Thailand Statistician
January 2008; 6(1) : 91-100
http://statassoc.or.th
Contributed paper

Robust Fuzzy Discriminant Analysis in Presence of Outliers by Genetic Algorithms

Chutima Hongsawat and Saengla Chaimongkol*
Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Phathum Thani, 12121, Thailand.
***Author for correspondence**; e-mail: saengla@mathstat.sci.tu.ac.th

Received: 1 September 2007
Accepted: 5 November 2007.

Abstract

This paper studies a robust fuzzy discriminant analysis (RFDA) to deal with outliers in the crisp data. The difference between RFDA and classic fuzzy discriminant anlaysis (FDA), which is based on the distance, is that RFDA uses robust distance to measure the similarity between data points. The performance of RFDA is evaluated by a simulation study. The data are generated by using the Monte Carlo simulation technique. The data are arbitrarily set to 2, 3, and 4 groups with the sample size of 30 each. In each group, there are 3 and 10 independent variables and there are 1 and 3 outliers. The genetic algorithm implement by MATLAB version 6.5 is used with 100 runs for each setting condition. Based on the mean of apparent error rates and correct identification rates of outliers, the results reveal that RFDA is more satisfactory for group classification and outlier detection than the FDA.

Keyword: fuzzy discriminant analysis, fuzzy shapes, genetic algorithm, MCD-estimators, outlier detection.

1. Introduction

Discriminant analysis [2] is a statistical technique of classifying an individual into one of two or more populations on the basis of measured variables. To conduct discriminant analysis, one can use maximum likelihood method, distance function or linear classification function. The main objectives of discriminant analysis are separation of groups as much as possible based on the measured variables and classification of new individual into a labeled group. The idea of discriminant analysis technique is that an individual can only belong to one group. However, outcome variables or measured variables are often vague and ambiguous. Individuals do not necessarily belong only to a single group. In this situation, the discriminant analysis will be a problematic method for classification. Therefore, there is a need to develop new discriminant technique for the classification problem on the basis of fuzzy variables and it is called fuzzy discriminant analysis (FDA).

The basic idea of FDA is that an individual that belongs to one group may simultaneously belong to other groups with different membership degrees. The aim of FDA is to find a membership function for each group and use it to determine the membership degree. Recently, several methods for FDA have been developed. For example, Watada et al. [9] introduced a fuzzy discriminant analysis method only for fuzzy data in fuzzy groups. Chen et al. [1] presented a fuzzy discriminant analysis for crisp chemical data sets with a few overlapping data points that are considered equally important for all groups in ordinary discriminant analysis. Lin and Chen [5] proposed a fuzzy method for two-group discriminant analysis where the membership function of the groups to be discriminated is obtained by minimizing the sum of square of classification rates. Furthermore, Lin and Chen [6] also proposed a method for classifying fuzzy multiple groups' discriminant analysis of crisp data. Based on the distance between individuals and centroids of groups, their method can detect membership function of each group by minimizing the classification error using genetic algorithm. If an individual is close to the centroid of all groups, a sum of degrees of membership will be high, implying that the individual is more likely to belong to more than one group simultaneously. In contrast, if an individual has zero degree of membership to each group, then it can be considered as an outlier.

Because discriminant rules are based on estimates of the population parameters, outliers might have an influence on the results of discriminant analysis. The outliers might shift the estimated mean and might blow up the dispersion matrices. To prevent this problem, one may use robust estimate of the population parameters. Many researchers have considered robust linear discriminant analysis (e.g., Hawkins and McLachlan [3], He and Fung [4]) but the method of robust estimators in fuzzy discriminate analysis has not been found in the literature.

This paper robustifies FDA by using robust distance, called robust fuzzy discriminant (RFDA). Robust Mahalanobis distance, a classical Mahalanobis distance computed by using robust estimators, MCD-estimators, of mean and covariance, is used in this paper to measure the similarity between individuals. The FAST-MCD algorithm of Rousseeuw and Van Driessen [7] is used to calculate the MCD-estimators. To conduct RFDA, a genetic algorithm is used to determine the membership function of each group by minimizing the classification error. The objective of this paper is to study effect of outliers on the RFDA and FDA of the crisp data by comparing their apparent error rates and correct identification rates of outliers.

2. Genetic Algorithm

Genetic algorithm [8] is a stochastic search technique based on the mechanism of natural selection and genetics. Genetic algorithms start with potential solutions of individuals forming as an initial population. Each individual in the population called chromosome is evaluated using its fitness, which is related to the objective function of the problem. Through reproduction, crossover, and mutation, the population of the next generation is created. Each generation, the fitter chromosomes are more likely to be selected and used to produce new chromosome called offspring through crossover and mutation operators. The crossover operation merges two chromosome selected by exchanging some genes between the chromosomes with a crossover rate. Mutation alters one or more genes of a selected chromosome to form a new chromosome with a mutation rate. The procedure continues in this fashion until the termination condition is satisfied. Hopefully, the best chromosome obtained is the optimal solution to the problem.

The genetic algorithm used in this study follows Lin and Chen's procedures [6]. A population that consists of 100 chromosomes representing potential solutions to the problem is used . Each chromosome has a length of 12 genes and each gene is represented by a digit from 0 to 9. This means that the decimal encoding is used rather than the binary coding. The first six digits and last six digits of each chromosome represent the two real numbers with four decimal points. The first and second real numbers are mapped onto domains of a_i and b_i, respectively. The a_i and b_i form a fuzzy boundary that surrounds to group *i*. For this study, a_i is set to lie in an interval [-*k*, 99.9999-*k*], where *k* is any positive number, and b_i is set to, lie in [0, 99.9999]. The fitness, the link between genetic algorithm and discriminant analysis, is to minimize the sum of squared classification errors for each group. That is, minimize $E_i = \sum_{j=1}^{n} (\mu_i(D_{ij}^2) - y_{ij})^2$ for *i= 1, 2,..., m* (number of group), (1)

where y_{ij} is the dummy variable which y_{ij} = 1 if observation *j* belongs to group *i* and y_{ij} =0, otherwise. D_{ij}^2 is the distance between observation *j* and the centroid of group *i*. $\mu_i(D_{ij}^2)$ is the membership function of group *i*. Two shapes of fuzzy set, Z-shaped and Semi-Right bell shaped fuzzy set, are used to represent the membership function of each group. Thus, the membership function of group *i* for Z-shaped and Semi-Right bell shaped fuzzy set can be defined as in equation (2), and (3), respectively.

$$\mu_i(D_{ij}^2) = \begin{cases} 1, & \text{if } D_{ij}^2 \le a_i, \\ \dfrac{b_i - D_{ij}^2}{b_i - a_i}, & \text{if } a_i < D_{ij}^2 \le b_i, \\ 0, & \text{if } b_i < D_{ij}^2. \end{cases} \tag{2}$$

and

$$\mu_i(D_{ij}^2) = \begin{cases} 1, & \text{if } D_{ij}^2 \le a_i, \\ 1 - 2\left(\dfrac{D_{ij}^2 - a_i}{(a_i + b_i)/2 - a_i}\right) & \text{if } a_i < D_{ij}^2 \le (a_i + b_i)/2, \\ 2\left(\dfrac{D_{ij}^2 - b_i}{(a_i + b_i)/2 - a_i}\right) & \text{if} (a_i + b_i)/2 < D_{ij}^2 \le b_i, \\ 0, & \text{if } D_{ij}^2 > b_i. \end{cases} \tag{3}$$

The objective of genetic algorithm is to find a_i and b_i that can maximally differentiate the observations that belong and do not belong to group *i.* The highly fitted chromosomes will reproduce, called reproduction. To create offspring for the next generation, the roulette selection method is used to select the chromosome in the population. The basic idea of the roulette selection [8] is that a selection probability for each chromosome is determined by using the proportion of its fitness value to the total fitness value for all of chromosomes in the population.

Originally, to form a new generation of genetic algorithm, all parents will be replaced by their offspring. However, this replacement may produce degenerating population; in other words, offspring may be less fit than their parent. To prevent the populations from degenerating, for each time of producing an offspring, if it fits better than the worst parent, this offspring will replace a randomly chosen parent which is not the best one.

After reproduction, the crossover and mutation operators are applied to the population to create the new population of the next generation. This paper chooses one point crossover technique and the crossover rate, which determines how often the crossover operator is invoked, is set to 0.9. The probability of performing mutation operators is set to 0.5. The total of 5,000 generations is used to terminate the procedure of genetic algorithms.

3. Robust Fuzzy Discriminant Analysis

Robust fuzzy discriminant analysis (RFDA) mentioned in this paper is a fuzzy discriminant analysis (FDA) using robust distances based on the MCD-estimators. FDA applies fuzzy theory to discriminant analysis with its objective to determine membership degree of belonging in each group for each individual. Suppose that *n* individuals are to be discriminated into *m* groups. The membership degree that individual *j* belongs to group *i* can be determined from the group's membership function, μ_i An effective μ_i should give a small difference between the membership degree and the actual membership value of individual *j*, while retaining a specific degree of fuzziness. In other words, μ_i should minimize the classification error at individual *j*. Thus, the objective function of FDA is defined in equation (1). That is, it minimizes the sum of squared classification errors for each group. In this paper, Z-shaped and Right-bell shaped fuzzy set are used to represent the membership function of each group expressed in equations (2) and (3), respectively.

The *a_i* and *b_i* in equation (2) and (3) are a fuzzy radius $\tilde{R}_i = [a_i, b_i]$ from the centroid of group *i*. The region between a_i and b_i indicates the fuzzy boundary around group *i*. The a_i and b_i are obtained from the genetic algorithm described in section 2.

Difference between classic FDA and RFDA is that RFDA replaces the mean and pooled covariance matrix of Mahalanobis's distance which are used in the membership function by using their MCD-estimators. The MCD-estimators are given by the subset of size *h* for which the determinant of its covariance matrix is minimal. The MCD-estimator of location is then given by the mean of these *h* observations and the MCD-estimator of covariance is given by their covariance matrix. The distance between point x_j to the centroid of group *i* is defined as

$$D_{ij}^2 = (x_j - c_i)' G (x_j - c_i), \qquad (4)$$

where c_i is MCD-estimator of location and *G* is MCD-estimator of pool covariance matrix, which is symmetric and positive-definite. For the MCD-estimators, the FAST-MCD algorithm proposed by Rousseeuw and Van Driessen [7] is used. The basic ideas of FAST-MCD are an inequality involving order statistics and determinant, and techniques of selective iteration and nested extension.

4. Simulation Studies

The primary objective of this study is to investigate the effect of outliers on apparent error rate and correct identification rate of outliers of RFDA, comparing to those under FDA. The outliers deviating in location from the original distributions of each population are studied. In addition, the number of populations varies to 2, 3, and 4 with the sample size of 30 for each and the number of independent variables (*p*) set to 3, and 10, are considered. Thus, we generate 30 random numbers normally distributed of *p*-variate observations from each population,

where $\mathrm{P}_1 \sim N_p(\mu_1, \Sigma_1)$, $\mathrm{P}_2 \sim N_p(\mu_2, \Sigma_2)$, $\mathrm{P}_3 \sim N_p(\mu_3, \Sigma_3)$, and $\mathrm{P}_4 \sim N_p(\mu_4, \Sigma_4)$. We consider in case of $\mu_1, \mu_2, \mu_3,$ and μ_4 the mean vectors with *p*-dimension of <u>5</u>, <u>10</u>, <u>15</u>, and <u>20</u>, respectively. The covariance matrices are assumed to be equal, that is $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \Sigma$. The variances of dimension *p* are set to 1, and the correlations between two independent variables are 0.1. After the data from each group are generated, we add outliers of 1 and 3 to the original generated data. The outliers in each group are generated by shifting 5 times of the original group's mean vector, but keeping equal covariance matrix. We generate 100 different data sets for each condition, from which we apply both FDA and RFDA. For each data set, the apparent error rate (*APER*) and the correct identification rate of outliers are computed. Then, over the 100 data sets, the mean of apparent error rate ($\overline{APER}$) and mean of correct identification rate of outliers $(\overline{CR})$ are computed. These fours values are defined as:

$$APER = \frac{\sum_{i=1}^{m} n_{iM}}{\sum_{i=1}^{m} n_i} \text{ X } 100,\ \overline{APER} = \frac{\sum_{t=1}^{100} APER_t}{100}, \qquad (5)$$

where n_{iM} is the number of observations classified into other groups rather than group *i*, and n_i is the total number of observations of group *i*. *t* is the number of run or total number of data set.

$$CR = \frac{\sum_{i=1}^{m} nout_{iM}}{\sum_{i=1}^{m} nout_i} \text{ X } 100,\ \overline{CR} = \frac{\sum_{t=1}^{100} CR_t}{100}, \qquad (6)$$

where $nout_{iM}$ is the number of outliers identified correctly in group *i*, and $nout_i$ is the total number of outliers of group *i. t* is the number of run or total number of data set.

The values of $\overline{APER}$ and $\overline{CR}$ are used to examine the effect of the outliers on the performance of FDA and RFDA.

5. Results

The results of $\overline{APER}$ for both FDA and RFDA are presented in Table1. From this table, it can be concluded that the Z-shaped and Semi-Right bell shaped membership function revealed the same of $\overline{APER}$ across the number of populations, independent variables, and outliers in each group in both FDA and RFDA. RFDA had much smaller $\overline{APER}$ than the FDA across number of population, independent variables, and outliers. The values $\overline{APER}$ of FDA are proportional to the number of outliers in each group, and the number of populations. In contrast, the number of populations and the number of independent variables do not have any effect on the values of $\overline{APER}$ for RDFA. The values of $\overline{APER}$ for RFDA do only depend on the number of outliers in each group, where the three outliers in each group had the $\overline{APER}$ higher about three times of a single outlier.

Table1. Mean of apparent error rates over 100 runs for both FDA and RFDA under all conditions.

Number of populations	Number of independent variables	Number of outliers in each group	FDA		RFDA	
			Z-shape	Semi-Right Bell Shape	Z-shape	Semi-Right Bell Shape
2	3	1	22.7742	21.0323	3.4032	3.3226
		3	43.303	43.2273	9.1818	9.0909
	10	1	16.3065	16.6452	3.5484	3.4355
		3	24.4848	25.3636	9.2273	9.1970
3	3	1	40.2473	38.0000	3.3978	3.3118
		3	55.8687	54.2929	9.2121	9.2121
	10	1	27.1505	26.3226	3.5914	3.5591
		3	38.7172	36.5758	9.3030	9.2828
4	3	1	48.3065	46.5161	3.4032	3.3710
		3	62.3106	59.8333	9.2652	9.1364
	10	1	35.5887	33.0403	3.4839	3.5323
		3	44.7273	45.0076	9.2955	9.2197

The results of $\overline{CR}$ in Table2 show that RFDA has almost perfectly identifying the outliers regardless of the membership function, the number of populations, number of independent variables, and number of outliers. In contrast, FDA had zero identifying outliers when there are three outliers in each group. For a single outlier, the semi-right bell shaped membership function had much smaller $\overline{CR}$ than the Z-shaped membership functions. However, the Z-shaped membership function showed below 25% for the values of $\overline{CR}$, which are considered very low ability in detecting outliers.

Table2. Mean of correct identification rates for outliers over 100 runs for both FDA and RFDA under all conditions.

Number of populations	Number of independent variables	Number of outliers in each group	FDA		RFDA	
			Z-shape	Semi-Right Bell Shape	Z-shape	Semi-Right Bell Shape
2	3	1	21.0	12.0	100	99.5
		3	0.00	0.00	99.5	99.4
	10	1	11.0	0.50	100	100
		3	0.00	0.00	99.8	100
3	3	1	15.7	9.30	100	99.7
		3	0.00	0.00	99.7	100
	10	1	16.3	1.00	99.7	100
		3	0.90	0.00	100	99.7
4	3	1	20.3	11.3	99.8	98.5
		3	0.00	0.00	99.3	98.8
	10	1	23.0	3.50	99.5	97.5
		3	0.00	0.00	99.3	99.0

6. Conclusion and recommendation

The results of the study reveal that RFDA is a more suitable method for group classification and outlier detection than FDA in presence of outliers in location. Its superior performance can be seen from the small error rates and almost perfect identifying capability outliers. Even if no outlier is present in the populations, there is no loss of using RFDA in classification because the FAST-MCD algorithm can be used to reduce computational time.

This study also shows that genetic algorithm is a good approach to minimize the classification errors, especially when the objective function is nonlinear. The Z-shaped and Semi-Right bell shaped fuzzy set used to represent the membership function of each group do not have any effect on the performance of genetic algorithm. Thus, the simplest and widely used, Z-shaped fuzzy set, should be chosen.

References

[1] Chen, Z.P., Jiang, J.H., Liang, Y.Z., and Yu, R.O. Fuzzy Discriminant Analysis for Chemical Systems, Chemometrics Intelligent Laboratory Systems, 45; 1999: 295-302.

[2] Johnson, R. A., and Wichern, D. W. Applied Multivariate Statistical Analysis. 5th edition. NJ: Prentice Hall, 2002.

[3] Hawkins, D.W., and McLachlan, G.J. High-Breakdown Linear Discriminant Analysis. Journal of the American Statistical Association, 92: 1997: 136-143.

[4] He, X.M., and Fung, W.K. High-Breakdown Estimation for Multiple Population with Applications to Discriminant Analysis. Journal of Multivariate Analysis, 72; 2000: 151-162.

[5] Lin, C.C., and Chen, A.P. A Method for two group fuzzy discriminant analysis. International Journal Fuzzy Systems, 3; 2001: 341-345.

[6] Lin, C.C., and Chen, A. P. Fuzzy Discriminant Analysis with Outlier Detection by Genetic Algorithm. Computers and Operations Research, 31; 2004:877-888.

[7] Rousseeuw, P.J., and Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41; 1999: 212-223.

[8] Sakawa, M. Genetic Algorithms and Fuzzy Multiobjective Optimization. Boston: Kluwer, 2002.

[9] Watada, J., Tanaka, H., Asai, K. Fuzzy discriminant analysis in fuzzy groups. Fuzzy Sets and Systems, 19; 1986: 261-271.