Different approaches to missing data in discrimination applied to medical problems

Ćwiklińska-Jurkowska M[1], Jurkowski P[2], Kukulska-Pawluczuk B[3], Kołtan A[4], Dróżdż W[5], Hilemann W[2]

[1] Department of Theoretical Background and Medical Informatics; Collegium Medicum; Nicolaus Copernicus University, Poland
[2] Department of Informatics and Research Methodology; Collegium Medicum; Nicolaus Copernicus University, Poland
[3] Department of Neurology; Collegium Medicum; Nicolaus Copernicus University, Poland
[4] Department of Pediatrics, Hematology and Oncology; Collegium Medicum; Nicolaus Copernicus University, Poland
[5] Clinical Neuropsychology Unit; Collegium Medicum; Nicolaus Copernicus University, Poland

Abstract

Purpose: **The aim of the work was the comparison of different methods' usefulness for managing incomplete explanatory data in aiding the recognition for new patients, by different methods of multidimensional discriminant analysis, and in visualizing medical classification problems.**

Material and methods: **Different methods as casewise deletion, single and multiple imputations for different discriminant methods and the distance-based discrimination built on Gower's distance, modified for missing data, were used for supporting diagnosis. The examination was done on three real medical data sets (recognition of childhood asthma, predicting the endpoint of childhood leukemia and diagnosis of neurological stroke kind.**

Results: **The performance of classification was a little worse for discrimination after single and multiple imputation than for casewise deletion. Distance-based discrimination with Gower's distance was next applied. We obtained only 1% decreasing of the patients correctly classified for examined medical data sets comparing the starting complete data set to the set with even about 30% of incomplete data patients – with randomly generated places for removing data, i.e. for artificial creating missing data. Multidimensional metric scaling using Gower's distance in visualizing neurological stroke groups, with mixed variables and missing values, gave similar visualization of patients as ordinary canonical discrimination, which is appropriate for complete data sets.**

Conclusions: **Gower's distance, handling missing values, can be successfully used for distance-based discrimination and visualization of classification problems.**

Key words: aiding medical diagnosis, discrimination, single and multiple imputation, missing values, mixed variables.

ADDRESS FOR CORRESPONDENCE:
Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Background
and Medical Informatics; Collegium Medicum
Nicolaus Copernicus University
ul. Jagiellońska 13-15, 85-067 Bydgoszcz; Poland
Fax: 052-585-3308;
e-mail: mjurkowska@cm.umk.pl

Introduction

Characteristic features of medical data are occurrences of values being lacking – "missing values" and also the mixed quantitative-qualitative character of variables. We examined different methods of handling missing data to use incomplete sets with known clinical diagnosis for the purpose of aiding diagnosis for new patients by discrimination procedures and for the aim of visualizing multidimensional medical classification problems.

Material and methods

Different known methods of treating missing values assume that data are missing at random (MAR – the missing observation may depend on the observed values, but not on the missing values) or even more – missing completely at random (MCAR – the probability of having a missing value is unrelated to the value of this variable or to any other variables in the data set). In other words: for MCAR cases with complete data not differ from cases with incomplete data and for MAR cases with complete data differ from cases with incomplete data and the probability, that the values are missing, depend on the observed values, but not on the missing values. If the pattern of missing data is not random, we call it nonignorable. There is no way to check if the condition of MAR is met, but the MCAR assumption is testable by Little's [1] chi-square test. Missing values in discriminant analysis can be treated in the following ways:

1) casewise (list-wise) deletion: the method is biassed and admissible only, when the number of missing data points is very small (i.e. <5%); assumption – MCAR SPSS, Statistica;
2) omit the most incomplete variables;
3) filling in missing data with plausible values;

3a) Single Imputation: means in groups substituting (or modes for categorical variables) SAS, SPSS, Statistica; multiple regression methods (for categorical variable – discriminant analysis) Solas; Last Value Carried Forward (longitudinal data); Hot decking (identify the most similar case to the case with missing value) Solas; EM (Expectation Maximization) SPSS, SAS

3b) Multiple Imputation (MI) – assuming MAR SAS, Solas: predictive imputation technique Solas, Propensity score SAS, Solas; MCMC (Monte Carlo Markov Chains) SAS.

EM method (Expectation-Maximization) [2,3] is an iterative procedure of missing data estimation by maximizing the likelihood. Rubin's [4] multiple imputation procedure MI replaces each missing value with a set of plausible values that represents the uncertainty about the right value to impute. Just 3-5 imputations are sufficient to obtain good results of filling in data. These multiple imputed data sets are then analysed by using the same discriminant procedures for obtained completed data sets. Next the results of discrimination are combined. MI is forgiving for departures from imputational model [5,6].

The used discriminant methods were: linear, quadratic, kernel and nearest neighbor one [7]. Radii for kernel methods and number of neighbors were chosen to obtain the smallest leave-one-out classification error.

We studied real medical data from Collegium Medium Departments – with mixed variables and many really or artificially missing values. Examined medical data sets are: Treated leukemic children with naturally missing values (21 mixed variables, 114 patients, classification into 2 groups – predicting relapse or death and without the event); Asthmatic children – complete set and the one for discrimination's results comparison – with randomly generated missing values (28 mixed clinical variables and chosen 7 the most discriminating ones, 170 patients, 2 groups – asthma, no asthma); Neurological patients – complete set and the one with randomly generated missing values (25 mixed variables, 170 patients, 4 groups of different cerebral stroke kinds).

The casewise deletion and also single imputation EM and multiple imputation for filling in missing data were used. The applied method of multiple imputation was for example EM is (Expectation-Maximization with importance resampling) by King et al. [8]. The number of imputed data sets is five.

Results

The best cross-validation result of all considered discriminant classifiers was obtained for the kernel normal discrimination as well for incomplete data, after casewise deletion, as for single and multiple imputed data. The results of variables' selection, for data sets completed by multiple imputations, are nearly the same as for real data set after casewise deletion. We obtained similar results of discrimination effectiveness for single imputation and multiple imputation.

Discrimination after single and multiple imputation gave little worse percentage of correctly classified patients than after casewise deletion. Thus, we used modified algorithm for calculation of Gower's distance, also appropriate in the situation with missing values. Complete data of patients from Department of Neurology were entered to our program and we compared results with the same set, where 33 randomly selected patients have missing values. Good classification results for using this method are also confirmed for set with missing values in about 30% of patients – for the diagnosis of asthmatic children. At this point, we obtained only 1% decreasing of correct classifications.

Using multidimensional metric scaling (MDS) [9], technique and Gower's distance [10] we visualized four group neurological patients, described by mixed variables and with missing values. First we obtained a plot of four neurological group means. Next we incorporated individual variability of four neurological diseases onto the same plot using Gower's [11] adding-a-point technique and Gower's distance. The procedure was done for full data set and for the set with randomly rejected 33 values of variables, as described above. In this way we obtained scatterplots of neurological patients after MDS on general diversity distances [12] and add-a-point procedure. These two scatterplots are similar to the plot of patients in two canonical variates space for full data set.

Discussion

We obtained little worse results of classification correctness for MI and EM, than for casewise deletion. It may be caused by changed related size of smaller group. However, recently the questions for validity of multiple imputations are raised [13]. So we also suggest another approach to missing values – the programmed by us modification of Gower's [10] distance between two observations (patients) with possible missing values in mixed variables. This enables using discrimination with missing values by applying distance-based discrimination [14], based on inter-observations (patients) distance.

Using Gower's measure for distance-based discrimination we obtained decreasing percentage of the correct classifications for examined medical data set only by 1%, comparing the beginning complete data set and the set with even about 30% of patients – chosen randomly – with generated randomly places for removing data.

Gower's distance can be also very successful, as applied to visualize groups of patients, when occur missing data.

References

1. Little RJA. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 1988; 83: 1198-202.
2. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B, 1977; 39: 1-38.
3. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.
4. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley & Sons Inc; 1987.

5. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research, 1998; 33: 545-71

6. Kott PS. A Paradox of Multiple Imputation. Proceedings of the Survey Research Methods Section, American Statistical Association, 1995; 380-3.

7. Webb AR. Statistical pattern recognition. New York: Wiley &Sons; 2002.

8. King G, Honaker J, Joseph A, Scheve K. Analysing Incomplete Political Science Data: An Alternative for Multiple Imputation. American Political Science Review, 2001; 95: 49-69.

9. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academia Press; 1979.

10. Gower JC. A general coefficient of similarity and some of its properties. Biometrics, 1971; 27: 857-74

11. Gower JC. Adding a Point to Vector Diagrams in Multivariate Analysis. Biometrika, 1968; 55: 582-5.

12. Krzanowski WJ. Ordination in the presence of group structure for general multivariate data. Journal of Classification, 1994; 11: 195-207.

13. Fay R. When are inferences from multiple imputations valid? Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria; 1999: 227-32.

14. Cuadras CM. A distance based approach to discriminant analysis and its properties. Mathematic Preprint Series, No 90, Second version, Univ of Barcelona; 1991.