Causal Model of by Association Rule

Aryut Ruangtong
Department of Computer Engineering,
Faculty of Industry Education,
Rajamangala University of Technology Suvarnabhumi,
Suphanburi Campus, Thailand
r.aryut@hotmail.com

Abstract **- The main objective for this research is to develop a prototype of a casual model and a coefficient of correlation between independent variables. Also, the main technique used in this research is an association rule for a logical relationship evaluation. As well, a generating all possible association rule, a computer support and confidence, and a filter association rule by Kardi Tenomo are used to evaluate between cause and effect variables. As a result, an outcome prototype can be used for a guideline in order to select an independent variable to forecast for a dependent variable effectively.**

Keywords **- Causal Model, Association Rule**

I. INTRODUCTION

There are some problems occur in a research about prediction or about a dependent variable forecast (Y-Prediction). It explains for a relation between variables as well as an influent effect together. Moreover, it means that if the researcher inputs an indicator or an independent variable to calculate for a dependent variable, a prediction will be discrepant or not accurate. Since, if one independent variable changes, it will effect to another independent variable. In addition, any indicators are influent and have a correlation together. In consequence, it effects to a prediction of a dependent variable (Y) by these indicators (Xi). Because of some wrong assumptions such as a dependency between independent variables.

Technically, it has so many methods to prove for a relation between independent variables or indicators. An information theory also Bayes Theorem is already used to prove for a variable relation. But in this research, the researcher will present another technique as an association rule that can prove for a variable relation the same. As well, it can evaluate for a trend of an independent variable logical relationship.

Then, this research supposes to prove or find for a relation in a group of independent variables. The main issue is to specify for a trend of a relation between independent variables. Furthermore, it uses an association rule technique to find out which variable is a cause or effect. Consequently, it creates a casual model to consider for increasing a prediction accuracy of a dependent variable. After that, it selects a variable to predict for a dependent variable (Y).

II. OBJECTIVE

1) To create a casual model using an association rule technique in order to specify what variable is a cause and what variable is an effect.

2) To find for a coefficient of correlation between a cause variable and an effect variable.

III. SCOPE

1) A data using in a research is a data from a standard data set.

2) A relation size of an independent variable for a dependent choosing decision will select to connect with an independent variable according to an association rule.

3) A variable deletion will consider from a relation that has not a significance value at α 0.05 by a technique of chi-square test.

4) A relation between variables supposes to depend in linearity.

IV. LITERATURE REVIEW

A. Normalization Method

A normalization method adjusts a scope of a continuous data to be in a proper range. So, that data can be used in any calculation such as a relation value calculation. Then, it is necessary to change in a proper range by a min-max normalization method as shown in the equation below.

$$v_n = (\mathrm{v} - \mathrm{Min(V)}) \frac{nMax - nMin}{Max(\mathrm{V}) - \mathrm{Min(V)}} + nMin \quad (1)$$

v_N is a data value in a new data range.

V is a data value in an old data range.

Min(V) is the lowest data value in an old data range of variable *V*.

Max(V) is the highest data value in an old data range of variable *V*.

nMin is the lowest value in a new data range.

nMax is the highest value in a new data range.

B. Association Rules by Kardi Teknomo

This method generates all possible association rules. Independent variable (X) is a combination of items up to d-1, where d is the number of items. Dependent variable (Y) is a combination of the set difference between all items and items listed on the dependent variable. In general, the total number of possible association rule, R, is exponential to the number of items, d, which is according to the following formula.

$$R = 3^d - 2^{d+1} + 1 \quad (2)$$

R is an amount of all outcome event.
D is an amount of outcome variable.

C. Computing Support and Confidence

This method supports count denoted by $n(\mathrm{X} \cup Y)$. It supports (in percent) for each association rule that simples a ratio between support count and the number of transaction.

$$\mathrm{support}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{n(\mathrm{X} \cup \mathrm{Y})}{N} \quad (3)$$

n is an amount of case.
N is an amount of observation.

D. Correlation Significant Test Method

A correlation significant test method proves a correlation between variables that are connected or influent together. It investigates that having significance or not. It can implement by the equation as shown below.

$$\mathrm{H0}: \ \rho = 0$$
$$\mathrm{H1}: \ \rho = 1$$
$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (4)$$

ρ **r** is a correlation value between two sample variables.

n is a notification value.

t is a score t.

V. METHODOLGY

A. Data Collection

This process studies and gathers indicators from a database of an open data set that is reliable, accurate, and accepted. A data set having N Indicator in this research uses a CM-1 data from a database of PROMISE software engineering repository. This kind of data widely uses in a research about a prediction efficiency test. There are 100 notification values in a data.

B. Data Preparation

1) ***Variable Name Changing***

The name is changed in each variable because of more convenient to refer to that variable.

2) ***Data Normalization***

Because each data has different scopes also wide ranges. It is necessary to adjust a data range to be between 0 to 1 by a min-max normalization method as shown in Table I below.

TABLE I
DATA EXAMPLE BEFORE AND AFTER RANGE ADJUSTMENT

A			B			C			D		
1	-0.6359	0	1	-0.4124	0	1	-0.1919	0	1	-0.3142	0
1	-0.6383	0	1	-0.4618	0	1	-0.328	0	1	-0.3911	0
24	-0.0839	0	5	0.03157	1	1	-0.328	0	3	-0.0065	0
20	-0.1803	0	4	-0.0918	0	4	0.69271	1	2	-0.1988	0
24	-0.0839	0	6	0.15491	1	6	1.37318	1	2	-0.1988	0
24	-0.0839	0	6	0.15491	1	6	1.37318	1	2	-0.1988	0
7	-0.4937	0	1	-0.4618	0	1	-0.328	0	1	-0.3911	0
12	-0.3732	0	2	-0.3384	0	1	-0.328	0	2	-0.1988	0
25	-0.0598	0	5	0.03157	1	5	1.03295	1	5	0.37804	1
46	0.44638	1	15	1.26491	1	3	0.35248	1	1	-0.3911	0
34	0.15713	1	5	0.03157	1	5	1.03295	1	1	-0.3911	0
10	-0.4214	0	2	-0.3384	0	1	-0.328	0	1	-0.3911	0

3) ***Data Collection in Form of Association Rule Creation***

The researcher adjusts a data range in a form of an association rule creation. After that, the researcher collects those adjusted data in a table form for convenient to create an association rule as shown in Table II.

TABLE II
DATA SAMPLE FOR ASSOCIATION RULE CREATION

Transaction Records

Transaction ID	A	B	C	D
1	0	0	0	0
2	0	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	1	1	0
6	0	1	1	0
7	0	0	0	0
8	0	0	0	0
9	0	1	1	1
10	1	1	1	0
11	1	1	1	0
12	0	0	0	0
13	0	1	1	0
14	0	1	1	0
15	1	1	0	0

C. Variable Selection for Testing

In this research, the researcher selects some variables from a database. Since, it is an example to investigate for a relation between variables using an association rule technique.

Then, the researcher chooses 4 variables as Variable A, Variable B, Variable C, and Variable D.

D. Association Rule

1) ***Generating All Possible Association Rule***

From a case study in this research, the researcher randomizes a variable and selects a data from a database for a logical model creation in 100 values of notification. Moreover, there are 4 variables. Then, there is a probability for an event occurs at 50 events as shown in Table III.

TABLE III
OPPORTUNITY OF PROBABILITY FROM ALL 100 VALUES OF NOTIFICATION

No.	X	Y
1	A	B
2	A	C
3	A	D
4	A	B C
5	A	C D
6	A	B D
7	A	B C D
8	B	A
9	B	C
10	B	D

No.	X	Y
11	B	A C
12	B	C D
13	B	A D
14	B	A C D
15	C	A
16	C	B
17	C	D
18	C	A B
19	C	B D
20	C	A D

No.	X	Y
21	C	A B C
22	D	A
23	D	B
24	D	C
25	D	A B
26	D	B C
27	D	A C
28	D	A B C
29	A B	C
30	A B	D

No.	X	Y
31	A B	C D
32	A C	B
33	A C	D
34	A C	B D
35	A D	B
36	A D	C
37	A D	B C
38	B C	A
39	B C	D
40	B C	A D

No.	X	Y
41	B D	A
42	B D	C
43	B D	A C
44	C D	A
45	C D	B
46	C D	A B
47	A B C	D
48	A B D	C
49	A C D	B
50	B C D	A

2) ***Computing Support and Confidence***

For a case study in this research, the researcher specifies for a support value at 15% and specifies for a confidence value at 60% as shown in Table IV. In addition, if an event in a data set of a random database follows by the rule (True), it will use a code 1. In the opposite way, if it does not follow by the rule (False), it will use a code 0 as shown in Table V.

TABLE IV
PERCENTAGE VALUE OF SUPPORT AND CONFIDENCE

Minimum Support	15%
Minimum Confidence	60%

TABLE V
CODE VALUE SUBSTITUTION

is in Rule	
TRUE	1
FALSE	0

The researcher specifies for a support value and a confidence value. Furthermore, the researcher specifies for a code value substitution of true probability and false probability. After that, the researcher computes for an association rule. As a result, there are 9 events that have a probability to be occurred as shown in Table VI.

TABLE VI
ASSOCIATION RULE COMPUTATION RESULT

No.	X	Y	n (X U Y)	N	%Support	n(X)	Confidence	is in Rules?
1	A	B	24	100	24.00%	30	80.00%	1
8	B	A	24	100	24.00%	35	68.57%	1
16	C	B	17	100	17.00%	21	80.95%	1
22	D	A	16	100	16.00%	22	72.73%	1
23	D	B	18	100	18.00%	22	81.82%	1
25	D	AB	16	100	16.00%	22	72.73%	1
30	AB	D	16	100	16.00%	24	66.67%	1
35	AD	B	16	100	16.00%	16	100.00%	1
41	BD	A	16	100	16.00%	18	88.89%	1

The results from a calculation by an association rule can be concluded in a prototype model as shown in Fig. 1.

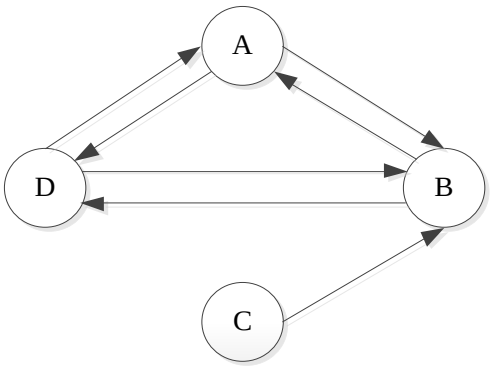


Fig. 1 Prototype Model from Calculation by Association Rule

E. Linearity Test

As a result of linearity and deviation from linearity, it has F-Test also a statistical significance value. It means that both of these variables have a linearity relation. As well, they do not have a relation in a curve line as shown in Table VII, VIII, IX, and X.

TABLE VII
LINEARITY TEST RESULT BETWEEN VARIABLE A AND B

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
A * B	Between Groups	(Combined)	165131.243	15	11008.750	175.590	.000
		Linearity	156917.864	1	156917.864	2502.853	.000
		Deviation from Linearity	8213.379	14	586.670	9.357	.000
	Within Groups		5266.430	84	62.696		
	Total		**170397.674**	**99**			

TABLE VIII
LINEARITY TEST RESULT BETWEEN VARIABLE A AND D

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
A * D	Between Groups	(Combined)	157990.269	11	14362.752	101.868	.000
		Linearity	154622.243	1	154622.243	1096.664	.000
		Deviation from Linearity	3368.027	10	586.670	2.389	.015
	Within Groups		12407.405	88	140.993		
	Total		**170397.674**	**99**			

TABLE IX
LINEARITY TEST RESULT BETWEEN VARIABLE B AND D

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
B * D	Between Groups	(Combined)	6075.784	11	552.344	112.353	.000
		Linearity	5900.286	1	5900.286	1200.181	.000
		Deviation from Linearity	175.498	10	17.550	3.570	.001
	Within Groups		432.622	88	4.916		
	Total		**6508.406**	**99**			

TABLE X
LINEARITY TEST RESULT BETWEEN VARIABLE B AND C

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
B * C	Between Groups	(Combined)	5172.527	7	738.932	50.889	.000
		Linearity	4819.599	1	4819.599	331.919	.000
		Deviation from Linearity	352.927	6	58.821	4.051	.001
	Within Groups		1335.879	92	14.520		
	Total		**6508.406**	**99**			

As a result of a linearity dependency testing from 4 pairs of variables A, B, C, and D, it refers that all pairs by an association rule are dependent in linearity according to a statistical principle.

F. Significance Calculation of Relation between X_i and X_j by Test Correlation Significance Method

This process computes for checking that each variable has a significance value altogether.

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

Also, it investigates that each independent variable has a statistical significance value or not.

In a case of this research, a degree freedom (df) has 100 values. Then, a t-test value is between -0.1660 and 0.1660 ($t_{0.05} \rangle \pm 1.660$) as shown in fig. 2.

t-distribution

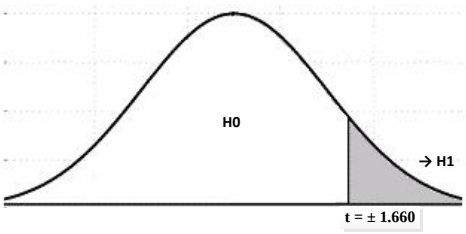
	Confidence Level						
	60%	70%	80%	85%	90%	95%	98%
	Level of Significance						
2 Tailed	0.40	0.30	0.20	0.15	0.10	0.05	0.02
1 Tailed	0.20	0.15	0.10	0.075	0.05	0.025	0.01
df							
1	1.376	1.963	3.133	4.195	6.320	12.69	31.81
2	1.060	1.385	1.883	2.278	2.912	4.271	6.816
3	0.978	1.250	1.637	1.924	2.352	3.179	4.525
4	0.941	1.190	1.533	1.778	2.132	2.776	3.744
5	0.919	1.156	1.476	1.699	2.015	2.570	3.365
80	0.846	1.043	1.292	1.453	1.664	1.990	2.374
90	0.846	1.042	1.291	1.452	1.662	1.987	2.369
100	0.845	1.042	1.290	1.451	1.660	1.984	2.365
∞	0.842	1.036	1.282	1.440	1.645	1.960	2.327

Fig. 2 Sample of T Score Value in Each α 0.05 Level from T-Test Table

Next, the researcher calculates for evaluating a significance value of a relation between Variable X_i and Variable X_j from an association rule creation. It can be shown in Table XI.

TABLE XI
RELATION AND SIGNIFICANCE BETWEEN VARIABLES FROM ASSOCIATION RULE

Relationship partner	α 0.05	T-Score	Hypothesis
A→B	0.373	4.9100	H_0
A→D	0.198	4.1870	H_0
B→A	0.001 *	0.8065	H_1
B→D	0.462	4.4250	H_0
C→B	0.871	9.8714	H_0
D→B	0.408	7.1893	H_0
D→A	0.003*	0.8387	H_1



G. Logical Model Development

This process refers from an association rule creation. Next, it selects a relation pair between variables by an association rule. After that, it tests for linearity between variables. And then, it investigates for a significance value between variables by a Pearson Correlation Technique. Finally, it can summarize in a logical model as shown in Fig. 3.

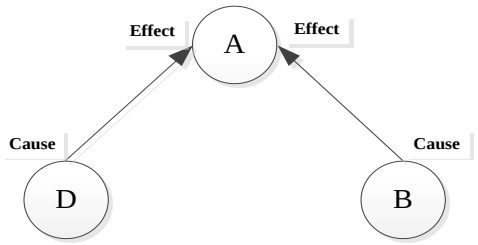


Fig. 3 Causal Model of Variables by Association Rule

VI. CONCLUSION

An association rule development can be used to create for a logical model.

1) An association rule can use to specify for a relation trend between independent variables.

2) A logical model using an association rule method can use to analyze and predict for a being occurred event.

This research can apply a relation calculation prototype development technique between these variables to use in a database or a research that has a lot of variables. It is in order to consider selecting an independent variable for a dependent variable prediction.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Vanitbancha, K. (2003). "Advanced Statistical Analysis by SPSS for Window (3rd Ed.)". Bangkok: Dhammachad Press.

[2] Teknomo, K. (2017). "Association Rules". <http://people.revoledu.com/kardi/tutorial/MarketBasket/AssociationRules.htm>. Accessed 5 October 2017.

[3] Teknomo, K. (2017). "Support and Confidence". <http://people.revoledu.com/kardi/tutorial/MarketBasket/SupportConfidence.htm>. Accessed 5 October 2017.

[4] Significance of the Correlation Coefficient. (2016). <http://janda.org/c10/Lectures/topic06/L24-significanceR.htm>. Accessed 1 August 2017.

[5] Teknomo, K. (2017). "Filtering Association Rule". <http://people.revoledu.com/kardi/tutorial/MarketBasket/Filtering.htm>.

Accessed 5 October 2017.

[6] Schisterman, E.F. (2003). "Estimation of the correlation coefficient using the Bayesian Approach and its applications for epidemiologic research". <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC155684>. Accessed 6 August 2017.

[7] Teknomo, K. (2017). "Conditionnal Probablity". <http://people.revoledu.com/kardi/tutorial/Questionnaire/Conditional%20Probability>. Accessed 5 August 2017.

[8] Massy, J.L. (1996). "Causal Model". Translat from Problemy Peredachi Informatsii, Vol. 32(1), pp. 131-136.