BALKAN MEDICAL JOURNAL
THE OFFICIAL JOURNAL OF TRAKYA UNIVERSITY FACULTY OF MEDICINE

Balkan Med J 2013; 30: 28-32 • DOI: 10.5152/balkanmedj.2012.070
Available at www.balkanmedicaljournal.org

Original Article

Affected States Soft Independent Modeling by Class Analogy from the Relation Between Independent Variables, Number of Independent Variables and Sample Size

Emine Arzu Kanık[1], Gülhan Orekici Temel[1], Semra Erdoğan[1], İrem Ersöz Kaya[2]

[1]*Department of Biostatistics and Bioinformatics, Faculty of Medicine, Mersin University, Mersin, Turkey*
[2]*Department of Computer Systems, Mersin University, Mersin, Turkey*

ABSTRACT

Objective: The aim of study is to introduce method of Soft Independent Modeling of Class Analogy (SIMCA), and to express whether the method is affected from the number of independent variables, the relationship between variables and sample size.

Study Design: Simulation study.

Material and Methods: SIMCA model is performed in two stages. In order to determine whether the method is influenced by the number of independent variables, the relationship between variables and sample size, simulations were done. Conditions in which sample sizes in both groups are equal, and where there are 30, 100 and 1000 samples; where the number of variables is 2, 3, 5, 10, 50 and 100; moreover where the relationship between variables are quite high, in medium level and quite low were mentioned.

Results: Average classification accuracy of simulation results which were carried out 1000 times for each possible condition of trial plan were given as tables.

Conclusion: It is seen that diagnostic accuracy results increase as the number of independent variables increase. SIMCA method is a method in which the relationship between variables are quite high, the number of independent variables are many in number and where there are outlier values in the data that can be used in conditions having outlier values.

Key Words: Classification, multicollinearity, outlier values

Received: *13.02.2012* ***Accepted:*** *09.07.2012*

Introduction

Many classification methods have been discussed and tried within the classification literature of the health sciences. Among them, the traditional methods, logistic regression and discriminant analysis are used most widely. However, as known, these methods are to provide a series of assumptions in practice. Furthermore, softer models based on iteration, such as Classification and Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS), have come out along with the developments in the computer technology. Although these models are also softer in comparison with the discriminant analysis and logistic regression analysis, some problems occur in practice, arising from the balance between insufficient data or number of independent variables and sample size. The biggest handicap of the CART and MARS methods is that the model established lacks of a test statistics (i.e. the confidence interval of the calculated statistics; hypothesis test control). When the health data of the data group to be classified is taken into consideration, it is nearly impossible to study with p number of independent variable and to form the independency among these variable as well as to compensate between the number of independent variable and sample size. Therefore, these problems cannot be eliminated through traditional methods widely applied. Even though the soft models could eliminate them, the fact that the model established lacks of statistical reliability level culminates in criticism (1-3).

The aim of this study is to introduce the Soft Independent Modeling of Class Analogy (SIMCA) that hasn't been used for the health sciences so far; is not influenced by multicollinearity and tests the significance of the model according to a F test. Furthermore, the number of independent variable in the model highlights whether it is influenced by the relationship among the variables and also sample size.

Material and Methods

Soft Independent Modeling of Class Analogy

The classification model comes out in two stages regardless of the type of classification model applied. At the first stage, a classification model is created, while the inspection whether the new object or observation belongs to this class is carried out at the second stage. SIMCA Model is included

Address for Correspondence: *Dr. Semra Erdoğan, Department of Biostatistics and Bioinformatics, Faculty of Medicine, Mersin University, Mersin, Turkey Phone: +90 324 361 06 84 e-mail: semraerdogann@gmail.com*

within this machine learning method. Suggested first by Wold in 1970s, this method is also known as Supervised Pattern Recognition (4). The establishment of the SIMCA model occurs within two stages. At first stage, the Principal Components Analysis (PCA) is carried out for each group of relevant observations separately. The number of components is determined through cross validation technique. Furthermore, the number of components in the PCA doesn't have to be equal. At the second stage, by means of SIMCA model formed through PCA, the classification of the new object or observation is carried out (5-7).

The Advantages and Disadvantages of the Model

In order to inspect the sufficiency of the sample size within the methods applied for classification, we should pay attention to the proportion between the number of variable and the sample size. However, there is no such limitation in terms of SIMCA method. In this method, an observation cannot be assigned only to a class; sometimes discusses the membership of two and more classes in this model. The term Soft is derived from this point. Within this framework, the outlier values of the data are eliminated (8). A level of statistical significance is calculated according to a F test (5). It is also ideal in terms of classifying these high-level data (9).

Statistical Model

Granted that we have measurable p numbers of variable for n number of observation; J symbolizes the number of group; X^j is the matrix of data. In terms of the matrix of data, i symbolizes the number of observation and j the number of group. The set of learning is shown as follows:

$$x_i^j = (x_{i1}^j, x_{i2}^j, ..., x_{ip}^j) \quad (1)$$

$X_{ik(q)}$; k in the q class. i of the object symbolizes the measurable value in the variable and is shown as in the Equation 2.

$$X_{ik}(q) = a_i(q) + \sum_{a=1}^{Aq} b_{ia}(q) * t_{ak}(q) + e_{ik}(q) \quad (2)$$

Here,
k=1,2…,p (number of variable),
i=1,2,..,n (number of observation),
a=1,2..,A (number of components),
q:1,2,..g,..,r (number of group).

$$S_0(q)^2 = \sum_{k=1}^{q} \sum_{i=1}^{np} \frac{e_{ik}(q)^2}{(p - A_q)(n_q - A_q - 1)} \quad (3)$$

$$S_p(q)^2 = \sum_{k=1}^{p} \frac{e_i(q)^2}{(p - A_q)} \quad (4)$$

$$F = \left(\frac{S_p^{(q)}}{S_0^{(q)}} \right)^2 \quad (5)$$

In classifying a new observation;
If

$$F < F_{p-A_q, (p-A_q)(n_p - A_q - 1); 0.95} \quad (6)$$

Then, the observation can be said to be in q group (5, 9, 10).

Model Statistics

Various classification statistics are calculated in order to discriminate the object or observation from each other in terms of their class membership in SIMCA modeling (8).

The Discrimination Power of the Variable

The discrimination power of a variable indicates the extent of each effect of independent variable in order to discriminate patient and control group. If the discrimination power equals to 1; no discrimination is observed; if equal to bigger than 1, a discrimination can be observed. If the value equals to 3 and more, it is observed that the relevant variable is of vital importance in discriminating the patient and control group.

$$d_k^{(r,g)} = \sqrt{\frac{S_{k,r}^2(g) + S_{k,g}^2(r)}{S_{k,r}^2 + S_{k,g}^2}} \quad (7)$$

$d_k^{(r,g)}$ in the Equation 7 symbolizes the discrimination power of the variable k in terms of r and g group. Other representations in the Equation are stated in the Equation 8-11 (11, 12).

$$S_{k,r}^2 = \sum_{i=1}^{n_r} \frac{e_{ik}^2}{(n_r - A_r - 1)} \quad (8)$$

$$S_{k,g}^2 = \sum_{i=1}^{n_g} \frac{e_{ik}^2}{(n_g - A_g - 1)} \quad (9)$$

$$S_{k,r}^2(g) = \sum_{i=1}^{n_r} \frac{e_{kf}^2(g)}{n_r} \quad (10)$$

$$S_{k,g}^2(r) = \sum_{i=1}^{n_g} \frac{e_{kf}^2(r)}{n_g} \quad (11)$$

Distance Between Groups

Distance between groups symbolizes the extent of the distance among the models formed for patient and control. d (r,g) symbolizes the extent of the distance between r and g group and is calculated as indicated in the Equation 12.

$$d(r.g) = \sqrt{\frac{\sum_{k=1}^{p}\left(S_{k,r}^2(g) + S_{k,g}^2(r)\right)}{\sum_{k=1}^{p}\left(S_{k,r}^2 + S_{k,g}^2\right)}} \quad (12)$$

It is concluded that no classification can be made, if the value equals less than 1 and the discrimination of the classes can be made successfully if the value equals to more than 3-4. The further the distance is, the more successful it is made (11, 12).

Modeling Power

The modeling power is formed for patient and control models separately. It symbolizes the extent of the effect of the variable on the model. $MPOW_k$; k. symbolizes the modeling power of the variable on g model and is calculated as indicated in the Equation 13 (11, 12).

$$MPOW_k = 1 - \frac{\sqrt{\sum_{i=1}^{n_g}\frac{e_{ik}^2}{\left(n_g - A_g - 1\right)}}}{\sqrt{\sum_{i=1}^{n_g}\frac{\left(x-\mu_k\right)^2}{n_g - 1}}} \quad (13)$$

This measurement takes a value between 0 and 1. If the modeling power of the variable equals nearly to less than 0.30; this implies that it is less significant in terms of the model. If this value equals close to 1, this implies that the variant is of importance for the model and has more effects on the model, too.

Simulation Study

The multicollinearity among independent variable in the studies carried out for classification is observed frequently. This situation becomes problematic by means of traditional multivariate statistical methods in practice. Furthermore, it is known that various classification methods are influenced by the sample size. In order to determine whether the method is influenced by the number of independent variable with the SIMCA model, the multicollinearity between variable and sample size, simulations were done.

For this purpose, a trial plan is considered putting forward that the average of independent variable for the first group and that of the second group are 0 and the standard deviations of both groups are 1; i.e. one of the groups does have no discrimination power.

It is also taken into consideration that the sample size of both groups are equal and equal to 30, 100 and 1000 and the number of variable is 2, 3, 5, 10, 50 and 100, as well as the relationships among the variants are of high level (0.95); medium level (0.50) and very low level (0.05). In this manner, the trial plan has 54 combinations, each of which has been tried 1000 times. MATLAB 6.0 Package Program has been applied for the data sets production and SIMCA model applications of the trial plans.

Results

Table 1 indicates the results of average classification accuracy of simulation results carried out 1000 times for each possible situation within the trial plan. In Table 1, p symbolizes the number of variable; R the amount of multicollinearity among the variable and N the size of sample. Figure 1, on the other hand, indicates the visual representation of these simulation results for all combinations.

In terms of the amounts of relationships between the sizes of all possible samples and variable; if the number of independent variable is 2; the diagnostic accuracy results range from 50% and 53%. When the number of independent variable is taken 3; the diagnostic accuracy results range between 51% and 56% and when it is 5 the results range between 52% and 63% (Table 1, Figure 1).

If the number of independent variable is 10; the diagnostic accuracy results range from 54% and 78%. When the number of independent variable is taken 50; the diagnostic accuracy results range between 63% and 96% and when it is 100 the results range between 72% and 100%, However, If the amount of multicollinearity among the variable is 0.05, 0.50 and 0.95 and the size of the sample is 30, the model's classification performance is approximately zero due to the unbalance between the size of sample and the number independent variable (Table 1, Figure 1).

Discussion

For all situations where the multicollinearity among the variable is 0.05, 0.50 and 0.95 and the size of sample is 30,100 and 1000, if the number of independent variable is taken 2.3 and 5; statistics measuring the model's performance of classification cannot reach the sufficient performance. However, the classification capabilities can reach the sufficient level (70% and above), independently of the multicollinearity of variable within the size of the sample, if the number of independent variable is 10, 50 and 100. The fact that the discrimination power cannot reach the sufficient performance implies a parallelism with the literature. Branden et al. (9) suggest that SIMCA method should be applied in the studies carried out for classification, in which the number of independent variable is a lot.

Furthermore, if the number of variable is 2, 3, 5 and 10, the more the size of the sample is, without influencing from the multicollinearity among the independent variable, the less the accuracy value will be. A similar situation can be observed if the number of the independent variable is 100. In this line, there are classification methods in the literature indicating the decrease of accuracy results through the increase of the size of the sample (13).

Sorensen et al. (5) suggest that there is no necessary to inspect the sufficiency of the sample size, as other multivari-

Table 1. Mean and standard deviation values of accuracies for each combination (1000 simulation for each combination)

		P					
R	N	2	3	5	10	50	100
0.05	30	0.53±0.05	0.56±0.06	0.63±0.06	0.78±0.05	0.73±0.06	0.00±0.00
	100	0.51±0.03	0.53±0.03	0.56±0.03	0.64±0.03	0.96±0.01	1.00±0.00
	1000	0.50±0.01	0.51±0.01	0.52±0.01	0.53±0.01	0.63±0.01	0.72±0.01
0.50	30	0.53±0.06	0.56±0.06	0.63±0.06	0.77±0.05	0.78±0.06	0.00±0.00
	100	0.51±0.03	0.53±0.03	0.57±0.03	0.64±0.03	0.96±0.01	1.00±0.00
	1000	0.50±0.01	0.51±0.01	0.52±0.01	0.54±0.01	0.68±0.01	0.79±0.01
0.95	30	0.53±0.06	0.56±0.06	0.63±0.06	0.78±0.05	0.78±0.06	0.00±0.00
	100	0.51±0.03	0.53±0.03	0.57±0.03	0.65±0.03	0.96±0.01	1.00±0.00
	1000	0.50±0.01	0.51±0.01	0.52±0.01	0.55±0.01	0.70±0.01	0.80±0.01

R: The size of multicollinearity among the variables, N: Sample size, p: The number of variable

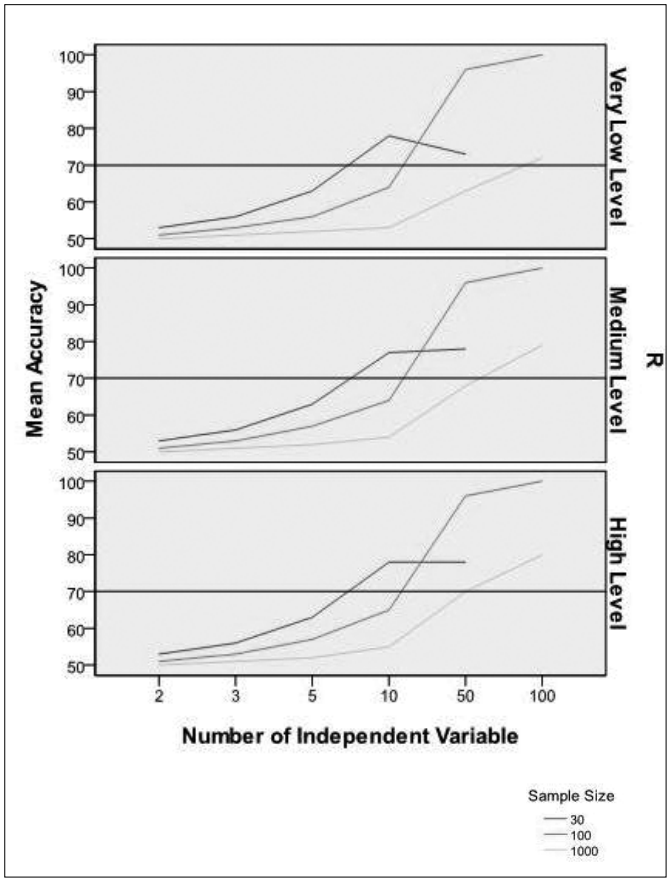

Figure 1. Visual representation of mean and standard deviation values of accuracies for each combination (1000 simulation for each combination)

ate models of the SIMCA modeling. Chen et al. (14) indicate that learning rate of the model decreases when the number of sample size is smaller than the dimensionality of the size. This situation is called small sample size problem by Chen et al. (14). According to the simulation results, although it is observed that the diagnostic accuracy results are influenced by the sample size, model's performance of classification is approximately zero if the number of the independent variable is 100 and the sample size is 30.

Furthermore, if the number of the independent variable and the sample size are bigger than 100, the model fails and culminates in over-fitting.

It is significant that the model sets the importance level of each variable within the framework of the studies carried out for classification. There are methods, such as CART and MARS that can set the importance level of variable; however, they have no test statistics; the confidence interval (15, 16).

The reliability testing of SIMCA method to be applied as alternative grants superiority to this model according to F test. SIMCA method is a method at which the multicollinearity among the variable and the number of independent variable are of high level and which can be applied for outlier values within data and has a statistical significance value (5, 8).

In order to apply the SIMCA modeling for classification, we should first pay attention to the balance between the number of independent variable and the size of sample. If the number of independent variable is taken 2, 3 and 5; statistics measuring the model's performance of classification cannot reach the sufficient performance regardless of the multicollinearity among the variable and the size of sample. Therefore, it symbolizes a method that can be applied if the number of the independent variable is so high, even though a multicollinearity exists among the variable. However, the fact that the number of the independent variable reaches a very high number, like 100, and the sample size becomes 30 and 100 culminates in problems within this context.

Last, but not least, we can conclude that the method cannot be influenced by the multicollinearity among the independent variable; however, the number of the independent variable and the sample size should be taken into consideration.

Ethics Committee Approval: *N/A.*

Informed Consent: *N/A.*

Peer-review: *Externally peer-reviewed.*

Author contributions: *Concept - E.A.K., G.Ö.T., S.E., İ.E.K.; Design - E.A.K., G.Ö.T., S.E., İ.E.K.; Supervision - E.A.K., G.Ö.T., S.E., İ.E.K.; Resource - E.A.K., G.Ö.T., S.E., İ.E.K.; Materials - E.A.K., G.Ö.T., S.E., İ.E.K.; Data Collection&/or Processing - E.A.K., G.Ö.T., S.E., İ.E.K.; Analysis&/or Interpretation - E.A.K., G.Ö.T., S.E., İ.E.K.; Literature Search - E.A.K., G.Ö.T., S.E., İ.E.K.; Writing - E.A.K., G.Ö.T., S.E., İ.E.K.; Critical Reviews - E.A.K., G.Ö.T., S.E., İ.E.K.*

Conflict of Interest: *No conflict of interest was declared by the authors.*

Financial Disclosure: *No financial disclosure was declared by the authors.*

References

1. Srivastava MS. Methods of Multivariate Statistics. Eds:Balding DJ, Bloomfield P, Cressie NAC, A John Wiley&Sons Inc, Canada, 2002. p.246-65.
2. Mcclave JT, Benson GP, Sincich T. Statistics for Business and Economics. 7 th ed, Upper Saddle River N. J, Prentice Hall, 1998. p.551-552.
3. Lattin MJ, Carroll DJ, Green P. Analyzing Multivariate Data. Brooks/Cole-Thompson, Pacific Grove CA, 2003. p.426-428.
4. Wold S. Pattern Recognition by Means of Disjoint Principal Components Models. Pattern Recogn 1976;8:127-39. [CrossRef]
5. Sørensen B, Falk ES, Wisløff-Nilsen E, Bjorvatn B, Kristiansen BE. Multivariate analysis of Neisseria DNA restriction endonuclease patterns. J Gen Microbiol 1985;131:3099-104.
6. Bylesjö M, Rantalainen M, Cloarec O, Nicholoson JK, Holmes E, Trygg J. OPLS Discriminant Analysis:Combining the Strengths of PLS-DA and SIMCA Classification. J Chemomet 2006;20:341-51. [CrossRef]
7. Lopez-de-Alba P, Lopez-Martinez L, Cerda V, Amador-Hernandez A. Simulaneous Determination and Classification of Riboflavini Thiamine, Niotinamide and Pyridoxine in Phamaceutical Formulations, by UV-Visible Spectrophotometry and Multivariate Analysis. J Braz Chem Soc 2006;17:715-22. [CrossRef]
8. Maesschalck RD, Candolfi A, Massart DL, Heuerding S. Decision criteria for soft independent modelling of class analogy applied to near infrared data. Chemometr Intell Lab 1999;47:65-77. [CrossRef]
9. Branden KV, Hubert M. Robust classification in High Dimensions based on the SIMCA Method. Chemometr Intell Lab 2005;79:10-21. [CrossRef]
10. Gemperline P, Webber LD. Raw materials testing using soft independent modelling of class analogy analysis of near-infrared reflectance spectra. J Am Chem Soc 1989;61:138-44.
11. Esbensen KH. SIMCA:An Introduction to Classification. Houmoller LP, eds. Multivariate data analysis in practice:An Introduction to multivariate Stata Analysis and Experimental Design. 5 th ed, Camo process AS, 2005.p:348-51.
12. Dunn WJ, Emery SL, Glen WG. Preprocessing, variable selection and classification rules in the application of simca pattern recognition to mass spectral data. Environ Sci and Technol 1989;23:1499-505. [CrossRef]
13. Zhu M, Shi Y, Li A, He J. A dinamic committee scheme on multiple-criteria linear programming classification method. Computational Science ICCS. 2007;4488:401-8.
14. Chen LF, Liao HYM, Ko MT, Lin JC, Yu GY. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn 2000;33:1713-26. [CrossRef]
15. Breiman L, Friedman JH, Olshen RA, Stone CJ. Introduction to Tree Classification. Classification and Regression Trees. 1st ed, London, Chapman & Hall, 2003. p. 18-55.
16. Friedman JH. Multivariate Adaptive Regression Splines. Ann Stat 1991;19:1-141. [CrossRef]