Study of graph patterns classification using maximum contrast methods

Naohiko Kinoshita, Toru Takiguchi, Kousuke Takano, Asuka Namiduka

Major in (Field of) Health Informatics and Business Administration, Niigata University of Health and Welfare (NUHW)

Abstract : Maximum contrast methods (MCMs) are developed by assuming various types of contrast vectors in the dose response relationship (DRR). Then, the maximum "goodness of fit" of the DRR pattern is identified using contrast statistics. MCM is generally used for data analyses in toxicological tests and is well-known that the method have the for its high detection power (1-β) for different types of DRRs. In the study on DRRs, researchers set various types of contrast vectors in advance and identified both dose and changing time for the effect (toxicity). However, the pattern consists of only an increase or a stagnation (non-reaction), there is no analysis method that considers a decreasing pattern in openly-available programs. In this study, one-way ANOVA data with a concave or convex type pattern (namely the trend) shows an increase or stagnation tendency, and a decreasing trend is also assumed. The focus herein is to classify a typical graph pattern with an integer ratio by setting both the number of groups and the increase-decrease displacement as a parameter. The classification of graph patterns as concave or convex by generating contrast vectors is considered. The aim of the study is to create a mathematical approach and program modules using contrast vectors corresponding to graph patterns categorized by the above-mentioned method. As a result, an improved binary-direction MCM (bd-MCM) prototype was developed. Bd-MCM presented herein can consider an increasing or decreasing tendency as well as a decreasing trend.

Key words : maximum contrast methods, comparison methods, classification

Introduction

Maximum contrast methods (MCMs) are one of the comparison methods used for trend analysis of dose response relationships (DRRs) in clinical studies, toxicological tests, and so on. In terms of multiple comparisons, Dunnett's test[1] and William test[2] have been used in the past. Nishiyama et al[3]. proposed a composite MCM as an analysis of dose response data in in-vitro toxicity tests. Furthermore, Hamada et al[4]. showed that MCMs had the best detection performance in the evaluation of statistical analyses of outliers and dose-response

【著者連絡先】
〒950-3198　新潟県新潟市北区島見町 1398 番地
新潟医療福祉大学
木下直彦
TEL：025-257-4490　FAX：025-257-4490
E-mail：kinoshita@nuhw.ac.jp
受付日：2018 年 11 月 15 日　受理日：2018 年 12 月 14 日

patterns. MCMs are used when the observation data are one-way ANOVA data with ordinal scale, such as DRR or the data exhibits a linear increase trend after a certain time. The approximate pattern type can be selected by setting the maximum value of the contrast statistic as the test statistic for various linear.

This method can be regarded as a kind of classification problem in which data trend graphs are classified into a number of patterns and the best "goodness of fitness" is selected from them when a change in observation data is represented by the graph.

"Supervised learning" is a class of methods applied to classification tasks in the field of machine learning. These methods involve classifying observational data into an appropriate class using various classifiers. Clustering of so-called unsupervised learning is employed as the method to classify observational data without making an assumed class. Researchers must establish the contrast vector that shows the shape of more than one assumed DRR beforehand when MCM is applied to DRRs.

There are no clear solutions on what kind of matrix must be made or how it must be made, when this contrast vector group is expressed as a matrix (hereinafter referred to as the definition matrix). Nishiyama[5] said that selecting and applying the definition matrix when the dose-response shape is unknown is unclear and indicated that the power of the test ($1-\beta$) should be compared. Each power of the test of five definition matrixes was compared, and an effective result was produced for the definition matrix of DRR.

The definition matrix verified by Nishiyama et al[5]. considers only DRR ; therefore, only the contrast vector with an increasing pattern is assumed in the aforementioned study. The contrast vector with the decrease pattern is not prepared in DRR. Hence, the development of the bi-directional MCM (bd-MCM) corresponding to both increasing patterns, including DRR, and decreasing patterns is considered here. Specifically, the possibility of highly precise class separation was challenged using a method for obtaining the definition matrix with a contrast vector that corresponds to everything with the typified chart pattern having an increasing and decreasing pattern and the observational data was classified.

The purpose of this study is to develop a mathematical model to obtain a typified model pattern from the number of groups and the extent of fluctuation of displacement in this study and to develop a program module that generates a definition line automatically for the chart pattern of everything. In this paper, we first explain the theory of bd-MCM and then the typing of graph patterns to be incorporated into the definition matrix is discussed.

Method

Herein, the classification problem based on the features of bd-MCM, the formulation of the mathematical models based on integer programming approach, and the development of a prototype program module was achieved based on a methodology for solving points. To develop the prototype, visual basic for applications (VBA) in Microsoft Excel was used due to its general versatility and scalability.

The theory of bd-MCM

Regarding the original MCM, the definition of Nishiyama et al[3]. was applied and the minimal quotation of the points are explained below. The observational data assumed in this paper are one-way data with an ordinal scale. At each level of the ordinal scale multiple observations exist, and each level is called a "group". Let k be the num-

ber of groups. There are n_i observation values y_{ij} ; j = 1, 2, ..., n_i in each of the i- group (i = 1, 2, ..., K) and { y_{ij} ; i = 1, 2, ..., K, j = 1,2, ... n_i} that conform to N (μ_i, σ^2) independently of each other. Given that the order of (μ 1, μ 2,... MK) does not change, it is assumed that a graph pattern is formed by the increase / decrease ratio with the adjacent group. For the remainder of this paper, it is assumed that there is no trend as a null hypothesis, that is, assume $H_0 = \mu_1 = \mu_2 = \ldots = \mu_K$. The verification of these tendencies are then considered. In this paper, this null hypothesis H_0 is called a "null hypothesis" following the definition of Yoshida[6].

For the constants c_1, c_2, ..., c_K that satisfy $\sum_{i=1}^{k} c_i = 0$, the statistical quantity defined by the formula (1) is called the "contrast statistics for the contrast vector: c = (c_1, c_2, c_3, ..., c_K)". The superscript ( ) ' indicates transposition.

$$t = \frac{\boldsymbol{c}'\bar{\boldsymbol{y}}}{\sqrt{\hat{\sigma}^2 \boldsymbol{c}' M \boldsymbol{c}}} \quad (1)$$

However,

$$\bar{\boldsymbol{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_K)', \ \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad (2)$$

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \ n = \sum_{i=1}^{K} n_i, \quad (3)$$

$$t_{max} = \max_{l=1,2,\ldots,m} t_l \quad (4)$$

$$C = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_m)' = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1K} \\ c_{21} & c_{22} & \cdots & c_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mK} \end{pmatrix} \quad (5)$$

Where M is a diagonal matrix with i diagonal elements of 1 / n_i. It is well-known that this is the best test statistic for testing the null hypothesis H_0 : c' μ = 0 against the alternative hypothesis H_1 : c' μ > 0. However, μ = (μ_1, μ_2, ... μ_K) '.

Consider the contrast vectors c_1, c_2, ..., c_m corresponding to the types for graph patterns with various types, prepare comparative statistics t_1, t_2, ..., t_m for them(4),to the test statistic is called MCM. Furthermore, a contrast coefficient matrix C obtained by arranging multiple comparison vectors to be used in the form of an m × k matrix as shown in formula (5) is the definition matrix of this MCM.

Classification of graph pattern

When the accuracy is increased, infinite contrast vectors are generated in the classification of graph patterns. In this research, classification is performed by focusing on graph patterns. It is necessary to classify typical graph patterns in advance for this reason. Therefore, the ratio between a certain group and the adjacent group (hereinafter referred to as "ratio of adjacent group (R-AG)") is limited to an integer ratio, and the increase / decrease ratio (D) of the ratio is set as an integer constant. The ratio of adjacent group can be obtained using the following formula:

R-AG: Ratio of adjacent group (r) = {(μ_2 - μ_1) : (μ_3 - μ_2): ...: (μ_{K-1} - μ_K)}

As an example, consider a graph pattern in which the group number K = 3 and the increase / decrease ratio D = 3.

In this case, the number of graph patterns is 3×3×3=27 as shown in Figure. 1.

Among these graphs, the graphs with the same R-AG have the same contrast vector. For this reason, the group ratio is into a graph pattern with relatively prime R-AG with the same condition as that shown in Figure. 1. The graph pattern is then categorized into 13 patterns. (Figure 2)

Furthermore, the definition matrix: C×(-1) is same as that of the definition matrix C owing to the following property of the contrast vectors : c_1 + c_2 + ... c_m = 0. The graph corresponding to the contrast vector obtained by multiplying C ×(-1) is a linear moving average model (x - axis) along the x axis. This indicates that these two graphs have the same contrast statistics. Figure 3 shows the graph patterns of the same contrast statistics.

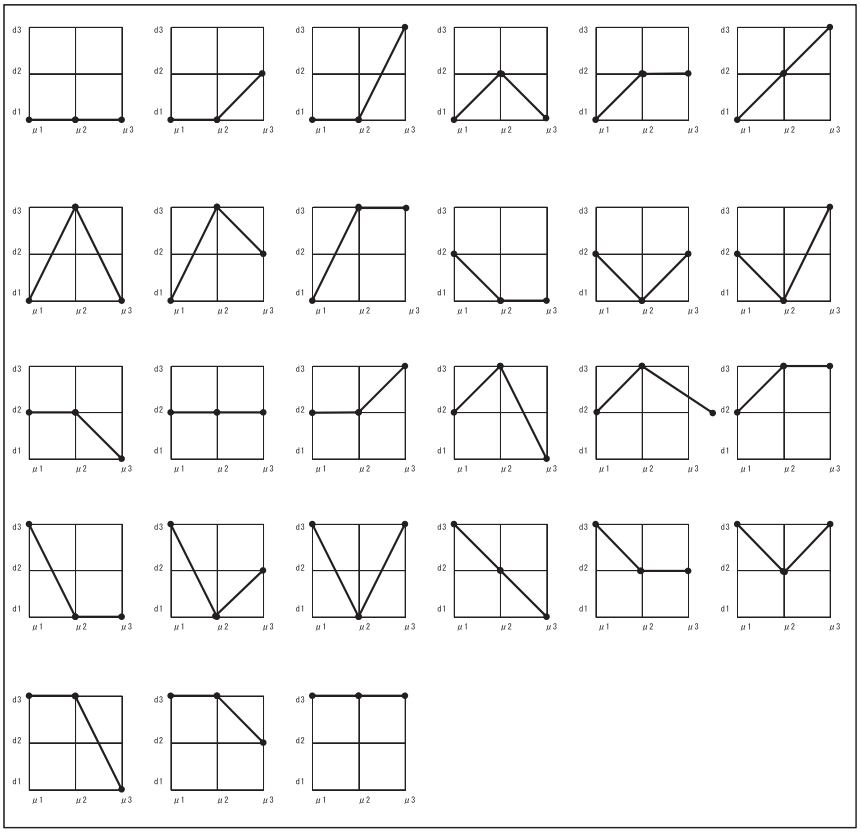

Fig. 1　All graph patterns（K=3，D=3）

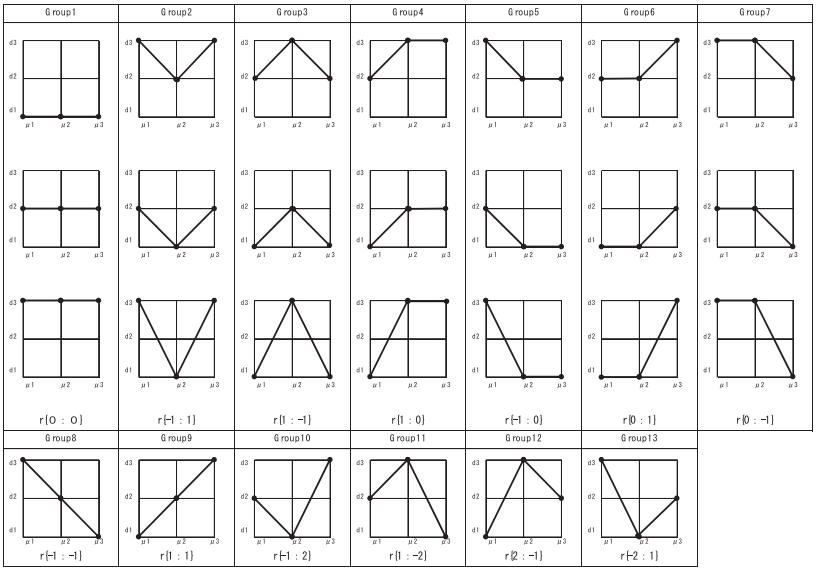

Fig. 2　classifying in graph patterns（K=3，D=3）

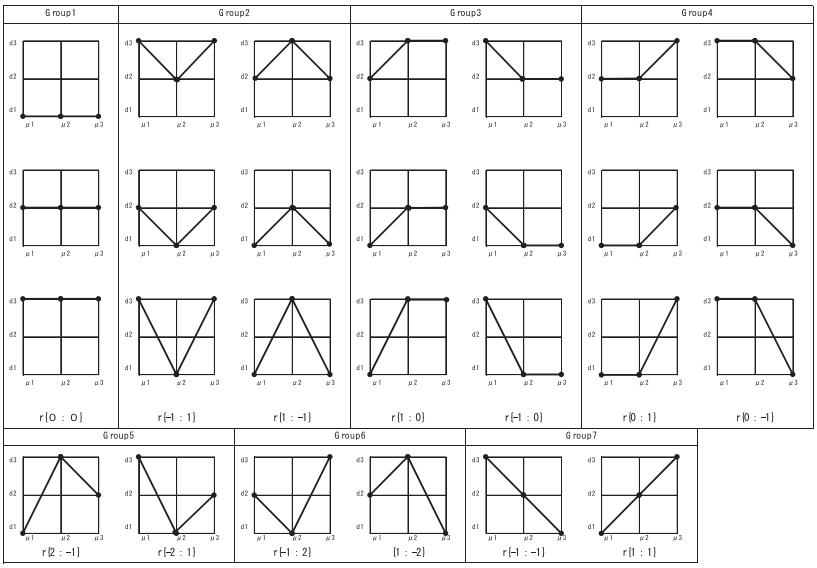

Fig. 3 Group of same contrast statistics using MCMs (maximum contrast methods)

When contrast statistics are considered, the graph patterns are categorized into seven categories; however, the linear shift of the graphs along the x axis is meaningless and become the same class in the classification focusing on increase and decrease . Therefore, it can be said that it is reasonable to adopt the linear categorized into 13 types shown in Figure 2. For this reason, it is necessary to determine which graph pattern is based on the positive / negative information of the adjacent group of observation data for the linearly shifted graph pattern with the same contrast statistics.

Mathematical model

A mathematical model based on integer programming method was developed for classification of graph patterns using the result of typification of graph patterns

Number of groups: K (t_1 ... t_K)
Increase / decrease width: D (maximum = D, minimum = - D). * D is an integer
R-AG: Ratio of adjacent group: r = (r_1: r_2: ....: r_K - 1), * r is an integer satisfying - D <r <D
* r is relatively prime.
Combination of integer numbers satisfying following constraint condition
$-(D-1) \leqq r_1 + r_2 + \cdots + r(K-1) \leqq D - 1$

As an example, Figure 4 shows the rejection area of the R-AG when K = 3 and D = 3. All areas except for the rhomboid area in Figure 4 are rejected. For group number K = 3, given that R-AG is two pairs, it is possible to show the combination region in a two-dimensional graph.

From the set of integers in the region shown in the example (Figure 4), the combination of increase / decrease ratios is obtained by eliminating non-disjoint combinations. Subsequently, all contrast vectors are created by setting constants that satisfy $\sum_{i=1}^{k} c_i = 0$, for these combinations

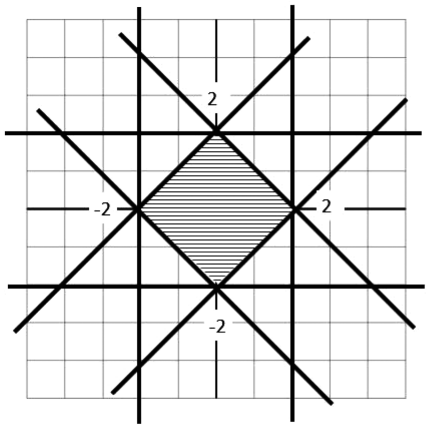

Fig. 4　Rejection area of ratio of adjacent group

Note：The part of the hatchet in Figure 4 shows the existence range of the adjacency increase / decrease ratio value in the case where the X axis and the Y axis are three levels respectively. Numerical values (the convex type is positive and the concavity is negative) with the pattern irregularity information added to the ratio of the distance between group 1　and group 2 and the distance between group 2 and group 3.

Prototype of program module

A prototype program was developed to automatically generate all categorized graph patterns based on mathematical models using an integer programming approach.

By specifying the number of groups (K) and the amount of Displacement (D), it is possible to generate contrast vectors corresponding to automatically categorized graph patterns. For categorized graph patterns, graphs are also automatically generated such that the shape of the graph can be visually recognized.

Conclusion

In this study, a high we attempted to develop a highly versatile classifier for unexpected graph patterns by setting exhaustive contrast vectors. It became clear that it was impossible to identify the

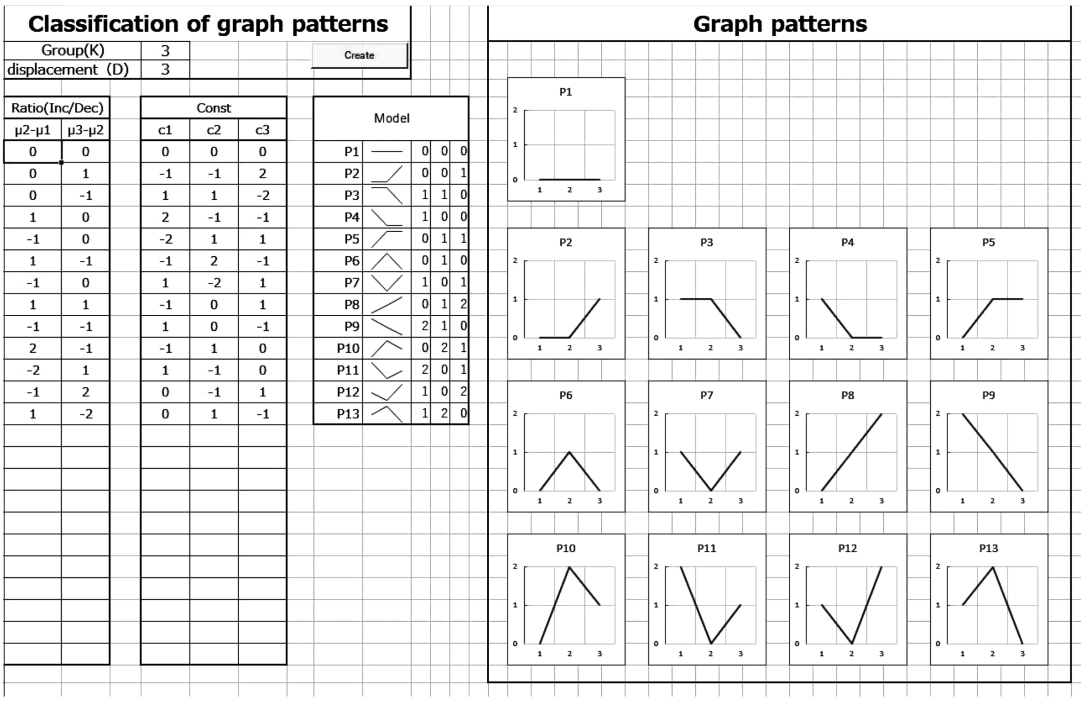

Fig. 5　Prototype of program module

pattern of linear shifted graphs using only bd-MCM. Additional processes were required to determine which graph pattern would be based on the positive / negative information.

By adding this process, it is possible to perform classification into a typed graph pattern; hence, it can be used for classification.

It is desirable that the program module created in this research be applied as that of Nishiyama et al[7]. However, this study focuses on the generation of categorized graph patterns; hence, it is a prototype specialized for generating corresponding contrast vectors. Regarding the release of the program module, it is necessary to check the quality of detection and the sample size design using the maximum comparison method. However, this is out of the scope of this study. We are intend to verify the detection capabilities and sample size design in the future and publish it as a package for programming languages such as R after improving the quality of the current prototype.

References

1) Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control, Journal of the American Statistical Association, 50 : 1096-1121, 1955.

2) Williams, D.A.: A test for differences between treatment means when several dose levels are compared with a zero dose control, Biometrics, 27 : 103-117, 1971.

3) Hiroshi, N., Isao, Y.: Proposal of the Composite Maximum Contrast Method and its Application to Toxicological Data Analysis, 計量生物学 Vol.25, No.1 : 1-18, 2004

4) Hamada, C.: A performance comparison of maximum contrast methods to detect dose dependency, Drug Information Journal, 31 : 423-432, 1997.

5) Nishiyama, H., Omori, T., Yoshimura, I.: A composite statistical procedure for evaluating genotoxicity using cell transformation assay data, Environmetrics, 14 : 183-192, 2003.

6) Yasushi N. and Michihiro Y.: Basis of a statistical multiple comparison way. Scientist company, 1997. (in Japanese)

7) Satoru N., ya Hirokazu Hara and Isao Y.: The SAS/IML program to utilize the biggest method of analog. Measure biology and 24 : 57-70, 2004. (in Japanese)