributes in a ruleset as in equation 1. Figure 3 shows the representation of rulesets $R^x = \{r_1^x, r_2^x, ..., r_n^x\}$ vertically where $r_n^x \in R^x$ and each rule $r_n^x = \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\}$, horizontally where $A_m^x \in r_n^x$. As an example we have used rule r_1^1 in ruleset 1 and rule r_1^2 in ruleset 2.

$$D^x \rightarrow R^x \rightarrow \begin{aligned} r_1^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_1^x|\} \\ r_2^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_2^x|\} \\ r_{...}^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_{...}^x|\} \\ r_n^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\} \end{aligned}$$

Figure 3. Representation rules and rulesets

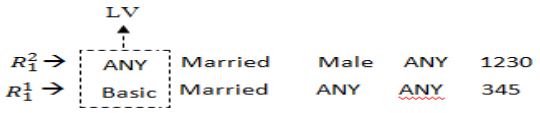

Figure 4. Comparation rule 1 of ruleset 2 $\{ R_1^2 \}$ and rule 1 of ruleset 1 $\{ R_1^1 \}$

$A_m^1 \in r_1^1$ where all attributes A_m^1 are member of rule r_1^1 in ruleset 1 and $A_m^2 \in r_1^2$ where all attributes A_m^2 are member of rule r_1^2 in ruleset 2. For example, if there are four attributes (m=4 in equation 1) then rule $r_1^1 = \{A_1^1, A_2^1, A_3^1, A_4^1, |r_1^1|\}$ and rule $r_1^2 = \{A_1^2, A_2^2, A_3^2, A_4^2, |r_1^2|\}$.

C. Metric similarity

The C$\{ R_i^1 , R_j^2 \}$ function as shown in figure 1 is a metric similarity function which apply cartesian product between rulesets R_i^1 and R_j^2 , and eliminate the cartesian product with non frequent pattern. The determining frequent pattern is applied by summing categorization of attribute comparison value and hierarchy level based on subsumption and overlap thresholds. To derive similarity hierarchy level value (SLV) in the HEP algorithm, firstly, we determine categories of attribute values between the rulesets as shown in figure 4. The categorization is based on similarity hierarchy level and the values shown in equation 1 as LV. Secondly, by summing the attribute categorizations or LV values, we get SLV (equation 1) as the similarity between the two rules. The two steps described above are shown between line numbers 4 and 8 in the HEP algorithm of figure 2.

$$\text{SLV} = \sum_{i=1}^{m} LV_i \quad (1)$$

where:

SLV=similarity value based on the similarity of attributes hierarchy level and values

m= number of attributes in a ruleset, where m > 1

(number of attributes in concept hierarchies - 1)

i=attribute position

LVi = categorization of attributes comparison based on similarity hierarchy level and values, and the options are

1) *If hierarchy level is different and the attribute in rule of ruleset R2 is subsumed by the attribute in rule of ruleset R1 (R2 ⊂ R1), LV=0.4.*
2) *If hierarchy level is different and the attribute in rule of ruleset R1 is subsumed by the attribute in rule of ruleset R2 (R1 ⊂ R2), LV=0.5.*
3) *If hierarchy level and values are the same and the attributes values are not ANY, LV=2.*
4) *If hierarchy level and values are the same and the attributes values are ANY, LV=2.1.*

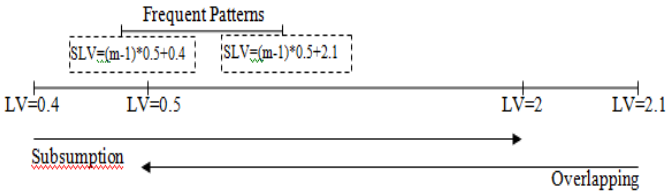

Figure 5. Composition subsumption and overlapping mining pattern

The four categorization of attribute comparisons or LV in equation 1 is based on two main categorizations i.e. subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2 or LV=2.1). Thus, the attributes will be categorized as subsumption when attributes comparison has different hierarchy level and value (LV=0.4 or LV=0.5). On the other hand, the attributes will be categorized overlapping when comparison between attributes has the same hierarchy levels and values (LV=2 or LV=2.1). For each LV option values 0.4,0.5,2 and 2.1 are user defined number, where option numbers 0.4 and 0.5 as values for subsumption categorization (minimum categorization) and option numbers 2 and 2.1 as values for overlapping categorization (maximum categorization). LV=0.4 is minimum value for subsumption categorization and if ruleset R2 is subsumed by ruleset R1 (R2 $\subset$ R1). On the other hand LV=0.5 is maximum value for subsumption categorization and if ruleset R1 is subsumed by ruleset R2 (R1 $\subset$ R2). LV=2 is minimum value for overlapping categorization and if the attributes values are not ANY, on the other hand LV=2.1 is maximum value for overlapping categorization and if the attributes values are ANY. Finally, LV=0.4 and LV=2.1 are taken as the minimum and maximum values of LV values respectively.

After the similarity between the two rules (SLV) has been derived, then we can determine frequent pattern which create discriminant rules show the discrimination between two rules in rulesets. Frequent pattern has minimum and maximum SLV values which can be derived with equations 2. In equations 2, m is the number of attributes in ruleset similar as m in equation 1, c and c1 are LV value in equation 1 where c is 0.5 and c1 is options between 0.4 and 2.1. The equation 2 indicates the frequency c for m-1 times plus c1 where c as LV value has similar frequency subsumption m-1 times plus c1 as combination c.

$$(m-1)*c + c1 \quad (2)$$

where:

m = m in equation 1

c =LV option 0.5 in equation 1

c1 =combination c, LV options (0.4 and 2.1) in equation 1

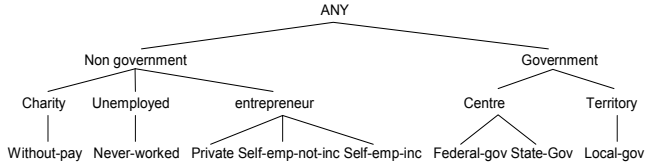

Figure 6. Workclass concept hierarchy

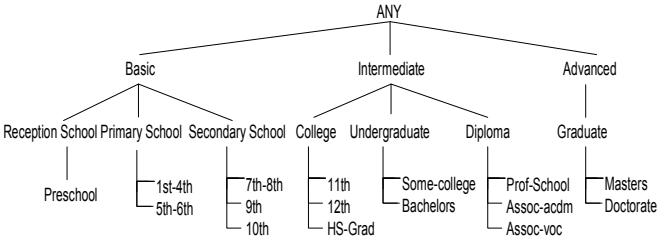

Figure 7. Education concept hierarchy

The two arrow lines in figure 5 shows the influence of two main categorizations subsumption and overlapping. The overlapping arrow line shows the influence overlapping from LV=2.1 (maximum value for overlapping categorization) until LV=0.5 (maximum value for subsumption categorization). Whilst subsumption arrow line shows the influence subsumption from LV=0.4 (minimum value for subsumption categorization) until LV=2 (minimum value for overlapping categorization).

D. Mining frequent pattern

Frequent pattern is a combination of feature patterns that appear in dataset with frequency not less than a user-specified threshold[3] and the frequent pattern synonym with large pattern was first proposed for market basket analysis in the form of association rules[6]. With frequent pattern we can have strong/sharp discrimination power where have large growth rate and support in target (D2) dataset and other support in contrasting (D1) dataset is small [7,8,9]. In AOI-HEP, the frequent pattern is shown by the subsumption LV=0.4 or LV=0.5 and as mention previously when LV=0.4 then ruleset R2 is subsumed by ruleset R1 (R2 $\subset$ R1) where R2 is subset rule and R1 is superset rule. On the other hand when LV=0.5 then ruleset R1 is subsumed by ruleset R2 (R1 $\subset$ R2) where R1 is subset rule and R2 is superset rule. R2 is in target (D2) dataset and R1 is in contrasting (D1) dataset (D2/D1=target/contrasting=R2/R1) and it is as accordance with HEP growth rate in equation 3. Superset rule is a frequent pattern since subset rule is part of the superset rule and for instance when SLV has the same LV values (SLV=0.5+0.5+0.5+0.5=2) then certainly the number of instances in superset rule is larger than in its subset rule. Thus, that instance condition SLV=0.5+0.5+0.5+0.5=2 shows that superset rule (frequent pattern) has high support (large pattern) and subset rule (infrequent pattern) has low support. in Emerging Pattern (EP), patterns will be recognized as EP if have high support (frequent pattern) in one class and low support (infrequent pattern) in other one [8,10].

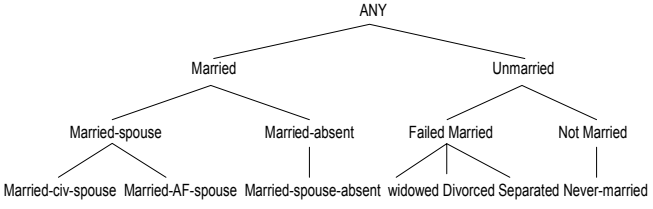

Figure 8. Marital-status concept hierarchy

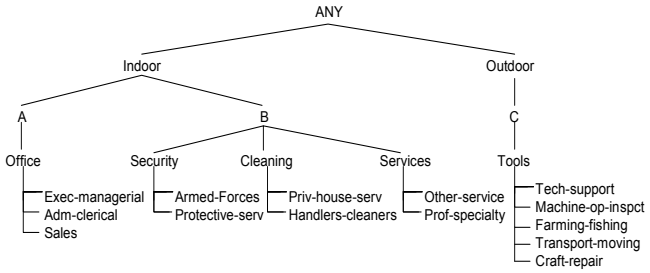

Figure 9. Occupation concept hierarchy

From frequent patterns, we can create a discrimination rule and are interested in mining the frequent pattern with strong/sharp discrimination power. In EP, the strength of discrimination power is expressed by its large growth rate and support in target (D2) dataset [7,8,9]. This is called an essential Emerging Patterns (eEP) [8]. In AOI-HEP, the strength of discrimination power is expressed by its large growth rate and support in target (D2) dataset as well. Certainly, to make large growth rate can be happened when have large support in target (D2) dataset and low support in contrasting (D1) dataset. Indeed, in EP, patterns will be recognized as EP if have high support in one class and low support in other one [8,10]. Moreover, support in contrasting (D1) dataset must be less than support in target (D2) dataset where by the end will create large growth rate.

In AOI-HEP, the strength of discriminant power is expressed by subsumption LV=0.5 where R2 in target (D2) dataset is superset and R1 in contrasting (D1) dataset is subset. The strength of discrimination power with subsumption LV=0.5 shows that have large support in target (D2) dataset and low support in contrasting (D1) dataset, where by the end will create large growth rate. Thus, for discriminant rule from frequent pattern which SLV value with all similarity subsumption LV=0.5 (SLV value with similarity subsumption LV=0.5, for instance SLV=0.5+0.5+0.5+0.5=2) will have frequent pattern with strong discrimination power. Meanwhile, there is SLV value with nearly all subsumption LV=0.5 and recognized as SLV value with frequent subsumption LV=0.5. However, SLV value with frequent subsumption LV=0.5 will be interested to be explored.

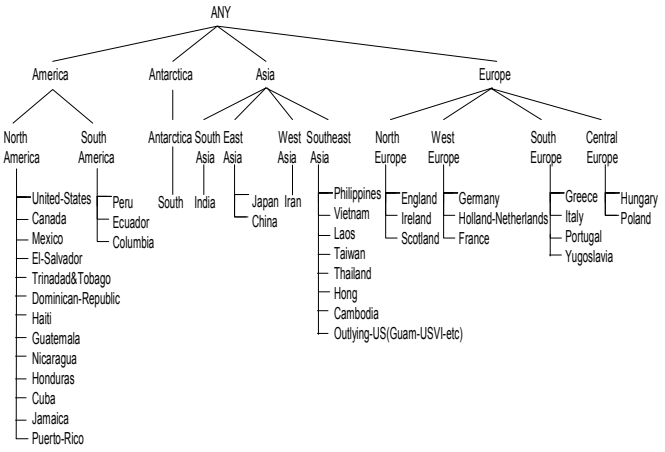

Figure 10. Native-country oncept hierarchy

TABLE I. **RULESET R2 FOR LEARNING GOVERNMENT**

No	Education	Marital	Occupation	Country	Instances
0	Intermediate	ANY	ANY	ANY	3454
1	ANY	ANY	ANY	America	786
2	Advanced	ANY	ANY	Asia	30
3	Advanced	ANY	ANY	Europe	17
4	Basic	Married-spouse	Services	Europe	1
5	Advanced	Married-spouse	Services	Antartica	1

This is because when two parts of objects are similar if they are similar in all features (full matching similarity) or if the percentage of similar features is greater than the 80%[11] or if they are similar in at least 90% of the features[12].

Since there are SLV value with all subsumption LV=0.5 where have full similarity subsumption LV=0.5, then there are frequent pattern with strong discrimination power for SLV value with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100 where m as in equation 1. Since the strength of discriminant power is expressed by subsumption LV=0.5 and frequent pattern has minimum and maximum SLV values of (m-1)*c+c1 with equation 2 where c=0.5,c1=0.4 and c=0.5,c1=2.1 then (m-1)*0.5+0.4 and (m-1)*0.5+2.1 respectively as shown in figure 5. Minimum and maximum SLV value for frequent pattern are SLV=(m-1)*0.5+0.4 and SLV=(m-1)*0.5+2.1 show the frequent similarity subsumption (LV=0.5) in m-1 times at percentage value of (m-1)/m*100 ((m-1)*0.5) plus 0.4 as minimum subsumption and 2.1 as maximum overlapping LV value categorization respectively. Thus, minimum and maximum SLV value for frequent pattern show frequent similarity subsumption (LV=0.5) at percentage value of (m-1)/m*100 which express discrimination power plus minimum subsumption LV=0.4 and maximum overlapping LV=2.1 respectively. Finally, with AOI-HEP we can mine frequent pattern with strong discrimination power in optional conditions:

- SLV value with full similarity subsumption LV=0.5.
- SLV value with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100 where m as in equation 1.

TABLE II. **RULESET R1 FOR LEARNING NON GOVERNMENT**

No	Education	Marital	Occupation	Country	Instances
0	7th-8th	Widowed	Tools	United-states	1
1	HS-grad	Never-married	ANY	United-states	4
2	HS-grad	Married-civ-spouse	ANY	ANY	5
3	Assoc-adm	Married-civ-spouse	Tools	United-states	1
4	Some-college	Married-civ-spouse	ANY	United-states	2
5	Some-college	Married-spouse-absent	Tools	United-states	1

TABLE III. FREQUENT PATTERN FOR RULESETS R_3^1 TO R_0^2 WITH GR=(3454/4289)/(1/14)=0.80532/0.07143=11.27442

Rulesets	Education	Marital	Occupation	Country	Instances
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_3^1	Assoc-adm	Married-civ-spouse	Tools	United-states	1
LV	0.5	0.5	0.5	0.5	

Mining frequent pattern with that two optionals above between full similarity and frequent similarity subsumption LV=0.5 as mentioned above can be seen in HEP algorithm in figure 2 by using F attribute which control how many subsumption LV=0.5 where indicate elimination for non frequent pattern with F>=x-1 as shown in line number 9 HEP algorithm in figure 2.

E. HEP growth rate

$$GR(X,Y) = \frac{\text{Support D2(X)}}{\text{Support D1(Y)}} = \frac{\text{Count R2(X)} \;/\; |D2|}{\text{Count R1(Y)} \;/\; |D1|} \quad (3)$$

where:

X = High level rule of ruleset R2 in dataset D2.

Y = High level rule of ruleset R1 in dataset D1.

D2 = Dataset D2.

D1 = Dataset D1.

|D2| = Total number of instances in dataset D2.

|D1| = Total number of instances in dataset D1.

Count R2(X) = Number of high level rule X of ruleset R2 in dataset D2.

Count R1(Y) = Number of high level rule Y of ruleset R1 in dataset D1.

Support D2(X) = Composition number of high level rule X of ruleset R2 in D2.

Support D1(Y) = Composition number of high level rule Y of ruleset R1 in D1.

Besides eliminating patterns with similarity function $C\{R_i^1, R_j^2\}$, the large number of frequent pattern can be eliminated by the growth rate function $GR\{R_i^1, R_j^2\}$ with given a GrowthRate threshold.

TABLE IV. FREQUENT PATTERN FOR RULESETS R_5^1 TO R_0^2 WITH GR=(3454/4289)/(1/14)=0.80532/0.07143=11.27442

Rulesets	Education	Marital	Occupation	Country	Instances
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_5^1	Some-college	Married- spouse-absent	Tools	United-states	1
LV	0.5	0.5	0.5	0.5	

TABLE V. FREQUENT PATTERN FOR RULESETS R_1^1 TO R_0^2 WITH GR=(3454/4289)/(4/14)=0.80532/0.28571=2.81861

Rulesets	Education	Marital	Occupation	Country	Instances
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_1^1	HS-Grad	Never-married	ANY	United-states	4
LV	0.5	0.5	2.1	0.5	

Growthrate is a standard function used in EP[9], and the difference in our approach is discovering high level emerging pattern with the same or different itemset instead of low level pattern with the same itemset. As mentioned in sub section B, rulesets are AOI outputs and each of rule in ruleset has $|r_n^\kappa|$ as the number of tuples forming the rule (figure 3). Because of rule in ruleset has $|r_n^\kappa|$ as the number of tuples, then there is no Jumping High level Emerging Patterns (JHEP), where JHEP is related as a term of JEP. JEP is EP with support is 0 in one dataset and more than 0 in the other dataset or EP as special type of EP which is having infinite growth rate (∞).

Growth rate GR$\{R_i^1, R_j^2\}$ in AOI-HEP framework is shown in figure 1 and in line number 10 in the HEP algorithm in figure 2 is used to discriminate between datasets D2 and D1. This growth rate can be calculated using equation 3. We can define that a HEP is a ruleset whose support changes from one ruleset in dataset D1 to another ruleset in dataset D2. In other words, HEP is a ruleset whose strength of high level rule Y of ruleset R1 in dataset D1 changes to high level rule X of ruleset R2 in dataset D2.

III. AOI-HEP EXPERIMENTS

Experiments used adult dataset from the UCI machine learning repository with the number of instances are 48842[13]. The programs were run with attribute and rule thresholds of 6 which were chosen based on the preliminary experiments done on adult dataset such that to get meaningful numbers of rules, a higher threshold is preferable after trial experiments. The experiments showed that frequent pattern as rare patterns and are numerous if using attribute thresholds between 4 and 6, and rules thresholds between 5 and 10.Since it was rare to find frequent pattern, we decided to use a bigger attribute threshold of 6 for experiments. Similarly, 6 was chosen for the rules threshold, since 6 is median between 2 and 9. Moreover, we obtained numerous frequent pattern rules for thresholds between 5 and 10 as expected when thresholds are bigger.

TABLE VI. FREQUENT PATTERN FOR RULESETS R_4^1 TO R_0^2 WITH GR=(3454/4289)/(2/14)=0.80532/0.14286=5.63721

Rulesets	Education	Marital	Occupation	Country	Instances
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_4^1	Some-College	Married-civ-spouse	ANY	United-states	2
LV	0.5	0.5	2.1	0.5	

Adult dataset has concept hierarchies built from five chosen attributes with a minimum concept level of three. The attributes in concept hierarchies for adult dataset are workclass, education, marital-status, occupation, and native-country attributes, as shown between figure 6 and 10 respectively. The instances of adult dataset was divided into two sub datasets based on learning the high level concept in one of their attributes. Adult dataset was leanerd by discriminating between the "government" (4289 instances) and "non government" (14 instances) concepts of the "workclass" attribute (as shown in figure 6) in datasets D2 and D1 respectively. Learning the high level concept in one of their five chosen attributes for concept hierarchies, makes the parameter m in equation 1 have value 4, where value 4 comes from five chosen attributes for concept hierarchies minus 1 and 1 is the attribute for the learning concept.

Experiments were carried out by a Java and tested on Intel(R) Atom(TM) CPU N550 (1.50 GHz) with 1.00 GB RAM. The AOI-HEP application has an input dataset and corresponding concept hierarchies in the form of flat files. Running AOI-HEP application with input adult dataset will have rulesets R2 and R1 as shown in tables I and II which have 6 tuples (rules) include number of instances for each tuple (rule). Each table has four attributes (m in equation 1) which are from five chosen attributes minus one attribute learning.

The results for running the AOI-HEP application for mining frequent pattern with adult dataset can be seen between tables III and VI where:

- Two SLV value frequent patterns with full similarity subsumption LV=0.5 as shown in tables III and IV.
- Two SLV value frequent patterns with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100 where m as in equation 1, as shown in tables V and VI.

Frequent pattern in table III has strong discrimination rule :

There are 11.2744 growth rates adult dataset with 80.53% frequent pattern in government workclass (with an intermediate education) and 7.14% infrequent pattern in non government workclass (with assoc-adm education, married-civ-spouse marital status, tools occupation and from the United States).

Frequent pattern in table IV has strong discrimination rule :

There are 11.2744 growth rates adult dataset with 80.53% frequent pattern in government workclass (with an intermediate education) and 7.14% infrequent pattern in non government workclass (with some college education, married-spouse-absent marital status, tools occupation and from the United States).

Frequent pattern in table V has strong discrimination rule :

There are 2.81861 growth rates adult dataset with 80.53% frequent pattern in government workclass (with an intermediate education) and 28.57% infrequent pattern in non government workclass (with HS-Grad education, Never-married marital status and from the United States).

Frequent pattern in table VI has strong discrimination rule :

There are 5.63721 growth rates adult dataset with 80.53% frequent pattern in government workclass (with an intermediate education) and 14.28% infrequent pattern in non government workclass (with some college education, married-civ-spouse marital status and from the United States).

Finally, experiments showed that adult dataset which learn on workclass attribute are interesting to mine since having four frequent patterns which are recognized as strong discrimination rules. Discriminating rules from table III to VI show as strong discriminating power where they have large growth rates (between 2.81861 and 11.2774) and supports in target (D2) datasets (80.53%). Moreover, they have small supports in contrasting (D1) dataset between 7.14% and 28.57% where each of the support in contrasting (D1) dataset is less than the support in target(D2) dataset.

REFERENCES

[1] S. Warnars, "Attribute Oriented Induction of High-level Emerging Patterns," in Proceedings of the IEEE International Conference on Granular Computing (IEEE GrC), Hangzhou, China, pp. 525–530, 11-13 August 2012.

[2] M.K. Muyeba, M. S. Khan, S. Warnars and J. Keane, "A framework to mine high-level emerging patterns by attribute-oriented induction," In Proceedings of the 12th international conference on Intelligent data engineering and automated learning (IDEAL'11), Norwich, United Kingdom, pp. 170-177, 7-9 September 2011.

[3] J. Han, H. Cheng, D. Xin and X.Yan, "Frequent pattern mining: current status and future directions," Data Min Knowl Disc, Vol.15, no.1, pp. 55-86, 2007.

[4] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Min. Knowl. Discov., Vol. 8, no.1,pp. 53-87, 2004.

[5] Y. Cai, N. Cercone and J. Han, "An attribute-oriented approach for learning classification rules from relational databases," In Proceedings of 6th International Conference on Data Engineering, pp. 281-288, 1990.

[6] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases,". ACM SIGMOD Rec, Vol. 22, no.2, pp. 207-216,1993.

[7] K. Ramamohanarao, J. Bailey and H. Fan, "Efficient Mining of Contrast Patterns and Their Applications to Classification," In Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing (ICISIP '05), IEEE Computer Society, pp. 39-47, 2005.

[8] H. Fan and K. Ramamohanarao, "A Bayesian approach to use emerging patterns for classification," In Proceedings of the 14th Australasian database conference (ADC '03), pp. 39-48, 2003.

[9] G. Dong and J. Li, "Efficient mining of emerging patterns: discovering trends and differences," In Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 43-52, 1999.

[10] R. Podraza and K. Tomaszewski, "KTDA: Emerging Patterns Based Data Analysis System," In Proceedings of XXI Fall Meeting of Polish Information Processing Society, pp. 213—221, 2005.

[11] R. Danger, J. Ruiz-Shulcloper and R.B. Llavori, "Objectminer: A new approach for Mining Complex objects," In Proceedings of the 6th international conference on Enterprise Information Systems (ICEIS '04), pp. 42-47, 2004.

[12] A.Y. Rodriguez-Gonzalez, J.F. Martinez-Trinidad, J.A. Carrasco-Ochoa and J. Ruiz-Shulcloper, "Mining Frequent Similar Patterns on Mixed Data," In Proceedings of the 13th Iberoamerican congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications(CIARP '08), pp. 136-144, 2008.

[13] A. Frank and A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.