IAOS Satellite Meeting on Statistics for
the Information Society
August 30 and 31, 2001, Tokyo, Japan

A New Statistical Tool for Extracting of the Causal Condition

KONO, Yasunari *Rikkyo University, Japan*
YAMAGUCHI, Kazunori *Rikkyo University, Japan*

Abstract

Modern network society has a large amount of data with development of the computer technology. These data are not utilized completely whereas several methods have been introduced especially in marketing field. The traditional statistical tools cannot keep up with the actual situation since it is not easy to detect the useful information from gigantic databases. Quine-McClusky method has some advantages for a huge data since the characteristics of the algorithm considering a higher order interaction make possible to apply for them. The method is suitable for a prediction as such kind of analysis to classify patterns of causal conditions with minimized equations. A statistical tool using the method enables to find groups whose response probabilities in dichotomous dependent variable are larger than a target value. This tool is quite effective to extract some interesting pattern from numerous piled data.

1 Introduction

The rapid progress of computer technology enables to gather a large amount of data easily. However it is difficult to construct the method to detect useful information from huge databases although various methods have ever been introduced for such data. Data mining is the search for valuable information in large volumes of data, which is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization (Cavena et al., 1997). Data mining is mainly used for 'confirmatory' and 'explanatory' analysis (Iwasaki, 1999). This paper attaches greater importance to former analysis including dependent variables for prediction.

Confirmatory techniques construct the model explaining a specific item of the data by plural other items, which is divided into two cases using the discrete and the continue dependent variables (Shiota, 1997). Data mining by classification is mainly analyzed by decision tree such as C4.5 (Quinlan, 1998) or CART (Breiman et al., 1984). Explanatory techniques (clustering, association rules), on the other hand, aim at the simple representation of the contents or the structure of the data to understand them when the data items do not have the difference between dependent and dependent variables.

Confirmatory causal analyses of behavioral patterns have some difficulties. First of all, they have complexities of the asymmetric information of the data. The causal analysis of ‘the automobile accidents’ or ‘the enterprise bankruptcy’ is given as this example. The data of the target group for the purpose of the research are very few, namely the data of non-target are in a large majority in these cases. Secondly, some methods such as logistic regression do not consider a higher order interaction. It is also hard to be screening huge independent variables. Lastly, the problem is about the multicollinearity. The estimate of the each coefficient becomes inaccurate when the data have much explanatory variables in which high correlations exist.

Kono (2001) proposed a modeling for detection of subgroup in a large data according to the response probabilities. The method finds target groups whose response probabilities in dichotomous dependent variable are larger than a target value. This algorithm stands for a selection of independent variables considering a higher order interaction. The method is quite effective to extract some interesting pattern from numerous piled data.

Purposes of this research are the classification of patterns of causal conditions using Quine-McClusky algorithm and the detection of the statistically significant groups in classified patterns.

2 Method

Dichotomous multi-valiate data matrix is assumed in this method. The goal of this analysis is the exploration of minimized equations on the condition $D(X_1, X_2, \cdots, X_p)$ that satisfied probability of positive response $Pr(Y=1|D) \geqq T$, which T shows a target value. The method is completed after repeated trial and error by using smaller value as cutoff than a specified response probability. The method finds the groups whose response probabilities in dichotomous dependent variable are larger than a target value.

The concrete algorithm of the new method consists of the following five steps.

Step 0: Set the target value.
Step 1: Set the percent cutoff value.
Step 2: Get the minimized equation.
Step 3: Set the confidence intervals of response probabilities for each group.
Step 4: Check the results (Back to Step 1 if the result leaves still room for improvement).

In Step 0, the percentage of the target group is set for the purpose of the research. It is possible to change the target value as the need arises. In Step 1, analysts make a cross table from the data. The first percent cutoff value is generally set to the same percentage as the target value. In Step 2, the target groups are nominated according to minimized equations detected by Quine-McClusky algorithm. In Step 3, analysts decide the level of confidence intervals, and find the lower limit of the confidence intervals for the nominated groups. In Step 4, it is checked whether the target vale is less than all of lower limit values of confidence intervals or not. Even if it is less than all of the values, it is desirable to repeat the above steps with smaller cutoff values.

3 Example

The first thing, the method needs making a cross table by reconstructing a data matrix. This hypothetical cross table has three causal conditions.

Table 1. Example Data With Three Independent Variables

Condition			# (Y=1)	Cases
X_1	X_2	X_3		
0	0	0	195	1965
0	0	1	120	150
0	1	0	225	872
0	1	1	69	101
1	0	0	0	34
1	0	1	43	65
1	1	0	387	501
1	1	1	4	5
			1043	3693

Minimized equations of the 60, 70 and 80 percent of the response probability are represented as follows:

(1) 60%: $Y = X_1 X_2 + X_3$,

(2) 70%: $Y = \overline{X}_1 \overline{X}_2 X_3 + X_1 X_2$,

(3) 80%: $Y = \overline{X}_1 \overline{X}_2 X_3 + X_1 X_2 X_3$,

where X represents positive response and $\overline{X}$ shows negative.

The patterns of the groups of dependent variables are found according to a response probability of dependent variables. Equation (1) shows the selecting of individuals who respond positively to X_3 or to both X_1 and X_2 when it is set to 60 percent. Quine-McClusky algorithm generally selects larger groups of positive respondent probability than the setting percentages. It represents 77 (74) percent in case of setting 60 percent that the individuals who respond positively to $X_1\ X_2$ (X_3) number 391 of 506 (236 of 321).

The following example shows four steps to find groups with the response probability more than 70 percent by the proposed method.

Step 0: Set the target value to 70 percent.
Step 1: Set the percent cutoff to 70 percent.
Step 2: Get the minimized equation and nominate the groups.
Step 3: Set the confidence intervals for response probabilities of the above two groups.
Step 4: Check the results

The minimized equation for 70 percent cutoff is $\overline{X}_1\overline{X}_2X_3 + X_1\ X_2$ as shown Equation (2). This equation nominates two groups, $\overline{X}_1\overline{X}_2X_3$ and $X_1\ X_2$, for target groups. In this example, the 80 percent confidence intervals for response probabilities are used. The result is as follows:

	Estimated response probability	Lower limit
$\overline{X}_1\overline{X}_2X_3$	80.0%	75.8%
$X_1\ X_2$	77.3%	74.9%

The both lower limit of the 80 percent confidence intervals for $\overline{X}_1\overline{X}_2X_3$ and $X_1\ X_2$ are larger than 70 percent. The results accept these two groups of Equation (2). Repeat the above steps with the 60 percent cutoff value to check whether larger groups than these two groups that meet the requirements exist or not.

Step 1': Set the percent cutoff to 60 percent.
Step 2': Get the minimized equation.
Step 3': Set the confidence intervals for response probabilities of the above two groups.
Step 4': Check the results

The minimized equation for 60 percent cutoff is $X_1\ X_2 + X_3$ as shown Equation (1). This equation nominates two groups, $X_1\ X_2$ and X_3, for target groups. In this example, the 80 percent confidence intervals for response probabilities are used. The result is as follows:

	Estimated response probability	Lower limit
$X_1\ X_2$	77.3%	74.9%
X_3	73.5%	70.4%

The both lower limit of the 80 percent confidence intervals for $X_1\ X_2$ and X_3 are larger than 70 percent. The results accept these two groups of Equation (1). These two groups of Equation (1), $X_1\ X_2$ and X_3, are also accepted the target value of the 70 percent. The iteration is stopped at this point because the output does not change until 26 percent cutoff.

This method finds $X_1\ X_2$ and X_3 as the groups that meet the requirements as stated above. The results show two target groups ($X_1\ X_2$ and X_3) with the response probability more than 70 percent in this data,

3 Discussion

This paper proposes a method to find groups whose response probabilities of dichotomous dependent variable are larger than a target value, namely the target groups are detected according to the response probability. Analysts enable to change the percentage of a target value to get the target group according to the purpose of the research. This statistical tool can be applied to fields that require the data mining technique like marketing because it is effective not only for selecting the target groups that will purchase the goods with high probability from the large database of customers, but for flexible setting for target groups according to the response probability.

The method uses the interval estimation to consider sampling errors, and constructs the 80 percent confidence intervals in Example. If the percent cutoff value is set to 80 percent, the result of the 80 percent confidence intervals is as follows:

	Estimated response probability	Lower limit
$\overline{X}_1 \overline{X}_2 X_3$	80.0%	75.8%
$X_1 X_2 X_3$	80.0%	57.1%

The lower limit of the 80 percent confidence intervals for the group of $X_1 X_2 X_3$ is smaller than 70 percent. The results cannot accept the group $X_1 X_2 X_3$ of Equation (3) although the group $X_1 X_2$ in Equation (1) and (2) implicates $X_1 X_2 X_3$ is accept. The method is quit effective for sampling errors, but how to decide the level of confidence needs further study.

This method, however, leaves room for improvement as follows:

1. The method has problems of multi-nominal and continuous variables. It uses dichotomous variables basically, but cannot use multi-nominal and continuous variables although it can deal with them by using dummy variables. The software that conducts this method should be improved to handle them directly.
2. The method must be considered about automatic stopping rule from Step4 to Step1' that is conducted by analysts on the present software.
3. How should be set the cutoff value for the target value? This is the question to find optimal way.
4. Screening logic has to be designed for the method when the data have a huge the independent variables.

These problems will be solved with comparing the other methods such as tree analysis or logistic regression.

References

Breiman, L., Friedman, J., Olshen, R. S. & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group. Belmont, CA.

Cavena, P., Hadjinian P., Stadler, R., Verhees, J. & Zanasi, A. (1997). *Discovering Data Mining, From Concept to Implementation*. Prentice-Hall, Inc. NJ.

Iwasaki, M. (1999). Data Mining and Knowledge Discovery, From a Statistical Viewpoint. *The Japanese Journal of Behaviormetrics*, 26(1), 46-58.

Kono, Y. (2001). Segmentation Procedure using Quine-McClusky Algorithm, *The Journal of Applied Sociology*, 43, 95-101.

Quinlan, J. (1998). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc. San Francisco, CA.

Shiota, C. (1997). Data mining - Techniques and applications. *Bulletin of The Computational Statistics of Japan*, 10(2), 127-144.