Causal Model of Variables by Information Theory

Aryut Ruangtong[1],
Nalinpat Porrawatpreyakorn[2],
Department of Information Technology, Faculty of Information Technology,
King Mongkut' University of Technology North Bangkok, Thailand
[1]r.aryut@hotmail.com
[2]nalinpat.p@it.kmutnb.ac.th

and Somchai Prakancharoen[3]
Department of Computer and Information Science, Faculty of Applied Science,
King Mongkut' University of Technology North Bangkok, Thailand
[3]spk@kmutnb.ac.th

Abstract - **The objective of this research was to find out the relationship or dependency of independence random variables. Data that was used to experiment in this technique was selected from CM-1 Dataset provided by Promise. There were many independence variables that were composed to a binary logistic regression model to predict if software project should be either defect or free of bugs. All of these independence variables (x_i) were used to predict dependence variable (y – defect or not defect). Chi- square test was used to choose which independence variables were related to dependence variable. Many independence variables were not significance related or dependence to class variable Y. After that, these unrelated independence variables were tried to figure out significance their relation and either cause or effect direction with all related independent variable (with Y). Information theory was used to find out direction of either cause or effect between two independence variables (one variable that related to Y and other that was not related to Y). These relations were considered their relation importance by correlation significance test-t-Test at significance level $\propto$ 0.05.**

Keywords - **Causal Model, Information Theory, Variable**

I. INTRODUCTION

In order to do a research, there is a data gathering to process in a research. Also, it has an indicator or an independent variable in a data gathering. This variables were used in a research about a forecasting or prediction of a dependent variable. However, any indicators influence and have a correlation together. Then, this problem frequently effects with a dependent variable (Y) prediction by an indicator (X_i) because of some false assumptions.

From these false assumptions, this research attempts to increase a precision of a dependent variable forecasting. This effort specifies on how to investigate or find out for a relation between groups of independent variables. What variable is a cause? What variable is an effect? Mathematical and statistical techniques by a principle of an information theory and casual model creation were used to solve these. Since, it decreases a discrepancy as well as increases accuracy for a prediction of a dependent class variable Y.

II. OBJECTIVE

There are two main objectives in this research. The first objective is to develop an investigation technique of a relation between each variable. And the second objective is to understand a relation form of variables. As well, to know what variable is a cause and

what variable is an effect.

III. SCOPE

There are three main scopes in this research. Beginning in the first scope, a data using for a research is a data from an open database or a standard dataset. The second scope, a variable relation size for a node choosing decision will choose a node connection that is the most relative only in one level. The third scope, a variable deletion considers from a relation size that does not have a significance value α 0.05 by a t-Test technique.

IV. THEORY AND METHODS

A. Entropy

Entropy is a quantity refers to a mess of the system. So, the impurity in the system, the higher in an entropy value. In contrary, the more purity in the system, the lower in an entropy value [5].

$$H(X) = -\sum_{x_i \in X} p(x_i) \log_2 p(x_i) \quad (1)$$

X is a variable to measure for an entropy value
X_i is a data set in a variable
$p(x_i)$ is a probability value of event X_i

B. Joint Entropy

Joint entropy is a measure of the uncertainty associated with a set of variables [5].

$$H(X, Y) = -\sum_{x_i \in x} \sum_{y_i \in y} p(x_i, y_i) \log_2 p(x_i, y_i) \quad (2)$$

X, Y are variables to measure for entropy values
x_i, y_i are data sets in variable X and variable Y
$p(x_i, y_i)$ is a probability value of event x_i and event y_i

C. Conditional Entropy

Conditional entropy considers to quantifies the amount of information needed to describe the outcome of a variable Y given that the value of another random variable X is known.

$$H(Y \mid X) = \sum_{x_i \in x} p(x_i) H(Y \mid x = x_i) \quad (3)$$

X, Y are variables to measure for a conditional entropy value

x_i is a data set in variable X

$p(x_i)$ is a function of a probability distribution in variable X

Moreover, conditional entropy can be calculated from this equation (4).

$$H(Y \mid X) = H(Y, X) - H(X) \quad (4)$$

D. Mutual Information

Mutual information is a measure of the mutual dependence between the two variables [4].

$$I(X; Y) = \sum_{x_i \in x, y_i \in y} p(x_i, y_i) \log_2 \frac{p(x_i, y_i)}{p(x_i)\, p(y_i)} \quad (5)$$

X, Y are variable to measure for a mutual information value

x_i, y_i are data set in Variable *X* and variable *Y*

$p(x_i)$ is a function of a probability distribution in variable *X*

$p(y_i)$ is a function of a probability distribution in variable *Y*

$p(x_i, y_i)$ is a function of a probability distribution in variable *X* and variable *Y*

Furthermore, mutual information can be calculated from this equation (6).

$$I(X; Y) = H(Y) - H(Y \mid X) \quad (6)$$

E. Linear Correlation Coefficient

A data that join together consists of a data about a linear dependence. Also, it is not linearity between two independent variables. It can calculate from this equation below [4].

$$I = -\frac{1}{2} \log(1 - \rho^2) \quad (7)$$

ρ is a correlation coefficient value
I is a mutual information
log is a logarithm base 10

F. Correlation Significant Test

It is a testing of a correlation between related or influenced variables if they have a significant correlation value or not. It can be calculated by this equation (8) [8].

$$\begin{aligned} &\text{H0: } \rho = 0 \\ &\text{H1: } \rho = 1 \\ &t = r\sqrt{\frac{n-2}{1-r^2}} \end{aligned} \quad (8)$$

ρ **or** r is a correlation value between two random variables

n is a observation number

t is a value of t score

G. Altogether Influence Consideration

It is a dependency trend between variables. It has a selection rule from a dependency value. As well, it helps to consider for cause and effect. Since, it considers only an independent variable that influences to Y with an independent variable that not influences to Y by each pair until completes in all independent variables [3].

$$\mathrm{H(Y \mid X) > H(X \mid Y)} \quad (9)$$

Here is a formula to calculate conditional entropy.

$$\mathrm{H(Y \mid X)} = \sum_{x_i \in x} p(\mathrm{x_i})\, \mathrm{H(Y \mid x = x_i)} \quad (10)$$

However, it can be calculated by this equation below.

$$H(\mathrm{Y \mid X}) = \mathrm{H(X, Y) - H(X)} \quad (11)$$

H. Mutual Information Consideration

It is a calculation for co information of an independent variable that influences to Y separately [5].

$$I(\mathrm{X; Y}) = \sum_{\mathrm{x_i}, \mathrm{y_i}} \mathrm{p(x_i, y_i)} \log \frac{p(\mathrm{x_i}, \mathrm{y_i})}{p(\mathrm{x_i}), \mathrm{p(y_i)}} \quad (12)$$

Moreover, it can be calculated by this equation (13).

$$I(\mathrm{X; Y}) = H(\mathrm{Y}) - \mathrm{H(Y \mid X)} \quad (13)$$

I. Normalization

Normalization is an adjustment of a data scope from a continuous data to a suitable ranged data (discrete). So, it can be used for any calculation such as a data relation calculation. Then, it is necessary to change in an appropriate range by a min-max normalization method. It can calculate by the equation below (14) [6].

$$v_n = (\mathrm{v - Min(V)}) \frac{nMax - nMin}{Max(\mathrm{V}) - \mathrm{Min(V)}} + nMin \quad (14)$$

v_N is a data value in a new data range

V is a data value in an old data range

Min(V) is a least data value in an old data range of a variable V

Max(V) is a most data value in an old data range of a variable V

nMin is a least value in a new data range

nMax is a most value in a new data range

V. RESEARCH METHODOLOGY

A. Data Gathering

Indicators (N) were collected from an open data set, reliable and accurate, CM-1 a data set of PROMISE software engineering repository. This data set is used in most of researches for an efficiency testing of a data prediction having 498 observation number.

B. Data Preparation by Data Name Changing

Replacing each variable name facilitate for a variable referring.

C. Data Normalization

Because each data have different data value scopes in a different range and weights. As well, it is complicate for an entropy value calculation. Then, it should adjust data ranges in each data variable to be from 0 to 10 by a

min-max normalization method.

TABLE I
AN EXAMPLE OF DATA NORMALLIZATION BETWEEN BEFORE AND AFTER DATA RANGE

A		B	
Original data	Normalized data	Original data	Normalized data
1	1	1	1
24	1	5	1
20	1	4	1
24	1	6	1
24	1	6	1
7	1	1	1
12	1	2	1
25	1	5	1
46	2	15	2

D. Relation Calculation between Independent Variable and Dependent Variable

There are 21 independent variables that are related to dependent Variable Y. This relation is calculated by chi-square test (χ^2) in α 0.05 significant value.

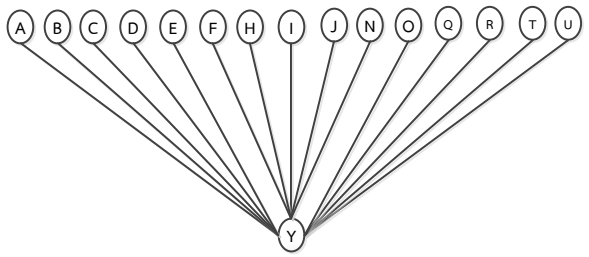

Fig. 1 All Variables that Significant Dependency to Y

E. Entropy Value Calculation

Calculate an entropy value in all independent variables with equation (15).

TABLE II
AN ENTROPY VALUE IN EACH INDEPENDENT VARIABLE

No.	Variable	Entropy (H(x))
1	A	0.9877
2	B	0.3117
3	C	1.2758
4	D	0.4844
5	E	1.0585
6	F	0.8394
7	G	2.7645
8	H	1.3971
9	I	1.7669
10	J	0.2362
11	K	2.4991
12	L	2.2377
13	M	0.751
14	N	0.5448
15	O	1.0803
16	P	1.8678
17	Q	2.2996
18	R	1.2096
19	S	2.2539
20	T	1.0743
21	U	0.8888

F. Relation Investigation between Independent Variable X_i and Independent Variable X_j

After calculate for a conditional entropy value, it investigates if independent variables has a relation together or not.

TABLE III
AN EXAMPLE CASE OF VARIABLE A AND SIX VARIABLES THAT IS NOT RELATED TO Y

A with G		A with K		A with L		A with M		A with P		A with S	
En.AH(A)	0.9877	En.AH(A)	0.9877	En.AH(A)	0.9877	En.AH(A)	0.9877	En.AH(A)	0.9877	En.AH(A)	0.9877
En.GH(G)	2.7645	En.KH(K)	2.4994	En.LH(L)	2.2377	En.MH(M)	0.751	En.PH(P)	1.8678	En.SH(S)	2.2539
En.H(A|G)	0.8322	En.H(A|K)	0.7692	En.H(A|L)	0.6736	En.H(A|M)	0.226	En.H(A|P)	0.5622	En.H(A|S)	0.6785
En.H(G|A)	0.2973	En.H(K|A)	0.2973	En.H(L|A)	0.2973	En.H(M|A)	0.2973	En.H(P|A)	0.2973	En.H(S|A)	0.2973
I : A ,G	0.1555	I : A ,K	0.2185	I : A ,L	0.3141	I : A ,M	0.7617	I : A ,P	0.4255	I : A ,S	0.3092
I : G,A	2.4477	I : K,A	2.2021	I : L ,A	1.9404	I : M ,A	0.4537	I : P ,A	1.5705	I : S ,A	1.9566
ρ : A ,G	0.4568	ρ : A ,K	0.5077	ρ : A ,L	0.5611	ρ : A ,M	0.6735	ρ : A ,P	0.6036	ρ : A ,S	0.5588
ρ : G,A	0.744	ρ : K ,A	0.7415	ρ : L ,A	0.7377	ρ : M ,A	0.6121	ρ : P ,A	0.7289	ρ : S ,A	0.738
t : I(A;G)	10.9215	t : I(A;K)	13.2466	t : I(A;L)	16.3209	t : I(A;M)	30.4879	t : I(A;P)	19.9647	t : I(A;S)	16.2912
t : I(G;A)	119.233	t : I(K;A)	100.3	t : I(L;A)	82.4488	t : I(M;A)	20.8366	t : I(P;A)	62.276	t : I(S;A)	83.471

Here is an example of a consideration that variable A relates to other variables which does not related to Y.

TABLE IV
A RELATION BETWEEN (A), H(A|λ), I, AND ρ WITH VARIABLE A BUT NOT RELATED TO Y

Other variable λ	H(λ)	H(A|λ)	H(λ|A)	I(A|λ)	$\rho_{A,\lambda}$	$t_{A,\lambda}$	I(λ|A)	$\rho_{\lambda,A}$	$t_{\lambda,A}$
Non-relation									
Class variable Y									
G	0.276	0.832	0.297	0.155	0.456	10.921	2.447	0.744	119.23
K	2.499	0.769	0.297	0.218	0.507	13.246	2.202	0.7415	100.3
L	2.237	0.673	0.297	0.314	0.561	16.32	1.94	0.737	82.448
M	0.751	0.226	0.297	0.761	0.673	30.487	0.453	0.6121	20.836
P	1.867	0.562	0.297	0.425	0.603	19.964	1.57	0.7289	62.276
S	2.253	0.678	0.297	0.309	0.558	16.291	1.956	0.738	83.471

Remark: H(A) = 0.98779

G. Relation Investigation between X_i and X_j

Investigate a relation trend between independent variables from a conditional entropy value [3].

$$H(\mathrm{A}\mid\mathrm{M}) = 0.2260$$
$$H(\mathrm{M}\mid\mathrm{A}) = 0.2973$$

Fig. 2 Example of Relation Trend

A consideration will investigate a definition of H(A|M) and H(M|A) conditional entropy values. H(A|M) means an entropy value of variable A. If value H(M|A) is high, variable A will change a lot. In consequence, value H(A|M) is a cause and variable A is an effect. Moreover, In contradict, H(A|M) is higher than H(M|A), variable A will be an effect and value M will be a cause.

H. Relation Size Calculation between X_i and X_j by Linear Correlation Method

Investigate a coefficient values that are related together with a variable. For example, coefficient correlation between A and M, δ, as shown here.

$$I(\mathrm{A};\delta) = 0.155$$
$$I(\delta;\mathrm{A}) = 2.447$$

$$\rho_{A,\delta} = 0.456$$
$$\rho_{\delta,A} = 0.744$$

I. Significance Value Calculation between X_i and X_j by t-Test Correlation Significance Method

Calculate a variable that has related significance values or not.

$$t_{A,\delta} = 10.921$$
$$t_{\delta,A} = 119.23$$

J. Significance Investigation

Investigate in an independent variable that has a significant value by statistics or not. About a data observation in this research, a degree freedom value (df) is 498. Then a t-Test critical value is between -0.19659 and 0.19659.

As a result, a correlation coefficient value is between A and δ. It means that $\rho_{A,\delta}$ and $\rho_{\delta,A}$ has the same significance in ($t_{0.05} \geq \pm 1.96$).

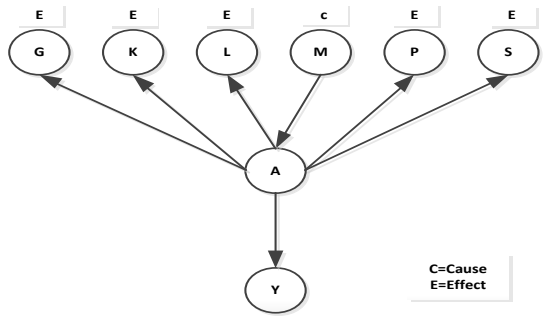

Fig. 3 Example of Relation Trend in Variable A with G, K, L, M, P, and S.

When consider from a relation trend in a variable, There is only variable M that effects to variable A. Since, M is a cause of A. Therefore M is kept as causal model of M and A. But G, K, L, P, and S are effects of A. So, they are deleted.

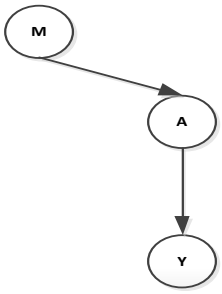

Fig. 4 Causal Model of M and A

Here is an analysis result of a cause variable and an effect variable.

After analyse for cause and effect, it investigates a linear correlation coefficient by a mutual information (I) calculation as shown in TABLE V.

TABLE V
A MUTUAL INFORMATION VALUE (i) AND CORRELATION COEFFICIENT VALUE (R) BETWEEN INFLUENCED INDEPENDENT VARIABLES

Related Pair Table				
No.	Variable	Related Pair	Value I(X;Y)	Value r
1	A	M	0.2260	0.8075
2	B	-	-	-
3	C	M	0.4537	0.6832
4	D	K	0.7543	0.8053
		P	0.1871	0.8233
5	E	-	-	-
6	F	G	0.1555	0.4403
7	H	-	-	-
8	I	S	0.3092	0.5904
9	J	G	2.4477	0.9383
10	N	K	0.2185	0.5112
11	O	S	1.9566	0.9662
		K	2.2021	0.9761
12	Q	G	2.4477	0.983
13	R	M	0.4537	0.6832
		S	1.9404	0.9654
14	T	K	0.2185	0.5112
		L	0.3141	0.5941
15	U	-	-	-

As a result, after perform whole independent variables cause and effect calculation there are relations between a pair of influenced independent variables, as shown in Fig. 5.

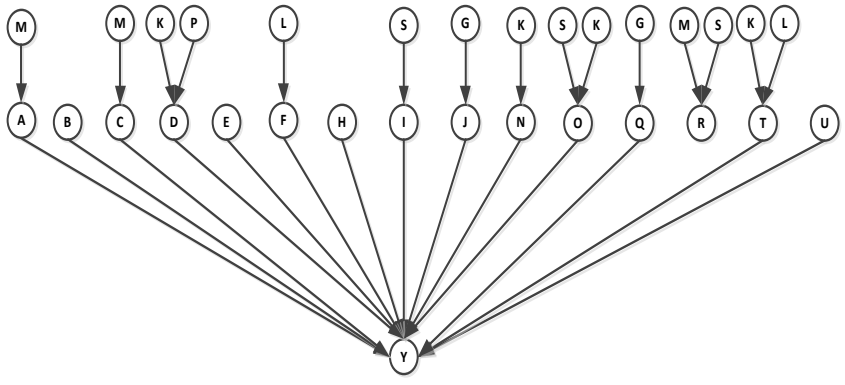

Fig. 5 Final Causal Model of all Influenced Independent Variables

VI. CONCLUSIONS

There are three conclusions in this research. First, causal model illustrate the relationship of whole independent variables. Second, increase accuracy for a prediction of a dependent class variable, under consideration of researcher. And third, can delete a variable that is not influent to a dependent variable.

REFERENCES

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

[1] Steuer, R., Kurths, J., and Selbig, J. (2002). "The mutual information detecting and evaluating dependencies between a variables". Oxford University Press.

[2] Kraskov, A., Stogbauer, H., and Grassberger, P. (2004). "Estimating Mutual Information". University of Wuppertal, Germany.

[3] Massy, J.L. (1996). "Causal Model". Translat from Problemy Peredachi Informatsii, Vol. 32, No. 1, pp. 131-136.

[4] Cover, T.M. and Thomas, J.A. (1911). "Elements of Information Theory". John Wiley & Sons, Inc., Print ISBN: 0-471-06259-6, Online ISBN: 0-471-20061-1.

[5] Keller, F. (2006). "Entropy, Joint Entropy, Conditional Entropy". University of Edinburgh, keller@inf.ed.ac.uk., Lecture: 6 March 2006.

[6] Jadsadathitikul, D. (2015). "Variable Groping by Information Theory". Department of Computer and Information Science, King Mongkt' University of Technology North Bangkok.

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