International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 6, December 2016, pp. 3037~3046
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i6.10579

Using Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) to Mine Frequent Patterns

Harco Leslie Hendric Spits Warnars
Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Mar 23, 2016
Revised Jul 18, 2016
Accepted Aug 1, 2016

Keyword:

AOI-HEP
Data mining
Frequent pattern
HEP frequent pattern
High level emerging pattern

ABSTRACT

Frequent patterns in Attribute Oriented Induction High level Emerging Pattern (AOI-HEP), are recognized when have maximum subsumption target (superset) into contrasting (subset) datasets (contrasting ⊂ target) and having large High Emerging Pattern (HEP) growth rate and support in target dataset. HEP Frequent patterns had been successful mined with AOI-HEP upon 4 UCI machine learning datasets such as adult, breast cancer, census and IPUMS with the number of instances of 48842, 569, 2458285 and 256932 respectively and each dataset has concept hierarchies built from its five chosen attributes. There are 2 and 1 finding frequent patterns from adult and breast cancer datasets respectively, while there is no frequent pattern from census and IPUMS datasets. The finding HEP frequent patterns from adult dataset are adult which have government workclass with an intermediate education (80.53%) and America as native country (33%). Meanwhile, the only 1 HEP frequent pattern from breast cancer dataset is breast cancer which have clump thickness type of About Aver Clump with cell size of Very Large Size (3.56%). Finding HEP frequent patterns with AOI-HEP are influenced by learning on high level concept in one of chosen attribute and extended experiment upon adult dataset where learn on marital-status attribute showed that there is no finding frequent pattern.

Corresponding Author:

Harco Leslie Hendric Spits Warnars,
Bina Nusantara University,
Kampus anggrek, Jl. Kebon Jeruk raya no. 27, Jakarta Barat-11530, Indonesia.
Email: shendric@binus.edu

1. INTRODUCTION

Frequent pattern is a combination of feature patterns that appear in dataset with frequency not less than a user-specified threshold [1-3] and the frequent pattern synonym with large pattern was first proposed for market basket analysis in the form of association rules [4]. 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 [5-7]. Frequent patterns have been implemented in applications such as: customer transaction analysis, web mining, software bug analysis, chemical and biological analysis and etc [8-10]. Frequent pattern in Attribute Oriented Induction High level Emerging Pattern (AOI-HEP), is recognized when have maximum subsumption target (superset) into contrasting (subset) datasets (contrasting ⊂ target) and having large High Emerging Pattern (HEP) growth rate and support in target dataset [11]. In the first AOI-HEP version [12] had been success to mine:

a. Total Subsumption HEP (TSHEP) which frequent in one rule but less frequent in another rule.
b. Subsumption Overlapping HEP (SOHEP) which are combination between subsumption and overlapping between rulesets.

This paper is continous from previous paper [11] where mining frequent patterns with AOI-HEP does not only on adult dataset but will be extended to other 3 datasets such as breast cancer, census and

IPUMS datasets from UCI Machine learning [13]. The experiments upon these 4 datasets show that adult and breast cancer datasets have frequent patterns while on other hand, census and IPUMS datasets do not have frequent patterns. In previous paper [11] there is no distinction between frequent pattern and strong discrimination rule, while in this paper there is distinction between finding frequent pattern and strong discrimination rules. AOI-HEP as data mining technique has opportunity to be more explored such as mining similar pattern [14], inverse discovery learning, learning more than 2 datasets, multidimensional view, learning other knowledge rules and so on [15].

2. AOI-HEP FREQUENT PATTERN ALGORITHM

AOI-HEP frequent pattern algorithm will consist 2 algorithms such as AOI characteristic rule algorithm [16] and HEP frequent pattern algorithm as seen in Figures 1 and 2 respectively. AOI characteristic rule algorithm will be run twice with input two datasets as horizontal partitions of the dataset and as usuall AOI characteristic rule algorithm, uses concept hierarchy as background knowledge for data generalization. AOI characteristic rule algorithm will eliminate distinct attributes and tuples until they are less or equal than attribute and rules thresholds respectively [17] and have output two rulesets for each two input datasets. These two rulesets will be input for HEP frequent pattern algorithm in Figure 2 which apply Cartesian product between these two rulesets and the non frequent pattern in Cartesian product result will be eliminated.

```
Input: dataset, concept hierarchies, attribute threshold,rule threshold
Output: characteristic rule of learning task, {R_i^1} , {R_j^2}, num_attr, |D2|,|D1|
1   For each of attribute Ai (1≤i≤n, where n= # of attributes)in the generalized relation GR
2   { While #_of_distinct_values_in_attribute_Ai > threshold
3     {If no higher level concept in concept hierarchy for attr Ai
4         {  remove attribute Ai   }
5       Else { substitute the value of Ai by its corresponding minimal generalized concept}
6      Merge identical tuples
7     }
8   }
9   While #_of_tuples in GR > threshold
10  { Selective generalize attributes
11    Merge identical tuples
12  }
```

Figure 1. AOI Characteristic Rule Algorithm

```
Input  :{R_i^1} , {R_j^2}, num_attr,|D2|,|D1|, GR_threshold
Output : R_j^2,|R_j^2|,(|R_j^2|/|D2|),R_i^1,|R_i^1|,(|R_i^1|/|D1|),HEP_GR
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_i^1[x] ⊂ R_j^2[x] ) SLV=SLV+0.4
8              If(R_i^1[x]!= R_j^2[x] and R_i^1[x] ⊂ 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>=num_attr-1
10           HEP_GR=(|R_j^2|/|D2|)/(|R_i^1|/|D1|)
11         If HEP_GR > GR_threshold
12           Print R_j^2,|R_j^2|,(|R_j^2|/|D2|),R_i^1,|R_i^1|,(|R_i^1|/|D1|),HEP_GR,SLV
13       }
14     }
15  }
```

Figure 2. HEP Frequent Pattern Algorithm

In Figure 2, GR_threshold has default between 0 and 100, attribute num_attr 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 2 and F is a counter for AOI-HEP frequent patterns which is indentified by SLV=0.5 as shown in line number 8 Figure 2. The outputs from HEP algorithm are R_j^2, $|R_j^2|$, ($|R_j^2|$/|D2|) as support target dataset, R_i^1, $|R_i^1|$, ($|R_i^1|$/|D1|) as support contrasting dataset, GrowthRate (HEP_GR) and SLV value. Moreover, 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, since rules with ANY values are less meaningful and do not offer meaningful interpretation. Furthermore, statement in line number 9 is used to eliminate non frequent pattern, where Equations SLV>=(num_attr-1)*0.5+0.4 and SLV<=(num_attr-1)*0.5+2.1 are recognized as minimum and maximum SLV value for frequent pattern.

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

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, the options are :

a. 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.
b. 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.
c. If hierarchy level and values are the same and the attributes values are not ANY, LV=2.
d. If hierarchy level and values are the same and the attributes values are ANY, LV=2.1.

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 ⊂ R1).

3. MINING FREQUENT PATTERN

Frequent pattern is a combination of feature patterns that appear in dataset with frequency not less than a user-specified threshold [1] and the frequent pattern synonym with large pattern was first proposed for market basket analysis in the form of association rules [4]. Mining frequent patterns has been done in data stream with DSCL algorithm [18] and Top-K Closed [19]. 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 [5-7]. 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 ⊂ 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 ⊂ 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 2. 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 [3], [6].

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 [5-7]. This is called an essential Emerging Patterns (eEP) [6]. 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 [3], [6]. 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. 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% [20] or if they are similar in at least 90% of the features [21].

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 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. 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:

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

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 frequent pattern 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 frequent pattern algorithm in Figure 2.

4. HEP GROWTH RATE

Besides eliminating patterns with similarity, the large number of frequent pattern will be eliminated by the growth rate function $GR\{R_i^1, R_j^2\}$ with given a GrowthRate threshold and 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 (∞) [22].

$$\text{GR(X,Y)} = \frac{\text{Support D2(X)}}{\text{Support D1(Y)}} = \frac{\text{Count R2(X)} \ / \ |\text{D2}|}{\text{Count R1(Y)} \ / \ |\text{D1}|} \quad (2)$$

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.

Growth rate GR{ R_i^1, R_j^2 } is shown in line number 10 of HEP algorithm in Figure 2 is used to discriminate between datasets D2 and D1. This growth rate which is calculated using Equation 2, 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.

5. AOI-HEP MINING FREQUENT PATTERN EXPERIMENTS

Experiments used adult, breast cancer, census and IPUMS datasets from the UCI machine learning repository with the number of instances are 48842, 569, 2458285 and 256932 respectively [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.

Each dataset has concept hierarchies built from five chosen attributes with a minimum concept level of three. The attributes in concept hierarchies for adult dataset include workclass, education, marital-status, occupation, and native-country attributes [11], and the attributes in concept hierarchies for the breast cancer dataset contains attributes i.e. clump thickness, cell size, cell shape, bare nuclei and normal nucleoli attributes. Meanwhile, class, marital status, means, relat1 and yearsch attributes, were given to concept hierarchies for the Census dataset and the attributes in concept hierarchies for the IPUMS dataset consists of relateg, marst, educrec, migrat5g and tranwork attributes.

Table 1. Ruleset R2 for Learning Government Concept at Workclass Attribute

No	Education	Marital	Occupation	Country	Instances	Support
0	Intermediate	ANY	ANY	ANY	3454	80.53%
1	ANY	ANY	ANY	America	786	18.33%
2	Advanced	ANY	ANY	Asia	30	00.70%
3	Advanced	ANY	ANY	Europe	17	00.40%
4	Basic	Married-spouse	Services	Europe	1	00.02%
5	Advanced	Married-spouse	Services	Antartica	1	00.02%

Table 2. Ruleset R1 for Learning Non Government Concept at Workclass Attribute

No	Education	Marital	Occupation	Country	Instances	Support
0	7th-8th	Widowed	Tools	United-states	1	07.14%
1	HS-grad	Never-married	ANY	United-states	4	28.57%
2	HS-grad	Married-civ-spouse	ANY	ANY	5	35.71%
3	Assoc-adm	Married-civ-spouse	Tools	United-states	1	07.14%
4	Some-college	Married-civ-spouse	ANY	United-states	2	14.29%
5	Some-college	Married-spouse-absent	Tools	United-states	1	07.14%

Table 3. Ruleset R2 for Learning About Aver Clump Concept from "Clump Thickness" Attribute of Breast Cancer Dataset

No	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	instances	Support
0	ANY	ANY	ANY	ANY	496	93.06%
1	Medium Size	Small Shape	ANY	About Aver Nucleoli	3	0.56%
2	Very Large Size	ANY	ANY	ANY	19	3.56%
3	Medium Size	Large Shape	Above Aver Nuclei	ANY	7	1.31%
4	Very Large Size	Medium Shape	ANY	Very Large Nucleoli	3	0.56%
5	Large Size	Very Large Shape	Very Large Nuclei	ANY	5	0.94%

Table 4. Ruleset R1 for Learning About Aver Clump Concept from "Clump Thickness" Attribute of Breast Cancer Dataset

No	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	instances	Support
0	ANY	ANY	ANY	ANY	277	95.85%
1	Small Size	Large Shape	Very Large Nuclei	Very Large Nucleoli	1	0.35%
2	Medium Size	Very Large Shape	ANY	Above Aver Nucleoli	5	1.73%
3	Large Size	Very Large Shape	ANY	ANY	4	1.38%
4	Very Large Size	Small Shape	Medium Nuclei	Very Large Nucleoli	1	0.35%
5	Large Size	Small Shape	Medium Nuclei	Large Nucleoli	1	0.35%

Each dataset was divided into two sub datasets based on learning the high level concept in one of their attributes. 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. In the adult dataset, we learn by discriminating between the "government" (4289 instances) and "non government" (14 instances) concepts of the "workclass" attribute [14] in datasets D2 and D1 respectively. In the breast cancer dataset, we learn by discriminating between "aboutaverclump" (533 instances) and "aboveaverclump" (289 instances) concepts of the "clump thickness" attribute in datasets D2 and D1 respectively. Meanwhile Census dataset learns "green" (1980 instances) and "no green" (809 instances) concepts of the "means" attribute for datasets D2 and D1 respectively. Finally, the IPUMS dataset learns "unmarried" (140124 instances) and "married" (77453 instances) concepts of the "marst" attribute as datasets D2 and D1 respectively.

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. The AOI-HEP frequent pattern application was run 4 times as the number of experimental datasets and with the attribute and rule thresholds 6 and have a running time of approximately 3, 3, 4 and 13 seconds respectively. By running AOI-HEP application with input adult, breast cancer, census and IPUMS datasets, we have rulesets R2 and R1 with 6 tuples (rules) each, include number of instances for each tuple (rule) and support for each rule. Each table has four attributes (m in Equation 1) which are from five chosen attributes minus 1 learning attribute. Incredibly, the extraordinary running time of 13 seconds with the input IPUMS dataset happened because IPUMS has huge instances learning dataset's unmarried and married concepts with 140124 and 77453 instances respectively.

Because of page limitation and the result of experiment then only rulesets R2 and R1 from adult and breast cancer datasets which are shown between Tables 1 and 4. The results of running the AOI-HEP frequent pattern application show that there are only 2 and 1 finding frequent patterns from adult and breast cancer datasets which are shown between Tables 5 to 12 and 13 respectively, while there is no frequent pattern from census and IPUMS datasets. Based on between Tables 5 and 12, the finding 2 frequent patterns from adult dataset are rules number 0 and 1, in Table 1 and they are:

a. Adult which have government workclass with an intermediate education (3454/4289=80.53%).
b. Adult which have government workclass with America as a native country (786/4289=18.33%).

Table 5. Frequent Pattern for Rulesets R_3^1 to R_0^2 with HEP GR= (3454/4289)/ (1/14) =0.80532/0.07143=11.27442

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_0^2	Intermediate	ANY	ANY	ANY	3454	80.53%
R_3^1	Assoc-adm	Married-civ-spouse	Tools	United-states	1	07.14%
LV	0.5	0.5	0.5	0.5	SLV=2	11.27%

Table 6. Frequent Pattern for Rulesets R_5^1 to R_0^2 with HEP GR= (3454/4289)/(1/14) =0.80532/0.07143=11.27442

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_0^2	Intermediate	ANY	ANY	ANY	3454	80.53%
R_5^1	Some-college	Married- spouse-absent	Tools	United-states	1	07.14%
LV	0.5	0.5	0.5	0.5	SLV=2	11.27%

Table 7. Frequent Pattern for Rulesets R_0^1 to R_1^2 with HEP GR= (786/4289)/(1/14) =0.1833/0.07143=2.57

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_1^2	ANY	ANY	ANY	America	786	18.33%
R_0^1	7th-8th	Widowed	Tools	United-states	1	07.14%
LV	0.5	0.5	0.5	0.5	SLV=2	2.57%

Table 8. Frequent Pattern for Rulesets R_3^1 to R_1^2 with HEP GR= (786/4289)/(1/14) =0.1833/0.07143=2.57

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_1^2	ANY	ANY	ANY	America	786	18.33%
R_3^1	Assoc-adm	Married-civ-spouse	Tools	United-states	1	07.14%
LV	0.5	0.5	0.5	0.5	SLV=2	2.57%

Table 9. Frequent Pattern for Rulesets R_5^1 to R_1^2 with HEP GR= (786/4289)/(1/14) =0.1833/0.07143=2.57

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_1^2	ANY	ANY	ANY	America	786	18.33%
R_5^1	Some-College	Married-spouse-absent	Tools	United-states	1	07.14%
LV	0.5	0.5	0.5	0.5	SLV=2	2.57%

Table 10. Frequent Pattern for rulesets R_1^1 to R_0^2 with HEP GR= (3454/4289)/(4/14) =0.80532/0.28571=2.81861

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_0^2	Intermediate	ANY	ANY	ANY	3454	80.53%
R_1^1	HS-Grad	Never-married	ANY	United-states	4	28.57%
LV	0.5	0.5	2.1	0.5	SLV=3.6	2.82%

Table 11. frequent pattern for rulesets R_4^1 to R_0^2 with HEP GR= (3454/4289)/(2/14) =0.80532/0.14286=5.63721

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_0^2	Intermediate	ANY	ANY	ANY	3454	80.53%
R_4^1	Some-College	Married-civ-spouse	ANY	United-states	2	14.29%
LV	0.5	0.5	2.1	0.5	SLV=3.6	5.64%

Table 12. Frequent Pattern for Rulesets R_4^1 to R_1^2 with HEP GR= (786/4289)/(2/14) =0.1833/0.1429=1.28

Rulesesets	Education	Marital	Occupation	Country	Instances	Support
R_1^2	ANY	ANY	ANY	America	786	18.33%
R_4^1	Some-College	Married-civ-spouse	ANY	United-states	2	14.29%
LV	0.5	0.5	2.1	0.5	SLV=3.6	1.28%

Table 13. Frequent Pattern for Rulesets R_4^1 to R_2^2 with HEP GR= (19/533)/(1/289) =0.356/0.035=10.30

Rulesesets	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	Instances	Support
R_2^2	Very Large Size	ANY	ANY	ANY	19	3.56%
R_4^1	Very Large Size	Small Shape	Medium Nuclei	Very Large Nucleoli	1	0.35%
LV	2	0.5	0.5	0.5	SLV=3.5	10.30%

Meanwhile, based on Table 13, the only finding frequent pattern from breast cancer dataset is rule number 2 in Table 3 and it is : Breast cancer which have clump thickness type of AboutAverClump with cell size of Very Large Size (19/533=3.56%).

The two of adult dataset's frequent patterns are the highest score rules with 3454 and 786 instances in Table 1, while the only one breast cancer dataset's frequent pattern is the second highest score rule with 19 instances which are much different with the first rule with 496 instances in Table 3. However, this breast cancer's frequent pattern fulfill of AOI-HEP frequent pattern where having:

a. Maximum subsumption target (superset) into contrasting (subset) datasets (contrasting $\subset$ target) Table 13 shows that rule R_2^2 as target (superset) dataset has maximum subsumption ($\subset$) into rule R_4^1 as contrasting (subset) dataset which is showed with maximum LV=0.5 and SLV value is 3.5.
b. Large HEP frequent pattern growth rate and support in target dataset [11]. Table 13 shows that frequent pattern has large HEP frequent pattern 10.30% (19/533)/(1/289) =0.356/0.035=10.30%) and large support rule R_2^2 as target (superset) dataset, where support R_2^2 as target (superset) dataset (3.56%) is large than R_4^1 as contrasting (subset) dataset (0.35%).

The results running of AOI-HEP frequent pattern application upon adult dataset can be seen between Tables 5 and 12 where:

a. There are 5 SLV value frequent patterns with full similarity subsumption LV=0.5 as shown between tables 5 and 9.
b. There are 3 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 between Tables 10 and 12.

Meanwhile, The results running of AOI-HEP frequent pattern application upon breast cancer dataset can be seen in Table 13 where: There are 1 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.

Based on finding frequent patterns between Tables 5 and 13, the strong discrimination rule can be formulated:

1. 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).
2. 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).
3. There are 2.57 growth rates adult dataset with 18.33% frequent pattern in government workclass (with an America as native country) and 7.14% infrequent pattern in non government workclass (with 7^{th}-8^{th} education, widowed marital status, tools occupation and from the United States).
4. There are 2.57 growth rates adult dataset with 18.33% frequent pattern in government workclass (with an America as native country) 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).
5. There are 2.57 growth rates adult dataset with 18.33% frequent pattern in government workclass (with an America as native country) 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).
6. 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).
7. 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).
8. There are 1.28 growth rates adult dataset with 18.33% frequent pattern in government workclass (with an America as native country) and 14.29% infrequent pattern in non government workclass (with some-college education, married-civ-spouse marital status and from the United States).
9. There are 10.30 growth rates breast cancer dataset with 3.56% frequent pattern in clump thickness type of AboutAverClump (with cell size of VeryLargeSize) and 0.35% infrequent pattern in clump thickness type of AboveAverClump (with cell size of VeryLargeSize, cell shape of SmallShape, Bare Nuclei of MediumNuclei and Normal Nucleoli of VeryLargeNucleoli).

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 between Tables 5 and 13 show as strong discriminating power where they have large growth rates (between 1.28 and 11.2774) and supports in target (D2) datasets (between 3.56% and 80.53%). Moreover, they have small supports in contrasting (D1) dataset between 0.35% and 28.57% where each of the support in contrasting (D1) dataset is less than the support in target (D2) dataset.

6. AOI-HEP JUSTIFICATION

Since AOI-HEP was proposed based on previous data mining techniques such as Attribute oriented Induction (AOI) and Emerging Pattern (EP) then AOI-HEP will be distinguished with AOI and EP. Since AOI-HEP is combination between two data mining techniques such as AOI and EP, then AOI-HEP is better than these two data mining techniques. Obviously, AOI-HEP is perfect since its mixture of strength of these two data mining techniques. Table 14 shows the performance metric with number of rules resulted and processing time among AOI-HEP, AOI and EP.

In number of rules resulted, Table 14 shows AOI-HEP has superiority rather than AOI and EP where AOI-HEP has a few number of rules resulted whilst AOI and EP have intermediate and many number of rules resulted respectively. AOI-HEP has superiority with a few number of rules resulted because AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm, as mentioned in Section 5. Moreover, the cartesian product are eliminated with frequent pattern. Meanwhile, EP

has weakness in many number of rules resulted since EP deals with low level data which have many low level rules. AOI-HEP and AOI use concept hierarchy to generalize from low level data into high level data, and as a result AOI-HEP and AOI mining high level rules which are less than low level rules. Thus, AOI-HEP has a few number of rules resulted because AOI-HEP mining high level rules which are less than low level rules, applies cartesian product and eliminates it by determining type of HEP.

However, in time to process as shown in Table 14, AOI-HEP has medium classification since AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm. Performance metric in Table 14 shows AOI-HEP and AOI have better performance in time to process against EP, since both of them deal with high level data. Since EP deals with low level data which have many low level rules then EP has weakness with slow performance in time to process, while AOI-HEP and AOI use concept hierarchy to generalize from low level data into high level data where high level data have less data rather than low level data. Obviously, time to process high level data will have better performance since deal with less data and the other hand, time to process low level data will have slow performance since deal with huge data. Rather than AOI, AOI-HEP has lower performance in time to process, since AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm, and cartesian product are eliminated with frequent patterns.

Table 14. Performance Metric Among AOI-HEP, AOI and EP

	AOI-HEP	AOI	EP
Number of rules resulted	Few	Intermediate	Many
Processing time	Medium	Fastest	Slow

7. CONCLUSION

Mining HEP frequent patterns with AOI-HEP are influenced by learning on high level concept in one of chosen attribute and extended experiment upon adult dataset where learn on marital-status attribute showed that there is no finding frequent pattern. The research for mining HEP frequent patterns with AOI-HEP is interested to be extended where mining HEP frequent patterns can be done by searching on every each attribute in dataset for finding possible frequent patterns. Moreover, since there are more than 2 concepts in high level attribute concept, then mining HEP frequent patterns need to be extended to discriminate more than 2 rulesets. Furthermore, the experiments showed that there are candidate HEP frequent patterns in census dataset in reverse condition, then mining HEP frequent pattern should be extended to mining inverse patterns. This research should need more extended research and experiments in order to find justification of this mining approach with other frequent pattern algorithms, the input datasets should be applied to other frequent pattern algorithm in order to find the differences in term of performance, type and kind of patterns, advantages and disadvantages.

ACKNOWLEDGEMENTS

This research is supported under Program of research incentive of national innovation system (SINAS) from Ministry of Research, Technology and Higher Education of the Republic of Indonesia, decree number 147/M/Kp/IV/2015, Research code: RD-2015-0020.

REFERENCES

[1] J. Han, *et al.*, "Frequent pattern mining: current status and future directions," *Data Min Knowl Disc*, vol/issue: 15(1), pp. 55-86, 2007.
[2] J. Han, *et al.*, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," *Data Min. Knowl. Discov.*, vol/issue: 8(1), pp. 53-87, 2004.
[3] 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.
[4] R. Agrawal, *et al.*, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec*, vol/issue: 22(2), pp. 207-216, 1993.
[5] K. Ramamohanarao, *et al.*, "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.
[6] 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.
[7] 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.

[8] C. C. Aggarwal, "An introduction to Frequent Pattern Mining," *Frequent Pattern Mining*, C. C. Aggarwal and J. Han (eds.), Springer, pp. 1-17, 2014.
[9] C. C. Aggarwal, *et al.*, "Frequent pattern mining Algorithm: A Survey," *Frequent Pattern Mining*, C. C. Aggarwal and J. Han (eds.), Springer, pp. 19-64, 2014.
[10] A. Zimek, *et al.*, "Frequent Pattern Mining Algorithm for Data clustering," *Frequent Pattern Mining*, C. C. Aggarwal and J. Han (eds.), Springer, pp. 403-423, 2014.
[11] S. Warnars, "Mining Frequent Pattern with Attribute Oriented Induction High level Emerging Pattern (AOI-HEP)," in *Proceedings of IEEE the 2nd International Conference on Information and Communication Technology (IEEE ICoICT 2014), Bandung, Indonesia*, pp. 144-149, 28-30 May 2014.
[12] 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.
[13] A. Frank and A. Asuncion, "UCI Machine Learning Repository," Irvine, CA, University of California, School of Information and Computer Science, 2010. [http://archive.ics.uci.edu/ml].
[14] S. Warnars, "Mining Frequent and Similar Patterns with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) Data Mining Technique," *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*, vol/issue: 3(11), pp. 266-276, 2014.
[15] S. Warnars, "Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) future research," in *Proceedings of IEEE the 8th International Conference on Information & Communication Technology and Systems (ICTS), Surabaya, Indonesia*, pp. 13-18, 24-25 September 2014.
[16] J. Han, *et al.*, "Knowledge discovery in databases: An attributed approach," in *Proceeding of the 18th International Conference on Very Large Data Bases*, pp. 547-559, 1992.
[17] Y. Cai, *et al.*, "An attribute-oriented approach for learning classification rules from relational databases," in *Proceedings of 6th International Conference on Data Engineering*, pp. 281-288, 1990.
[18] Z. C. Seng, *et al.*, "Frequent itemsets mining based on concept lattice and sliding windows," *Telkomnika*, vol/issue: 11(8), pp. 4780-4787, August 2013.
[19] M. Yimin, *et al.*, "An efficient algorithm for mining Top-k closed frequent item sets over data streams over data streams," *Telkomnika*, vol/issue: 11(7), pp. 3759-3766, July 2013.
[20] R. Danger, *et al*., "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.
[21] A. Y. R. Gonzalez, *et al.*, "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.
[22] J. Li, *et al*., "Instance-based classification by Emerging Patterns," in *proceeding of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pp.191-200, 2000.

BIOGRAPHY OF AUTHOR

Head of Information system concentration at Doctor of Computer Science, Bina Nusantara university (www.dcs.binus.ac.id) . Having bachelor degree in Computer Science between July 1991-April 1995 with information system topic, master degree in Information Technology from university of Indonesia between July 2004-January 2007 with data warehouse thesis topic. Between September 2008-December2012 did PhD computer science at the Manchester Metropolitan university, United Kingdom with data mining thesis topic. Have been IT lecturer since 1995 and have Indonesian national academic position rank (Jenjang jabatan akademik) Associate professor (LektorKepala, 550 points) since 2007. I have research interest on field such as BigData, Parallel computing, Data Mining, Machine Learning, intelligent application and so on. My publications can be reached at https://www.researchgate.net/profile/Harco_Leslie_Hendric_Spits_Warnars2.