Easy Understanding for Mining Discriminant Itemset with Emerging Patterns

Harco Leslie Hendric Spits Warnars[1]
Ford Lumban Gaol[2]
Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia
shendric@binus.edu[1],
fgaol@binus.edu[2]

Nesti Fronika Sianipar
Research Interest Group
Food Biotechnology
Bina Nusantara University
Jakarta, Indonesia
nsianipar@binus.edu

Bahtiar Saleh Abbas
Industrial Engineering
Bina Nusantara University
Jakarta, Indonesia
bahtiars@binus.edu

Horacio Emilio Perez Sanchez
Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC)
Universidad Catolica de Murcia (UCAM)
Guadalupe, Spain
hperez@ucam.edu

***Abstract*—This paper will give basic knowledge understanding how to discriminate between two datasets with Emerging Patterns (Eps) upon famous weather dataset. This paper didn't use previous data mining techniques such as border-based algorithm or so on, but only to give the systematic basic knowledge understanding to discriminate between two datasets by finding score of support, growthrate and confidence. The discrimination will give example to differ between two datasets with itemset/pattern which consist one attribute/dimension or multidimensional attributes and of course by finding Jumping EPs (JEPs) as strong discrimination with proofing of their finding growthrate with confidence score. The finding discrimination shows as uninterested or interested discrimination when having confidence below or greater than 50%. These discrimination results can be used for classification purposes as well.**

Keywords—Data Mining; Emerging Patterns; Jumping Emerging Patterns; Discrimination; Confidence; Growthrate

I. INTRODUCTION

Emerging Patterns (EPs) data mining technique which was found in 1999 has been mixed and implement in data mining world in order for finding pattern, discriminate and doing classification. EPs has been used and combined with other algorithms like EPs are combined with Decision tree technique CART-based algorithm in order to discover relevant EPs for classification that consist six steps where CART trees replace border-based algorithm function [4]. EP had been combined with Attribute Oriented Induction (AOI) [8,9] as Attribute Oriented Induction High level Emerging Patterns(AOI-HEP) in order to find frequent and similar patterns[5,6,7,16]. Also, EPs had been implemented in table of Database Management Systems (DBMS) particularly of relational table databases. MRDM (Multi Relational Data Mining) method [13,14,15] or can be called Mr-EP (Multi Relational Emerging Pattern) [10,11,12] are EPs algorithm which discovers EPs from data scattered in multiple tables of a relational database.

However, there was some difficulty from my bachelor and master students to understand this simple and powerful strong discrimination machine. Some previous textbook or papers didn't give detail example explanation for them. Hope this paper will open their mind, how easy for dealing with data mining particularly for doing discrimination to datasets and more than that by using EPs will be easily to discriminate between datasets.

Section 2 will give basic equations for support, growth and confidence for EPs technique. Section 3 give understanding for mining discriminant simple dimension, multidimensional and Jumping EPs (JEPs). Section 4 explains how to find confidentiality as justification for mining itemset/pattern with simple dimension, multidimensional and JEPs. In this explanation, we will not refer to previous EPs algorithms such as border-based MBD-LLBORDER algorithm[2] but prefer to discuss based on understanding knowledge how to implement support, growthrate and confidence.

II. SUPPORT, GROWTHRATE AND CONFIDENCE

EPs is scored with Support score as shown in equation (1) in order to count the number of itemset or pattern in dataset. Dataset is a set D of transaction D={i1, i2, …, iN}, where i is attribute and content N attributes from i1 to iN. For example, table 1 or 2 has N=5 since there are 5 attributes and represented as D={Outlook, Temperature, Humidity, Windy, Class}, where i1=Outlook, i2=Temperature, i3=Humidity, i4=Windy, i5=Class. Meanwhile, itemset or pattern is a set X of transaction X={j1,…,jM}, where M<=N, $X \subset D$ and j is attribute and content M attributes from j1 to jM. M<=N, since the number attributes of itemset X should <= than dataset D and $X \subset D$, where itemset X are a proper subset of dataset D, or dataset D are a proper superset of itemset X, where the number of records/tuples/rows of itemset X are part of dataset D.

$$\text{suppD(X)} = \frac{count_D(X)}{|D|} \quad (1)$$

where :
D = Dataset
X = Itemset or pattern, where $X \subset D$
$|D|$ = total number of instances in dataset D
suppD(X) = support in dataset D containing itemset X
countD(X) = the number of transactions in dataset D containing itemset X.

978-1-5090-4048-3/17/$31.00 ©2017 IEEE

As mentioned before, EPs as strong discrimination are defined with EPs(target) as GrowthRate score for itemset X as shown in equation (2) since itemset X will be mined from both datasets like target/positive/suppD2(X) and background/negative/suppD1(X). The EPs will be called as EPs of learning based on support of dividend or target or positive, which is showed as $\text{supp}_{\mathbf{D2}}(X) = \text{count}_{\mathbf{D2}}(X)/|D2|$. EPs(target) can be also called as EPs(D2) as long as the name of D2 as target dataset.

Growthrate for EPs(target) in equation (2) is used in order to score how many times or how many growth the differences of itemset X between 2 datasets. As shown in equation (2) where EPs are associated with division of two datasets, where dividend is called target or positive dataset which is showed as $\text{supp}_{\mathbf{D2}}(X)=\text{count}_{\mathbf{D2}}(X)/|D2|$ whilst divisor is called background or negative dataset which is showed as $\text{supp}_{\mathbf{D1}}(X) = \text{count}_{\mathbf{D1}}(X)/|D1|$.

Growthrate for EPs(target) in equation (2) as discriminator score should have non empty support score and fail for to do so will be subjected as Jumping EPs (JEPs) where because score of one of support=0 and literally make it become jumping score since one of support score=0, where 0/n=0 or n/0=undefined. GrowthRate for EPs(target) in equation (2) can have 3 different results and they are:

1. EPs(target)= ∞ (or 0), if dividend=suppD2(X)=0
   Its have been known that in division equation, if dividend= suppD2(X)=0 then will have division result=0, because any number which has dividend 0 will have result=0. For example: GR(X)=0/5=0. This is recognized as Jumping EPs since score of one of the support=0 or dividend=suppD2(X)=0 and it will scored with infinity(∞).
2. EPs(target)=∞ (or undefined), if divisor=suppD1(X)=0
   Thus, any number divided by 0 will have undefined(∞) result and there are 2 options for divisor=suppD1(X)=0 and they are:
   a) If suppD2(X)=0=dividend=0 then obviously we can not divide 0 with 0. For example: GR(X)=0/0=∞.
   b) If suppD2(X)=dividend ≠ 0 then the same as well we can not divide non 0 with 0. For example: GR(X)=5/0=∞.

   They are recognized as JEPs since score of one of the support=0.
3. EPs(target)>0
   if dividend=suppD2(X)=divisor=suppD1(X) ≠0. Since there will have many growthrate result, then using Growthrate threshold will limit the growthrate result and more than that we just only interest with strong growthrate.

$$\text{EPs(target)} = \frac{\text{target}}{\text{background}} = \frac{\text{Positive}}{\text{Negative}} = \frac{\text{supp}_{D2}(X)}{\text{supp}_{D1}(X)} = \frac{\frac{\text{count}_{D2}(X)}{|D2|}}{\frac{\text{count}_{D1}(X)}{|D1|}} \quad (2)$$

where :

∞ = infinity (JEPs), when (n/0=undefined) or (0/n=0)

suppD1(X) = support in dataset D1 containing itemset X (see equation (1))

suppD2(X) = support in dataset D2 containing itemset X (see equation (1))

Meanwhile, Growth rate of an EPs(target) as shown in equation 2 should be assessed in order to justify the finding growthrate score or justify the finding itemset between 2 datasets. There are 3 confidence of predictions equations such as (3), (4) or (5) that one can justify the confidence of EPs [1].

$$\text{Conf(EPs(target))} = \frac{\text{EPs(target)}}{\text{EPs(target)}+1} \quad (3)$$

$$\text{Conf(EPs(target))} = \frac{\text{supp}_{D2}(X)}{\text{supp}_{D2}(X)+\text{supp}_{D1}(X)} \quad (4)$$

$$\text{Conf(EPs(target))} = \frac{\text{EPs(target)}*\text{supp}_{D1}(X)}{\text{EPs(target)}*\text{supp}_{D1}(X)+\text{supp}_{D1}(X)} \quad (5)$$

III. Discrimination by Mining in Single dimensional, multidimensional and JEPs

For easy understanding purposes, then this paper discussion will use famous weather dataset as seen either in tables 1 or 2 below, with 5 attributes such as Outlook, Temperature, humidity, Windy and Class including 14 tuples/records/rows. Data in table 1 was saved in text file where between each attribute was split with sign of comma "," and the first line/row/tuple is the title/header of attributes which is not part of the data and its "**Outlook,temperature,Humidity,Windy,Class**". Meanwhile, data in table 2 was saved as table in Database Management Systems (DBMS) such as MySQL, Oracle, SQLserver and so on and data in DBMS will be accessed with select SQL statement. Since there is the different the way to save the input data where between text file dataset and DBMS table, then the process of EPs algorithm for accessing each type of data will be different. Each attribute in both of tables 1 and 2 have different distinct value, where attributes like Outlook and Temperature have 3 different distinct values such as Sunny, Overcast, Rain and Mild, Cool, Hot, respectively. Meanwhile, other attributes such as Humidity, Windy and Class have 2 different distinct values such as Normal and High, True and False, and P and N, respectively.

The process of learning of both of weather dataset in tables 1 and 2, have been known in data mining technique as supervised learning where the input data should be learned and managed in order to get the patterns as mining result. Meanwhile the unsupervised technique with this EPs can be applied since the input data do not need to be learned and managed where automatically will produce patterns as mining result. Thus, the EPs for mining pattern in data can be divided into 4 types EPs algorithms where each of them should have different algorithms. The first and second of EPs algorithms are used for mining text file dataset, while the third and fourth are used for mining DBMS table database. They are:

1. Supervised EPs Text File.
2. Unsupervised EPs Text Filet.
3. Supervised EPs DBMS table.
4. unSupervised EPs DBMS table.

TABLE I. WEATHER DATASET IN TEXT FILE

Outlook,temperature,Humidity,Windy,Class
Sunny,Mild,Normal,True,P
Sunny,Cool,Normal,False,P
Overcast,Mild,High,True,P
Overcast,Hot,High,False,P
Overcast,Cool,Normal,True,P
Overcast,Hot,Normal,False,P
Rain,Mild,High,False,P
Rain,Cool,Normal,False,N
Rain,Mild,Normal,False,N
Sunny,Hot,High,True,N
Sunny,Hot,High,False,N
Sunny,Mild,High,False,N
Rain,Mild,High,True,N
Rain,Cool,Normal,True,N

TABLE II. WEATHER DATASET IN TABLE DBMS

Outlook	Temperature	Humidity	Windy	Class
Sunny	Mild	Normal	True	P
Sunny	Cool	Normal	False	P
Overcast	Mild	High	True	P
Overcast	Hot	High	False	P
Overcast	Cool	Normal	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Hot	High	True	N
Sunny	Hot	High	False	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N
Rain	Cool	Normal	True	N

Before doing EPs then the weather dataset in both tables 1 and 2 should be learning based on their distinct value. For example, if attribute Class is selected then the data in table 1 or 2 is split into 2 sub datasets, P and N as shown in table 3 where attribute Class has 2 different distinct values such as P and N with number of records/tuples/rows 9 and 5 respectively. If attribute Outlook is selected then the data in table 1 or 2 is split into 3 sub datasets, Sunny, Overcast and Rain as shown in table 4 where attribute Outlook has 3 distinct values such as Sunny, Overcast and Rain with number of records/tuples/rows 5,4 and 5 respectively.

TABLE III. WEATHER DATASET IN TEXT FILE WITH CLASS ATTRIBUTE SELECTION

Class=P	**Outlook,temperature,Humidity,Windy**
	Sunny,Mild,Normal,True
	Sunny,Cool,Normal,False
	Overcast,Mild,High,True
	Overcast,Hot,High,False
	Overcast,Cool,Normal,True
	Overcast,Hot,Normal,False
	Rain,Mild,High,False
	Rain,Cool,Normal,False
	Rain,Mild,Normal,False
Class=N	**Outlook,temperature,Humidity,Windy**
	Sunny,Hot,High,True
	Sunny,Hot,High,False
	Sunny,Mild,High,False
	Rain,Mild,High,True
	Rain,Cool,Normal,True

TABLE IV. WEATHER DATASET IN TEXT FILE WITH OUTLOOK ATTRIBUTE SELECTION

Outlook=Sunny	**temperature,Humidity,Windy,Class**
	Mild,Normal,True,P
	Cool,Normal,False,P
	Hot,High,True,N
	Hot,High,False,N
	Sunny,Mild,High,False,N
Outlook=Overcast	**temperature,Humidity,Windy,Class**
	Mild,High,True,P
	Hot,High,False,P
	Cool,Normal,True,P
	Hot,Normal,False,P
Outlook=Rain	**temperature,Humidity,Windy,Class**
	Mild,High,False,P
	Cool,Normal,False,P
	Mild,Normal,False,P
	Mild,High,True,N
	Cool,Normal,True,N

The implementation experiment will be explained into 3 sub section where first sub section for mining 1 dimension itemset, second sub section for mining multidimensional itemset and third sub section for mining Jumping EPs. The process discrimination will only use the learning dataset in table 3 as learning dataset with attribute Class.

A. Discrimination by mining 1 dimension itemset

For example, the mining will mine itemset X for Outlook=Sunny. As shown in table 3, in class P there are 2 tuples in 1st and 2nd rows with attribute Outlook=Sunny. Meanwhile, in class N there are 3 tuples between 1st and 3rd rows with attribute Outlook=Sunny. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{\text{supp}_P(X)}{\text{supp}_N(X)}=\frac{\frac{\text{count}_P(X)}{|P|}}{\frac{\text{count}_N(X)}{|N|}}=\frac{\frac{2}{9}}{\frac{3}{5}}=\frac{0.22}{0.6}=0.37$$

$$\text{EPs(N)}=\frac{\text{supp}_N(X)}{\text{supp}_P(X)}=\frac{\frac{\text{count}_N(X)}{|N|}}{\frac{\text{count}_P(X)}{|P|}}=\frac{\frac{3}{5}}{\frac{2}{9}}=\frac{0.6}{0.22}=2.7$$

Another example, the mining will mine itemset X for Windy=False. As shown in table 3, in class P there are 6 tuples in 2nd,4th,6th till 9th rows with attribute Windy=False. Meanwhile, in class N there are 2 tuples in 2nd and 3rd rows with attribute Windy=False. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{\text{supp}_P(X)}{\text{supp}_N(X)}=\frac{\frac{\text{count}_P(X)}{|P|}}{\frac{\text{count}_N(X)}{|N|}}=\frac{\frac{6}{9}}{\frac{2}{5}}=\frac{0.67}{0.4}=1.67$$

$$\text{EPs(N)}=\frac{\text{supp}_N(X)}{\text{supp}_P(X)}=\frac{\frac{\text{count}_N(X)}{|N|}}{\frac{\text{count}_P(X)}{|P|}}=\frac{\frac{2}{5}}{\frac{6}{9}}=\frac{0.4}{0.67}=0.6$$

B. Discrimination by mining Multidimensional itemset

Mining multidimensional is extended from mining single dimension where process mining will include more than 1 attribute. For example, the mining will mine in 2 attributes/dimensions like Outlook and Windy with itemset X for Outlook=Sunny, Windy=False. As shown in table 3, in class P there is 1 tuple in 2nd row and in class N there are 2

tuples in 2nd and 3rd rows. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{1}{9}}{\frac{2}{5}}=\frac{0.11}{0.4}=0.28$$

$$\text{EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{2}{5}}{\frac{1}{9}}=\frac{0.4}{0.11}=3.6$$

Another example, the mining will mine in 3 attributes/dimensions like Outlook, Temperature and Humidity with itemset X for Outlook=Rain, Temperature=Mild, Humidity=High. As shown in table 3, in class P there is 1 tuple in 7th row and in class N there is 1 tuple in 4th row. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{1}{9}}{\frac{1}{5}}=\frac{0.11}{0.2}=0.56$$

$$\text{EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{1}{5}}{\frac{1}{9}}=\frac{0.2}{0.11}=1.8$$

C. *Discrimination by mining Jumping EPs (JEPs)*

As mentioned before, Jumping EPs will be found when one of support has 0 score, and JEPs can be mined both in 1 dimension or multidimensional. For example, the mining will mine in 1 attribute/dimension with itemset X for Outlook=Overcast. As shown in table 3, in class P there are 4 tuples in between 3rd and 6th rows and in class N there is none row. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{4}{9}}{\frac{0}{5}}=\frac{0.44}{0}=\infty$$

$$\text{EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{0}{5}}{\frac{4}{9}}=\frac{0}{0.44}=\infty$$

Another example, the mining will mine in 4 attribute/dimensions or multidimensional itemset X for Outlook=Rain, Temperature=Cool, Humidity=normal, Windy=true. As shown in table 3, in class P there is none row and in class N there is 1 tuple in 5th row. Equation (1) will be applied in both class P and N.

$$\text{EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{0}{9}}{\frac{1}{5}}=\frac{0}{0.2}=\infty$$

$$\text{EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{1}{5}}{\frac{0}{9}}=\frac{0.2}{0}=\infty$$

IV. CONFIDENCE OF DISCRIMINATION EPS

In order to justify the finding EPs then we need to convince the finding EPs score by running equations (3), (4) and (5) as confidence of EPs. The Equations will be applied based on mining in single dimension, multidimensional and Jumping EPs.

A. *Confidence EPs on discrimination of 1 dimension itemset*

We will justify the finding EPs with itemset X for Outlook=Sunny with EPs(P)=0.37 and EPs(N)=2.7. The finding of EPs(P) is uninterested EPs since the score below 1 and different for EPs(N) as interested EPs since the growthrate score greater than 1. Moreover, the strength of discrimination power is expressed by its large growth rate and support in target dataset and was called an essential Emerging Patterns (eEP) [2,3]. Next are implementation equations (3), (4), (5) in order to justify their finding EPs.

Firstly, $\text{EPs(P)}=\frac{\text{supp}_\text{P}(\text{X})}{\text{supp}_\text{N}(\text{X})}=\frac{\frac{\text{count}_\text{P}(\text{X})}{|\text{P}|}}{\frac{\text{count}_\text{N}(\text{X})}{|\text{N}|}}=\frac{\frac{2}{9}}{\frac{3}{5}}=\frac{0.22}{0.6}=0.37$

EPs(P)=0.37 is measured the confidentiality and they are:

Conf(EPs(P))=

$$\frac{\text{EPs(target)}}{\text{EPs(target)+1}}=\frac{\text{EPs(P)}}{\text{EPs(P)+1}}=\frac{0.37}{0.37+1}=\frac{0.37}{1.37}=0.270073=27\%$$

Conf(EPs(P))=

$$\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_P(X)}{supp_P(X)+supp_N(X)}=\frac{0.22}{0.22+0.6}=\frac{0.22}{0.82}=0.268293=27\%$$

Conf(EPs(P))=

$$\frac{EPs(target)*supp_{D1}(X)}{EPs(target)*supp_{D1}(X)+supp_{D1}(X)}=\frac{EPs(target)*supp_N(X)}{EPs(target)*supp_N(X)+supp_N(X)}$$

$$=\frac{0.37*0.6}{0.37*0.6+0.6}=\frac{0.222}{0.822}=0.270073=27\%$$

Secondly, $\text{EPs(N)}=\frac{\text{supp}_\text{N}(\text{X})}{\text{supp}_\text{P}(\text{X})}=\frac{\frac{\text{count}_\text{N}(\text{X})}{|\text{N}|}}{\frac{\text{count}_\text{P}(\text{X})}{|\text{P}|}}=\frac{\frac{3}{5}}{\frac{2}{9}}=\frac{0.6}{0.22}=2.7$

EPs(N)=2.7 is measured the confidentiality and they are:

Conf(EPs(N))=

$$\frac{EPs(target)}{EPs(target)+1}=\frac{EPs(P)}{EPs(P)+1}=\frac{2.7}{2.7+1}=\frac{2.7}{3.7}=0.72973=73\%$$

Conf(EPs(N))=

$$\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_N(X)}{supp_N(X)+supp_P(X)}=\frac{0.6}{0.6+0.22}=\frac{0.6}{0.82}=0.731707=73\%$$

Conf(EPs(N))=

$$\frac{EPs(target)*supp_{D1}(X)}{EPs(target)*supp_{D1}(X)+supp_{D1}(X)}=\frac{EPs(target)*supp_P(X)}{EPs(target)*supp_P(X)+supp_P(X)}$$

$$=\frac{2.7*0.22}{2.7*0.22+0.22}=\frac{0.594}{0.814}=0.72973=73\%$$

The implementation of equations (3),(4) and (5) between EPs(P)=0.37 and EPs(N)=2.7 show that EPs(P)=0.37 is uninterested since having confidence 27% below 50%. Meanwhile, EPs(N)=2.7 is interested since having confidence 73% greater than 50%. Moreover, the strength of discrimination power is expressed by its large growth rate and support in target dataset and was called an essential Emerging Patterns (eEP) [2,3].

B. Confidence EPs on discrimination of Multidimensional itemset

We will justify the finding EPs mining in 3 attributes/dimensions like Outlook, Temperature and Humidity with itemset X for Outlook=Rain, Temperature=Mild, Humidity=High with EPs(P)=0.56 and EPs(N)=1.8. The finding of EPs(P) is uninterested EPs since the score below 1 and different for EPs(N) as interested EPs since the growthrate score greater than 1. Moreover, the strength of discrimination power is expressed by its large growth rate and support in target dataset and was called an essential Emerging Patterns (eEP) [2,3]. Next are implementation equations (3), (4), (5) in order to justify their finding EPs. Firstly,

$$\text{EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{1}{9}}{\frac{1}{5}}=\frac{0.11}{0.2}=0.56$$

EPs(P)=0.56 is measured the confidentiality and they are:

$$\text{Conf(EPs(P))}=\frac{\text{EPs(target)}}{\text{EPs(target)}+1}=\frac{\text{EPs(P)}}{\text{EPs(P)}+1}=\frac{0.56}{0.56+1}=\frac{0.56}{1.56}=0.358974=36\%$$

$$\text{Conf(EPs(P))}=\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_P(X)}{supp_P(X)+supp_N(X)}=\frac{0.11}{0.11+0.2}=\frac{0.11}{0.31}=0.354839=35\%$$

$$\text{Conf(EPs(P))}=\frac{EPs(target)*supp_{D1}(X)}{EPs(target)*supp_{D1}(X)+supp_{D1}(X)}=\frac{EPs(target)*supp_N(X)}{EPs(target)*supp_N(X)+supp_N(X)}$$

$$=\frac{0.56*0.2}{0.56*0.2+0.2}=\frac{0.112}{0.312}=0.358974=36\%$$

$$\text{Secondly, EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{1}{5}}{\frac{1}{9}}=\frac{0.2}{0.11}=1.8$$

EPs(N)=1.8 is measured the confidentiality and they are:

$$\text{Conf(EPs(N))}=\frac{\text{EPs(target)}}{\text{EPs(target)}+1}=\frac{\text{EPs(P)}}{\text{EPs(P)}+1}=\frac{1.8}{1.8+1}=\frac{1.8}{2.8}=0.642857=64\%$$

$$\text{Conf(EPs(N))}=\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_N(X)}{supp_N(X)+supp_P(X)}=\frac{0.2}{0.2+0.11}=\frac{0.2}{0.31}=0.645161=64\%$$

$$\text{Conf(EPs(N))}=\frac{EPs(target)*supp_{D1}(X)}{EPs(target)*supp_{D1}(X)+supp_{D1}(X)}=\frac{EPs(target)*supp_P(X)}{EPs(target)*supp_P(X)+supp_P(X)}$$

$$=\frac{1.8*0.11}{1.8*0.11+0.11}=\frac{0.198}{0.308}=0.642857=64\%$$

The implementation of equations (3),(4) and (5) between EPs(P)=0.56 and EPs(N)=1.8 show that EPs(P)=0.56 is uninterested since having confidence 36% below 50%. Meanwhile, EPs(N)=1.8 is interested since having confidence 64% greater than 50%. Moreover, the strength of discrimination power is expressed by its large growth rate and support in target dataset and was called an essential Emerging Patterns (eEP) [2,3].

C. Confidence EPs on discrimination of Jumping EPs (JEPs)

Because JEPs have infinite/undefined (∞) score then the JEPs confidentiality can be measured with only equation (4) since there is no EPs score. We will justify the finding EPs mining in 1 attribute/dimension like Outlook with itemset X for Outlook=Overcast with EPs(P)=EPs(N)= ∞.

$$\text{Firstly, EPs(P)}=\frac{supp_P(X)}{supp_N(X)}=\frac{\frac{count_P(X)}{|P|}}{\frac{count_N(X)}{|N|}}=\frac{\frac{4}{9}}{\frac{0}{5}}=\frac{0.44}{0}=\infty$$

EPs(P)= ∞ is measured the confidentiality and it is:

$$\text{Conf(EPs(P))}=\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_P(X)}{supp_P(X)+supp_N(X)}=\frac{0.44}{0.44+0}=\frac{0.44}{0.44}=1=100\%$$

$$\text{Secondly, EPs(N)}=\frac{supp_N(X)}{supp_P(X)}=\frac{\frac{count_N(X)}{|N|}}{\frac{count_P(X)}{|P|}}=\frac{\frac{0}{5}}{\frac{4}{9}}=\frac{0}{0.44}=\infty$$

EPs(N)= ∞ is measured the confidentiality and it is:

$$\text{Conf(EPs(N))}=\frac{supp_{D2}(X)}{supp_{D2}(X)+supp_{D1}(X)}=\frac{supp_N(X)}{supp_N(X)+supp_P(X)}=\frac{0}{0+0.44}=\frac{0}{0.44}=0=0\%$$

The implementation of equation (4) for EPs(P)=EPs(N)= ∞ show that EPs(P) is interested since having 100% confidentiality, whilst EPs(N) is uninterested since having 0% confidentiality.

V. CONCLUSION

The Mining pattern with EPs can be extended to find frequent or similar patterns, by mining n-1 or n attribute in dataset respectively. Mining frequent and similar patterns are interested as well to be explored in order to find the similarity of the dataset.

REFERENCES

[1] Dong, G. and Li, J. 2005. Mining border description of emerging patterns from dataset pairs. Journal of Knowledge and Information Systems, 8(2), 178-202.

[2] Dong, G. and Li, J. 1999. Efficient mining of emerging patterns: discovering trends and differences. In Proc. of the 5th ACM SIGKDD int. conf. on Knowledge discovery and data mining, 43-52.

[3] Fan, H. and Ramamohanarao, K. 2003. A Bayesian approach to use emerging patterns for classification. In Proc. of the 14th Australasian database conference - Volume 17 (ADC '03), 39-48.

[4] Boulesteix, AL., Tutz, G, and Strimmer K. 2003. A CART-based approach to discover emerging patterns in microarray data. Bioinformatics. 19(18),2465-2472.

[5] Warnars, H.L.H.S. 2012. Attribute Oriented Induction of High-level Emerging Patterns. In Proc. of Granular Computing Int. Conf. on Granular Computing (GrC).

[6] Warnars, H.L.H.S. 2014. Mining Frequent and Similar Patterns with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) Data Mining Technique. Int, Journal Of Emerging Tech. in Computational and Applied Sciences (IJETCAS), 11(3), 266-276.

[7] Warnars, H.L.H.S. 2016. Using Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) to Mine Frequent Patterns. Int. Journal of Electrical and Computer Engineering (IJECE), 6(6), 3037-3046.

[8] Han, J., Cai,Y. & Cercone, N. 1992. Knowledge discovery in databases: An attribute-oriented approach. In Proceedings of the 18th International Conference Very Large Data Bases, 547-559.

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

[10] Appice, A., Ceci, M., Malgieri, C., and Malerba, D. 2007. Discovering Relational Emerging Patterns. In Proceedings of the 10th Congress of the Italian Association For Artificial intelligence on AI*IA 2007: Artificial intelligence and Human-Oriented Computing (Rome, Italy, September 10 - 13, 2007). R. Basili and M. T. Pazienza, Eds. Lecture Notes In Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, 206-217.

[11] Ceci, M., Appice, A., and Malerba, D. 2007. Discovering Emerging Patterns in Spatial Databases: A Multi-relational Approach. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases* (Warsaw, Poland, September 17 - 21, 2007). J. N. Kok, J. Koronacki, R. Lopez De Mantaras, S. Matwin, D. Mladenič, and A. Skowron, Eds. Lecture Notes In Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, 390-397.

[12] Ceci, M., Appice, A., and Malerba, D. 2008. Emerging Pattern Based Classification in Relational Data Mining. In *Proceedings of the 19th international Conference on Database and Expert Systems Applications* (Turin, Italy, September 01 - 05, 2008). S. S. Bhowmick, J. Küng, and R. Wagner, Eds. Lecture Notes In Computer Science. Springer-Verlag, Berlin, Heidelberg, 283-296.

[13] Knobbe, A., Blockeel, H., Siebes, A., Van der Wallen, D.M.G., *Multi-Relational Data Mining*, In Proceedings of Benelearn '99, 1999

[14] Knobbe, A., *Multi-Relational Data Mining*, Ph.D. Dissertation, Kiminkii, the Netherlands, 2004

[15] Džeroski, S. 2003. Multi-relational data mining: an introduction. *SIGKDD Explor. Newsl.* 5, 1 (Jul. 2003), 1-16.

[16] Warnars, H.L.H.S. 2014. Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) future research. The 8th Int. Conf. on Information & Communication Technology and Systems (ICTS), Surabaya, Indonesia, pp. 13-18, 24-25 September 2014.