Boolean and Cluster Analysis for Knowledge Discovery in Databases

Yasunari Kono¹, Naoko Sakurai², Kazunori Yamaguchi³

¹⁾Rikkyo University, Graduate School of Social Relations (kono@ess.rikkyo.ne.jp)
 ²⁾Waseda University, Institute of Asia-Pacific Studies (sakurai@wiaps.waseda.ac.jp)
 ³⁾Rikkyo University, Department of Industrial Relations (kyamagu@rikkyo.ac.jp)

Abstract

Various methods have ever been introduced for KDD (Knowledge Discovery in Databases) with the rapid development of computer technology. It is not easy to construct the method that enables to detect the useful information from gigantic databases. Data mining, the stage of finding valuable information in KDD, has lately attracted considerable attention. The aim of this paper is to propose a new method for KDD realized by the Boolean approach. The Boolean approach is designed by Ragin to integrate a caseoriented comparative method and a variable-oriented approach for essentially supporting qualitative analysis. This approach has not come to be widely used for the last decade whereas it has some advantages; for example, the characteristics of the algorithm considering a higher order interaction make possible to apply the approach to KDD. The technique of data mining is divided into confirmatory and explanatory analysis, and this paper lays stress on former analysis. The Boolean approach is suitable for a prediction as such kind of analysis to classify patterns of causal conditions with minimized equations. This approach, on the other hand, has some demerits to treat large amount of data. First of all, it has the problem of contradictory terms. Three approaches for contradictory terms suggested by Ragin are insufficient to resolve the problem. Secondly, this approach does not consider sampling errors. The reason of it is that both patterns of few and many frequencies have equal weight, and a frequency of a certain pattern will not affect results because the approach adopts the method of sampling patterns. These problems are solved by improving the algorithm of the Boolean approach. The proposed new method considers sampling errors by using the interval estimation. The method enables to find clusters or groups such that 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 new method for KDD is quite effective to extract some interesting pattern from numerous piled data.

1. Introduction

Development of computer technology enables to extract a large amount of data easily. Such data, however, must be processed to use them for the purposes of researches. Data mining is the search for valuable information in large volumes of data [17], which is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization[2].

Data mining is mainly used for 'confirmatory' and 'explanatory' analysis [5]. This paper attaches greater importance to former analysis including dependent variables. Several methods of data mining are introduced such as classification, regression, clustering, dependency modeling, change and deviation detection, decision tree, discovery of association rule and neural network [3]. Weiss and Indurkhya [17] have divided some generic types of data-mining problems into two general categories: (a) prediction (classification, regression, time series) and (b) knowledge discovery (deviation detection, database segmentation, clustering, association rules, summarization, visualization, text mining).

Confirmatory techniques construct the model explaining a specific item of the data (dependent variable) by plural other items (independent variables), which is divided into two cases using the discrete and the continue dependent variables [13]. Data mining by classification is mainly analyzed by decision tree such as C4.5 [11] or CART (Classification And Regression

Tree) [1]. 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 [13].

The Boolean approach is suitable for a prediction, especially for classification of patterns. Boolean algorithms are based on the work of electrical engineers who developed them in the 1950s to simplify switching circuits [12]. QCA (Qualitative Comparative Analysis), a software carrying out the Boolean approach, adopts the minimization of the logic function in the fields of the switching theory and the logic circuits [6]. This algorithm stands for a selection of independent variables considering a higher order interaction from a statistical point of view.

The key point of data mining is how to detect necessary information from vast data, which is symbolized by the name of Knowledge Discovery in Databases. First of all, data transformation and data cleaning is carried out at the stage of preprocessing. The Boolean approach needs the data transformation from continuous to dichotomous variables, but this approach does not consider data cleaning such as treating of missing value or outlier. Secondly, reduction and summarization of data should be conducted not to loss the information of the whole data set from a standpoint of saving time for the analysis [5]. The Boolean approach enables to summarize data by minimized equations without random sampling even if the numbers of the data is very large, but it is difficult to interpret the output when the numbers of independent variables becomes more than four.

This paper proposes a new method to find clusters or groups such that response probabilities of dichotomous dependent variable are larger than a target value. The method is constructed by improving the algorithm of the Boolean approach because it has some problems such as contradictory terms. The Boolean approach does not consider sampling errors. The reason of it is that both patterns of few and many frequencies have equal weight, and a frequency of a certain pattern will not influence results because the approach adopts the method of sampling patterns [4]. The proposed method considers sampling errors by using the interval estimation.

This paper is organized as follows. In the next chapter, the Boolean approach is introduced. Chapter 3 defines the new method and gives a numerical example. This paper concludes in Chapter 4 with proposals for future work.

2. Boolean approach

The Boolean approach was designed by Ragin [12] to integrate a case-oriented comparative method and a variableoriented approach since they are surprisingly complementary. The Boolean approach, however, is basically a method to support qualitative analysis like a case study although Ragin intended to integrate a qualitative and a quantitative method. The present Boolean approach should not handle a large number of quantitative variables because the other statistical methods are superior in this respect, but this approach will be extended its sphere of application when some problems are solved.

2.1 Characteristics of Boolean Approach

In this section, we introduce Ragin's basic idea for the Boolean approach at the beginning and then pick up the several comments to the approach.

(1) Ideal Features of a Synthetic Comparative Research Strategy

Ragin [12] stated five ideal features of a synthetic comparative research strategy as follow:

- An ability to examine a large number of cases
- An ability to address complex causal conjunctures
- An ability to produce parsimonious explanations (if desired)
- An ability to investigate cases both as wholes and as parts
- An ability to evaluate competing explanations

Though he explained that the Boolean approach meets these requirements, his explanation has some questions. First of all, the Boolean approach can examine a large number of cases, but the approach is more effective on a small number of cases rather than a large one comparing with the other statistical methods. Secondly, it is equivalent to the first problem to address complex causal conjunctures; statistical methods can explain complex phenomena in many cases more accurately than the Boolean approach.

Third feature, parsimonious explanations, is not doubtful for the approach because it is possible to explain complex causal conjunctures by minimized algebraic equations. Ragin mentioned about last feature that the typical end product of the Boolean analysis is a statement of the limits of the causal variables identified with different theories, not their mechanical rejection or acceptance.

(2) Problems of Boolean Approach

Kosaka [9] criticized the Boolean approach, and he indicated the problems about the approach as follows:

- Problems of gaps between logical possibility of causal combinations and actual causal combinations
- Formalization of event S about probabilities of combinations if it is unknown whether the event results or not because a certain causal combination has not realized empirically.
- Problems to handle contradictory cases
- Evaluation for existing theories

Ragin [12] discussed the first problem as limited diversity, actual causal combinations are described logical possibility of causal combinations as subset of maximum diversity and The Boolean techniques of qualitative comparative analysis provide a direct approach to the question.

The second problem links to the first one, Kosaka [9] deal with unknown cases by allotting 1 or 0 uniformly, or 1 and 0 to each case. He suggested to search for additional factors and to replace conditions in contradictory cases. This concept is similar to a new method, which is proposed in the present paper.

Last comment means examining discontinuity between current theories and empirical realizations, in other words, a comparative method based on Boolean algebra makes possible to verify existing theories, and to indicate what causal combination lack in current theories according to each empirical data [9].

(3) Thought experiment by Boolean analysis

Kanomata [7] mentioned an epochal significance of the Boolean approach comparing with typical qualitative comparative methods in social science fields as follows:

- A capability to analyze various causation including complex interactions
- An each capability of a comparison, a minimization and an introduction of objective assumption
- A capability to express results as equations

The first feature is equivalent to Ragin's second comment; also the third one is included in Ragin's third comment. The second characteristic shows same advantages as a statistical method.

Kanomata's methods corresponds to thought experiment and conceptual simulation judging from these points:

- Conceptual variables
- Analysts decide dependent variables.
- Minimized algorithm of Boolean approach

Boolean approach enables to be used on conceptual level in this way ([7], [10], [14], [15]).

(4) Merits and Demerits of the Boolean approach

Tarohmaru and Tanaka [16] pointed features of the Boolean analysis as follows:

Merits

- It is possible to analyze complex interactive effects among variables in detail.
- It is possible to comparatively analyze at once many cases contrasting with general case studies.
- It is possible to use few cases for quantitative analysis such as international comparative studies. Demerits
- It is impossible to test.
- Independent and dependent variables should be binary.

The second and the third merit show suitable numbers of cases for The Boolean approach. On the other hand, they commented that the first demerit is possible to be covered by logit analysis.

(5) Differences between the Boolean approach and Inferential Statistics

Inaba [4] stated that the differences between the Boolean approach and inferential statistics are as follows:

- The Boolean approach minimizes sample data as analytical objects and does not estimate parameter.
- Both patterns of few and many frequencies have equal weight, and a frequency of a certain pattern will not influence results because the approach adopts the method of a sampling patterns.
- Effects of interaction effect become examined objects from the beginning.

He added that the data of the Boolean approach must have such characteristics that the sample data are equal to the population; analysts should take interest in sampling some patterns including a few patterns; a causal condition needs to be set as a research subject.

He also indicated a limitation of the Boolean analysis: the more the dimensional number increases, the harder the interpretation for interaction, which is equal to the general analysis of variance. It limits two or three factors and two seems realistic.

2.2 Problems of Contradictory Terms

The Boolean approach has not come to be widely used for the last decade. One of the reasons is that the approach leaves much room for improvement, especially about the problems of contradictory terms ([4], [8], [9]). Ragin [12] indicated three approaches to deal with the problems: the case-oriented approach, the interpretative approach and the method applied statistical standards, but these approaches cannot solve the problem sufficiently.

Tarohmaru and Tanaka [16] suggested the Boolean analysis with multi-value variables as independent variables because the Boolean approach has a weak point as it can treat only dichotomous, 1 and 0. The basic questions, however, about contradictory terms are equivalent to the binary even if the analysis can treat multi-value variables. These variables must be change to dummy variables whereas the Boolean approach cannot handle continuous variables directly. The problem becomes substantial when continuous variables need to be divided since it causes contradictory terms.

In addition to this question, the data including dichotomous variables like gender also causes contradictory terms because a pattern on equal condition does not always lead same results. In this case, researchers have to divide the outputs of contradictory rows in a truth table into 1 or 0, and it becomes possible to change the outputs by a percent cutoff, but this method causes some demerits.

2.3 Regular Operations by Boolean Approach for contradictory terms

Table 1 has three independent variables (*X*, *Y*, *Z*), and "Occurrence" indicates the number of occurrences among cases in each row, and this table has four contradictory rows (i = 1, 3, 5, 7) in the output (*F*).

	Condition		Output	Occurrence	Cases	
i	X	Y	Ζ	F		
1	0	0	0	0(?)	1	10
2	1	0	0	0	0	2
3	0	1	0	1(?)	4	5
4	0	0	1	1	1	1
5	1	1	0	?	2	5
6	1	0	1	0	0	1
7	0	1	1	?	4	7
8	1	1	1	0	0	3
				Total	12	34

Table 1 Hypothetical Truth Table

The Boolean shows the output values of rows including contradictory terms as "C" and represents a percentage of output variables of 1 (or 0) in row's data. It is possible to deal with them by a percent cutoff [6]. Table 2 shows the frequencies in Table 1 analyzed by QCA.

Table 2 Frequencies in Table 1

i	X	Y	Ζ		Out	#0	%	# 1	%
1	0	0	0	,	С	9	90	1	10
2	1	0	0	,	0	2	100	0	0
3	0	1	0	,	С	1	20	4	80
4	0	0	1	,	1	0	0	1	100
5	1	1	0	,	С	3	60	2	40
6	1	0	1	,	0	1	100	0	0
7	0	1	1	,	С	3	43	4	57
8	1	1	1	,	0	3	100	0	0
6 7 8	1 0 1	0 1 1	1 1 1	, , ,	0 C 0	1 3 3	100 43 100	0 4 0	

QCA re-codes outputs of contradictory rows if an analyst input a cutoff value as follows:

Percent Cutoff: 75

Then, the output in the first row is re-coded to "0" since the percentage of "0" is over 75 percent, in the same way, the output in the third row is re-coded to "1," but the outputs in the fifth and the seventh row show still "C" because the

percentage of "0" in fifth row and of "1" in seventh row are less than 75% (see Table 3).

i	X	Y	Ζ		Out	#0	%	# 1	%
1	0	0	0	,	0	9	90	1	10
2	1	0	0	,	0	2	100	0	0
3	0	1	0	,	1	1	20	4	80
4	0	0	1	,	1	0	0	1	100
5	1	1	0	,	С	3	60	2	40
6	1	0	1	,	0	1	100	0	0
7	0	1	1	,	С	3	43	4	57
8	1	1	1	,	0	3	100	0	0

Table 3 Revised Frequencies

Analysts can change contradictory terms to 0 or 1 by a percent cutoff in this way, which percentage is decided by their own judgement.

Inaba [4] pointed out that the Boolean approach by a percent cutoff leads analysts to falling into difficulties in interpreting the results, and it causes conflicts with principles of the approach; both patterns of one and many frequencies have equal weight. Thus the Boolean approach is not suitable for the data assumed contradictory terms on the circumstances such as diversity is thought from the beginning.

Contradictions must be coped with in some ways because most of general data have contradictions actually. A percent cutoff is liable to be subjective although it seems an effective method, therefore analysts must judge carefully. Contradictory terms should be handled objectively in essence, but it is not mistake that final judgements are left to experts of case studies.

2.4 Ragin's Approaches for Contradictory Terms

Ragin [12] suggested three approaches to deal with the problems, the case-oriented approach, the interpretative approach and the method applied statistical standards.

First of all, the case-oriented approach is to follow the lead of case-oriented researchers. Secondly, the interpretative approach codes "0" (or "-" or "1") uniformly to all ambiguous causal combinations. The last method is divided into two ways: one is the way that judges from the results of testing statistical hypothesis; the other is the way that calculates expected values of each combinations of independent values using the additive logistic model. Ragin [12] mentioned that all solutions to the problems of contradictions, except the first, violate the spirit of case-oriented qualitative research and should be used only when it is impossible to return to the original cases and construct a better truth table. On the contrary, case-oriented approach lacks objectivity.

The interpretative approach among Ragin's three approaches is quoted in this paper to lead the new method. Table 1 revises to Table 4 after the percent cutoff (75%) procedure.

	Condition		Output	Occurrence	Cases	
i	X	Y	Ζ	F		
1	0	0	0	0	1	10
2	1	0	0	0	0	2
3	0	1	0	1	4	5
4	0	0	1	1	1	1
5	1	1	0	С	2	5
6	1	0	1	0	0	1
7	0	1	1	С	4	7
8	1	1	1	0	0	3
				Total	12	34

Table 4 Revised Hypothetical Truth Table

The interpretative approach is classified into three types by coding "0" (or "-" or "1") uniformly each to all ambiguous causal combinations.

(1) Approach gives "0" uniformly

The first approach changes "C" to "O" in Table 4 uniformly as if the lines had not contradictory terms. This approach is based on that if no clear tendency is apparent among the cases conforming to certain causal combination, then the output should be coded conservatively [12], which represents what combination of causal condition have a connection with the results definitely. Minimized algebraic equation in this way shows:

$$F = xYz + xyZ \tag{1}$$

(2) Approach gives "-" uniformly

The second approach treats combinations including contradictory terms as though the combinations do not exist; the effect of this procedure is allowing the algorithm to determine which final output value the contradictory rows should receive. If they help to produce a more minimal solution, they receive a coding of 1; if they do not, they receive a coding of 0 [12]. Minimized algebraic equation in this manner shows:

$$F = Yz + xY + xZ$$
 (Simplifying assumptions for $F: XYz + xYZ$) (2)

(3) Approach gives "1" uniformly

The last approach gives "1" about uncertain causal combinations. Then it is possible to include all combinations that are supposed to cause the results. This approach is based on that wide net should be cast [12]. Minimized algebraic equation in this procedure is as follows:

$$F = Y_Z + xZ \tag{3}$$

3. Method

3.1 Modeling for a Response Probability of Dependent Variables by Boolean Approach

Ragin's approaches are insufficient to resolve the contradictory problems since his idea of this approach assume to compare case studies as qualitative research.

The Boolean approach has some weak points besides the contradictory problems. It is that the result reflects the accidental outputs because both patterns of few and many frequencies have equal weight. The new method resolves the problem as follows:

- The method is completed after repeated trial and error by using smaller value as cutoff than a specified response probability.
- The method finds the clusters or the groups such that 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 and nominate the groups.
- Step 3: Make 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 truth 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 Boolean algorithm.

In Step 3, analysts decides 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.

3.2 Example

Table 5 has three conditions of the dichotomous variables, and this table uses 'Positive response' instead of 'Occurrence' in Table 1 and Table 4.

		Condition		Positive	Cases
	Х	Y	Ζ	response	
а	0	0	0	195	1965
b	0	0	1	120	150
С	0	1	0	225	872
d	0	1	1	69	101
е	1	0	0	0	34
f	1	0	1	43	65
g	1	1	0	387	501
h	1	1	1	4	5

Table 5 E	Example Data	With Three	Independent	Variables
-----------	--------------	------------	-------------	-----------

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

$$60\%: R = XY + Z \tag{4}$$

$$70\%: R = xyZ + XY \tag{5}$$

$$80\%: R = xyZ + XYZ \tag{6}$$

The patterns of the groups of dependent variables are found according to a response probability of dependent variables. When it is set to 60 percent, Equation 8 shows the selecting of individuals who respond positively to Z or to both X and Y. The Boolean approach always selects larger groups of positive respondent probability than the setting percentages. It represents 74 (77) percent in case of setting 60 percent that the individuals who respond positively to Z (XY) number 236 of 321 (391 of 506).

In addition to the question, the Boolean approach has the problem of reliability. It treat the 80 percent results of b (120/150) and h (4/5) row alike since they have equal weight.

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.

Table 6 shows the truth table when a cutoff value is 70 percent.

	X	Y	Ζ		Out	# 0	%	# 1	%
а	0	0	0	,	1	1770	90	195	10
b	0	0	1	,	1	30	20	120	80
с	0	1	0	,	0	647	74	225	26
d	0	1	1	,	0	32	32	69	68
е	1	0	0	,	0	34	100	0	0
f	0	0	1	,	0	22	34	43	66
g	0	1	0	,	1	114	29	387	71
h	1	1	1	,	1	1	20	4	80

 Table 6
 Truth Table with 70% Cutoff

Step 2: Get the minimized equation and nominate the groups.

The minimized equation for Table 6 is xyZ + XY as shown Equation (5). This equation nominates two groups, xyZ and XY, for target groups.

Step 3: Make confidence intervals for response probabilities of the above two groups.

In this example, the 80 percent confidence intervals for response probabilities are used. The result is as follows:

	Estimated response probability	Lower limit
xyZ	80.0%	75.8%
XY	77.3%	74.9%

The both lower limit of the 80 percent confidence intervals for two groups, xyZ and XY, are larger than 70 percent. The results accept these two groups of Equation (5).

Step 4: Check the results

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.

Table 7 shows the truth table when a cutoff value is 60 percent.

	X	Y	Ζ		Out	# 0	%	# 1	%
а	0	0	0	,	1	1770	90	195	10
b	0	0	1	,	1	30	20	120	80
с	0	1	0	,	0	647	74	225	26
d	0	1	1	,	1	32	32	69	68
е	1	0	0	,	0	34	100	0	0
f	0	0	1	,	1	22	34	43	66
g	0	1	0	,	1	114	29	387	71
h	1	1	1	,	1	1	20	4	80

Table 7 Truth Table with 60% Cutoff

Step 2': Get the minimized equation and nominate the groups.

The minimized equation for Table 7 is Z + XY as shown Equation (4). This equation nominates two groups, Z and XY, for target groups.

Step 3': Make confidence intervals for response probabilities of the above two groups.

In this example, the 80 percent confidence intervals for response probabilities are used. The result is as follows:

	Estimated response probability	Lower limit
Ζ	73.5%	70.4%
XY	77.3%	74.9%

The both lower limit of the 80 percent confidence intervals for two groups, xyZ and XY, are larger than 70 percent. The results accept these two groups of Equation (4).

Step 4': Check the results

These two groups of Equation (4), XY and Z, also accept the target value of the 70 percent. The iteration is stopped at this point because the output does not change until 26 percent cutoff.

As the result mentioned above, this method can find XY and Z as the groups that meet the requirements.

4. Discussion

This paper proposes a new method to find groups such that response probabilities of dichotomous dependent variable are larger than a target value. The method is constructed by improving the algorithm of the Boolean approach because it has some problems such as contradictory terms. The Boolean approach does not consider sampling errors. The reason of it is that both patterns of few and many frequencies have equal weight, and a frequency of a certain pattern will not influence results because the approach adopts the method of sampling patterns [4]. The proposed new method uses the interval estimation to consider sampling errors, and constructs the 80 percent confidence intervals in Chapter 3.

If the percent cutoff value is set to 80 percent for the example in Chapter 3, the result of the 80 percent confidence intervals is as follows:

	Estimated response probability	Lower limit
xyZ	80.0%	75.8%
XYZ	80.0%	57.1%

The lower limit of the 80 percent confidence intervals for the group of *XYZ* is smaller than 70 percent. The results cannot accept these two groups of Equation (6).

The new method is quit effective for sampling errors, but how to decide the level of confidence needs further study.

This paper shows the way of the solution for the problems of the Boolean approach such as contradictions by setting the response probability of dependent variables, but they are not resolved completely. The other problems of the Boolean approach remain besides the demerits picked up in this paper.

First of all, the Boolean approach does not take into account missing value because it is fundamentally for qualitative comparative case study. We cannot avoid the problems of missing values for the large amount of data set. We have to extend our proposed method to be applicable to the incomplete data.

Secondly, the Boolean approach perplexes analysts to interpret the results graphically when the numbers of independent variables are four or more. Venn diagram may illustrate just only for only two or three variables. Although Boolean approach has characteristics of representing by minimized equations, the results need to be expressed by some diagram.

Thirdly, the QCA program is hard to handle compared with other statistical packages, and cannot deal with multivalue variable directly. Yosano [18] pointed out that it is reasonable for analysts tend to apply a statistical method such as SPSS including a probability model rather than to analyze by Boolean approach using QCA, which is a software only for a single purpose. Each step of the proposed method is operated by hand using QCA in this paper. The software processing from Step 1 to Step 4 automatically is under development.

Acknowledgements

We are especially grateful Professor Akinori Okada for his advice on the paper and also thank Professor Michiko Watanabe and Mr. Tatsuki Inoue for their useful comments.

References

- [1] Breiman, L., Friedman, J., Olshen, R. S. & Stone, C. *Classification and Regression Trees*. Wadsworth International Group. Belmont, CA. 1984
- [2] Cavena, P., Hadjinian P., Stadler, R., Verhees, J. & Zanasi, A. Discovering Data Mining, From Concept to Implementation. Prentice-Hall, Inc. NJ. 1997
- [3] Fayyad U., Piatestsky-Shapiro, G. & Smyth, P. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press. Menlo Park, CA. 1996
- [4] Inaba, A. Application of Boolean approach to meta-analysis. Report of the Fruits of Study by Scientific Research Expenses, 1996-1997, pp. 43-60. 1998
- [5] Iwasaki, M. Data Mining and Knowledge Discovery, From a Statistical Viewpoint. The Japanese Journal of

Behaviormetrics. Vol. 26 No.1 pp.46-58. 1999

- [6] Kanomata, N. On the qualitative comparative analysis program. *Ritsumeikan Social Sciences Review*, Vol. 29, No. 2, pp.85-132 1993
- [7] Kanomata, N. "Self-fulfilling prophecy" and rationality: Thought experiment by Boolean analysis. *Japanese Sociological Review*, 47, pp.156-170. 1996
- [8] Kono, Y. A Design for Contradictory Terms in Boolean Approach: Head and Tail Analysis. Master Thesis, Rikkyo University, Tokyo. 1998
- [9] Kosaka, K. Comparative methods: On C. Ragin's proposals. In J. Kobayashi (Ed.), *Report of the Fruits of Study by* Scientific Research Expenses, 1989-1990, pp. 99-115. 1991
- [10] Misumi, K. Reconsideration for concept of role by Boolean approach. *Preprints of the 24th Conference of the Japanese Association for Mathematical Sociology*, pp.32-35. 1997
- [11] Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc. San Francisco, CA. 1998
- [12] Ragin, C. C. The Comparative method; Moving beyond Qualitative and Quantitative Strategies. The University of California Press. Berkeley, CA. 1987
- Shiota, C. Data mining Techniques and applications. Bulletin of The Computational Statistics of Japan, Vol.10, No.2, pp127-144. 1997
- [14] Tarohmaru. H. Reconstruction of Ehrlich's Die Rechtsfäigkeit: Formalization with Boolean algebra. Sociological Theory and Methods, Vol. 10, No. 1, pp.15-30. 1995
- [15] Tarohmaru, H. Boolean analysis of causes of capacity to hold rights. Nenpo Ningen Kagaku, Vol. 17, 35-47. 1996
- [16] Tarohmaru, H., & Tanaka, S. Boolean analysis with multi-value variables as independent variables. *Bulletin*, Vol. 23, 169-183. 1997
- [17] Weiss, S. M. & Indurkhya, N. Predictive DataMining, A Practical Guide. Morgan Kaufmann Publishers, Inc. San Francisco, CA. 1998
- [18] Yosano, A. Qualitative metric analysis as Boolean analysis and probability model: Comparative note about Boolean and logit analysis. *Report of the Fruits of Study by Scientific Research Expenses, 1996-1997*, pp. 15-24. 1998