Causal Model of Variables by Bayes Theorem

Aryut Ruangtong[1],
Department of Computer Engineering,
Faculty of Industry Education,
Rajamangala University of Technology Suvarnabhumi,
Suphanburi Campus, Suphanburi, Thailand
[1]r.aryut@hotmail.com

Somchai Prakancharoen[2],
Faculty of Technology,
Siam Technology College, Thailand
[2]somchip@siamtechno.ac.th

and Nalinpat Porrawatpreyakorn[3]
Department of Information Technology,
Faculty of Information Technology,
King Mongkut' University of Technology North Bangkok, Thailand
[3]nalinpat.p@it.kmutnb.ac.th

Abstract - **There are three main objectives in this research. The first objective is to create a causal model and a coefficient of correlation between each independent variable using Bayes Theorem. The second objective is to find for a logical relationship also a technique of Pearson Correlation. And the third objective is to find for a coefficient of correlation between cause and effect variables. As well, the outcome model can be used for a selection of an independent variable in order to predict for a dependent variable efficiently.**

Keywords - **Causal Model, Bayes Theorem**

I. INTRODUCTION

There are some problems occurred in a research about prediction or about a dependent variable forecasting (Y-Prediction). It explains for a relation between variables as well as an influent effect altogether. Moreover, it means that if the researcher inputs indicators or independent variables 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. Because any indicators are influenced and have a correlation altogether. In consequence, it effects to a prediction of a dependent variable (Y) by these indicators (X_i). 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. Bayes Theorem is one of the statistical principles using for an investigation also a trend specification of a reasonable relation in those independent variables.

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 a statistical technique by Bayes Theorem to find out which variable is a cause or effect. Consequently, it creates a casual model to consider for accuracy increasing a prediction of a dependent variable. After that, it could support researcher to selects a suitable variables to predict for a

dependent variable (Y).

II. OBJECTIVE

1. To create a casual model using Bayes Theorem 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 largest relation size of an independent variable for a dependent choosing decision will select to connect with an independent variable only one closet relation.

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.

IV. LITERATURE REVIEW

A. Normalization Method

Normalization 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 min-max normalization as shown in the equation below [2].

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

vN 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. Probability Method

Probability uses a statistical method to predict for an occurring event. It uses an occurred data as a basic for a prediction. Then, it needs to know for an opportunity of an event how much can be occurred. It is a natural attribute of that data. Because a data in each category has diffident in attribute and opportunity to be occur. If it expresses an opportunity value in scatter, it names that value as a probability as shown in the equation below [5].

$$P(\mathrm{E}) = \sum_{x \in E} f(\mathrm{x}) \quad (2)$$

$f(\mathrm{x})$ is a Probability Density Function (PDF).

C. Bayes Theorem

Bayes Theorem is a statistical theory that describes for a probability of a being occur event (A) if another event is occurred already (B) as shown in the equations (3, 4) [4].

Bayes Theorem can write in the equation as shown below.

$$P(\mathrm{A} \mid \mathrm{B}) = \frac{P(\mathrm{B} \mid \mathrm{A}) P(\mathrm{A})}{P(\mathrm{B})} \quad (3)$$

Or can find from the equation as shown below.

$$P(\mathrm{B} \mid \mathrm{A}) = \frac{P(\mathrm{B} \mid \mathrm{A}) P(\mathrm{B})}{P(\mathrm{A})} \quad (4)$$

D. Influence between Variable Consideration

It is a consideration about a dependency trend between variables. It has a regulation to select from a dependency value. Consequently, it helps to consider for cause and effect. Furthermore, it considers only for independent variables that are influent together with each pair of other independent variables [1].

After knowing Value X, it will know a high probability value of Value Y (left). In the same time, when knowing Value Y, a probability value of Value X is low (right).

It concludes that Value X is a cause and Value Y is an effect.

$$P(\mathrm{Y} \mid \mathrm{X}) > P(\mathrm{X} \mid \mathrm{Y}) \quad (5)$$

E. Correlation Significant Test Method

Correlation Significant Test proves a correlation between variables that are connected or influent together. It investigates that having significance or not. I can implement by the equation as shown below.

$$\mathrm{H0}: \ \rho = 0$$
$$\mathrm{H1}: \ \rho = 1$$
$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (6)$$

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 498 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 10 by a min-max normalization method as shown in Table I below.

TABLE I
DATA EXAMPLE BEFORE AND AFTER RANGE ADJUSTMENT

A (loc)		B (v(g))		C (ev(g))		D (iv(g))	
Before	After	Before	After	Before	After	Before	After
1.1	1	1.4	1	1.4	1	1.4	1
1	1	1	1	1	1	1	1
24	1	5	1	1	1	3	1
20	1	4	1	4	2	2	1
24	1	6	1	6	3	2	1
24	1	6	1	6	3	2	1
7	1	1	1	1	1	1	1
12	1	2	1	1	1	2	1
25	1	5	1	5	2	5	1
46	2	15	2	3	2	1	1

C. Variable Selection for Testing

In this research, the researcher selects some variables from a database, A-B-C and D. Since, it is an example to investigate for a relation between variables using Bayes Theorem.

D. Significant Test between Sampling Variables

This process tests for finding a significant value between all 4 sample variables as shown in Table II below.

TABLE II
TEST DEPENDENCY SIGNIFICANT (P-VALUE)

Variable	A	B	C	D
A	-	0.000 *	0.000 *	0.000 *
B	0.000 *	-	0.000 *	0.003 *
C	0.000 *	0.000 *	-	0.000 *
D	0.000 *	0.003 *	0.000 *	-

After a significant test between sampling variables, all these sampling variables have variable values equal to $\alpha <= 0.05$ using a technique of Chi-square test as shown in fig. 1 below.

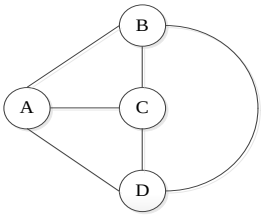

Fig. 1 Relation between Sampling Variables

E. Related Variables

In this research, the researcher will explain about an investigation of related variables only 1 pair as for instance. Because all 4 sample

variables have a significant value according to a statistical principle at all.

In this case, the researcher will give an example between Variable A and Variable B.

1) After adjust a data range for more proper to calculate as normalization, it counts on a range in a case of an observation value. Also, it specifies on an attribute as well as a quantity of that data. For example, if a notification value of Variable A is 1, a notification value of Variable B will be as shown in Table III below.

Remark: A1 means a notification value of Variable A in Range 1.

TABLE III
DATA RANGE VALUE COUNTING RESULT OF (A|B)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
B1	402	17	1	0	0	0	0	0	0	0
B2	5	37	11	4	0	0	0	0	0	0
B3	0	2	2	4	2	0	0	0	0	0
B4	0	0	1	5	1	0	0	0	0	0
B5	0	0	0	0	1	0	0	0	0	0
B6	0	0	0	0	0	0	0	0	0	0
B7	0	0	0	0	0	0	0	0	0	0
B8	0	0	0	0	0	0	0	0	1	1
B9	0	0	0	0	0	0	0	0	0	0
B10	0	0	0	0	0	0	0	0	0	1

2) The implantation is the same as (1) but swaps the matrix for finding the opposite relation between Variable B with Variable A as shown in Table IV below.

Remark: B1 means a notification value of Variable B in Range 1.

TABLE IV
DATA RANGE VALUE COUNTING RESULT OF (B|A)

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
A1	402	5	0	0	0	0	0	0	0	0
A2	17	37	2	0	0	0	0	0	0	0
A3	1	11	2	1	0	0	0	0	0	0
A4	0	4	4	5	0	0	0	0	0	0
A5	0	0	2	1	1	0	0	0	0	0
A6	0	0	0	0	0	0	0	0	0	0
A7	0	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	1	0	0
A10	0	0	0	0	0	0	0	1	0	1

F. Linearity Test

From a testing of linearity also deviation from linearity, it has f-test value as well as a statistical significant value. It expresses that both 2 variables have a linear relation. Moreover, both of these 2 variables have not a parabola relation. It is shown in Table V and Table VI.

TABLE V
LINEAR TEST RESULT BETWEEN VARIABLE A AND VARIABLE B

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
A * B	Between Groups	(Combined)	405.396	6	67.566	538.537	.000
		Linearity	400.647	1	400.647	3193.374	.000
		Deviation from Linearity	4.749	5	.950	7.570	.000
	Within Groups		61.602	491	.125		
	Total		**466.998**	**497**			

TABLE VI
LINEAR TEST RESULT BETWEEN VARIABLE B AND VARIABLE A

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
B * A	Between Groups	(Combined)	283.631	6	47.272	580.351	.000
		Linearity	277.644	1	277.644	3408.611	.000
		Deviation from Linearity	5.987	5	1.197	14.699	.000
	Within Groups		39.994	491	.081		
	Total		**323.624**	**497**			

G. Probability (P) Calculation

This process calculates to find a probability value in all independent variables.

1) A probability calculation of Variable A and Variable B in a case of (A|B) is shown in Table VII.

TABLE VII
PROBABILITY CALCULATION RESULT OF VARIABLE A AND VARIABLE B IN CASE OF (A|B)

A|B	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Prob
B1	0.807	0.034	0.002	0	0	0	0	0	0	0	0.062
B2	0.01	0.074	0.022	0.008	0	0	0	0	0	0	0.107
B3	0	0.004	0.004	0.008	0.004	0	0	0	0	0	0.034
B4	0	0	0.002	0.01	0.002	0	0	0	0	0	0.026
B5	0	0	0	0	0.002	0	0	0	0	0	0.005
B6	0	0	0	0	0	0	0	0	0	0	0
B7	0	0	0	0	0	0	0	0	0	0	0
B8	0	0	0	0	0	0	0	0	0.002	0.004	0.009
B9	0	0	0	0	0	0	0	0	0	0	0
B10	0	0	0	0	0	0	0	0	0	0.002	0.005
									(A|B) = 0.2507		

A probability calculation between Variable (A|B) can be shown in Fig. 2.

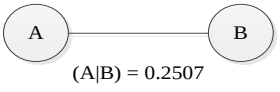

Fig. 2 Relation Diagram in Case of (A|B)

2) A probability calculation of Variable A and Variable B in a case of (B|A) is shown in Table VIII.

TABLE VIII
PROBABILITY CALCULATION RESULT OF VARIABLE A AND VARIABLE B IN CASE OF (B|A)

B|A	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	Prob
A1	0.807	0.01	0	0	0	0	0	0	0	0.817	0.071
A2	0.034	0.074	0.004	0	0	0	0	0	0	0.112	0.106
A3	0.002	0.022	0.004	0.002	0	0	0	0	0	0.03	0.045
A4	0	0.008	0.008	0.01	0	0	0	0	0	0.026	0.041
A5	0	0	0.004	0.002	0.002	0	0	0	0	0.008	0.016
A6	0	0	0	0	0	0	0	0	0	0	0
A7	0	0	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	0.002	0	0.002	0.005
A10	0	0	0	0	0	0	0	0.002	0	0.004	0.009
									(B|A) = 0.2974		

A probability calculation between Variable (B|A) can be shown in Fig. 3.

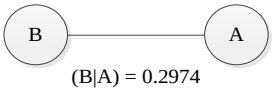

Fig. 3 Relation Diagram in Case of (B|A)

H. Reasonable Relation by Bayesian Theorem

Bayes Theorem refers to a probability theory and uses a statistical principle to analyze for a Probability (Pop). Also, it calculates and evaluates for a probability of an occurring event.

In an example of this research, a probability value between Variable A and Variable B using in this research is shown in 2 equations below.

(A|B) = 0.2507 and (B|A) = 0.2974

It means that a probability of an opportunity for an event occurring between (B|A) is higher than (A|B). In consequence, an example from this research can conclude following by Bayes

Theorem that Variable A is a cause variable and Variable B is an effect variable.

Fig. 4 Diagram of Cause and Effect Variables

I. Correlation Significance Test of Relation between Xi and Xj

This process calculates if variables having a significance value together.

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

$$tA,\delta = 14.3561$$

$$t\delta,A = 15.9063$$

J. Significance Testing

This process proves that independent variables have a significance value by a statistical principle or not.

In a case of this research, a value of degree freedom (df) is equal to 498. So, a t-test value is in a range between -0.19659 and 0.19659.

As a result, a correlation coefficient value between Variable A and Variable B has a significance altogether ($t_{0.05} \rangle \pm 1.96$).

K. Causal Model

According to a calculation process of values in a pair of Variable A and Variable B, it implements to calculate for finding values in the rest of variable pairs from the same sampling method.

TABLE IX
CONCLUSION OF CALCULATION RESULT IN EVERY VARIABLE PAIRS FROM SAMPLING METHOD (4 VARIABLES)

Variable Pair	Probability (P)	Relation Direction	Linear Test Result Between variable	ρ Value	t-test	Dependence Variable
(A|B)	0.2507	-	* (Significant)	0.5418	14.356	A is a cause.
(B|A)	0.2974	A → B	* (Significant)	0.5812	15.906	B is an effect.
(A|C)	0.3841	C → A	* (Significant)	0.6425	18.673	C is a cause.
(C|A)	0.2974	-	* (Significant)	0.5812	15.906	A is an effect.
(A|D)	0.2334	-	* (Significant)	0.5257	13.763	A is a cause.
(D|A)	0.2973	A → D	* (Significant)	0.5811	15.902	D is an effect.
(B|C)	0.384	C → B	* (Significant)	0.6424	18.668	C is a cause.
(C|B)	0.2507	-	* (Significant)	0.5418	14.356	B is an effect.
(B|D)	0.2334	-	* (Significant)	0.5257	13.763	B is a cause.
(D|B)	0.2507	B → D	* (Significant)	0.5418	14.356	D is an effect.
(C|D)	0.2334	-	* (Significant)	0.5257	13.763	C is a cause.
(D|C)	0.3841	C → D	* (Significant)	0.6425	18.673	D is a effect.

From a calculation result in Table IX, the outcome values can be created in a form of casual model as shown in Fig. 5.

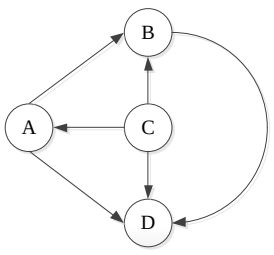

Fig. 5 Outcome Logical Model from 4 Variables by Sampling Method.

VI. CONCLUSION

1. The research can specify for a trend and a relation size between independent variables.

2. The 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. Since, it considers 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] Massy, J.L. (1996). "Causal Model". Translat from Problemy Peredachi Informatsii, Vol. 32(1), p. 131-136.

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

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

[4] Zellner, A. (2017). "Generalizing the standard product rule of probability theory and Bayes's Theorem". Journal of Econometrics, Vol. 138(2007), p. 14-23.

[5] Jiang, D. and et al. (2017). "Probability distribution pattern analysis and its application in the Acute Hypotensive Episodes prediction". Measurement, Vol. 104, p. 180-191.