Beyond Fisher's Linear Discriminant Analysis

- New World of Discriminant Analysis -

Shinmura, Shuichi
Seikei Univ., Faculty of Economics
3-3-1 Kichijoji-kitamachi,
Musashino City (Tokyo 180-8633), Japan
E-mail: shinmura@econ.seikei.ac.jp

1. Introduction

Fisher (1936) proposes Fisher's linear discriminant function (**LDF**), and opens the new world of discriminant analysis. It is very essential methodology in industry and science. It is approached from various research areas such as statistics, pattern recognition and mathematical programming (**MP**).

1) In 1930s, Fisher introduces LDF under the assumption that distributions of two classes are same variance-covariance matrices of normal distribution (**Fisher's assumption**). After his theory, many discriminant methods are developed such as quadratic discriminant function (**QDF**), Mahalanobis distance for multi-group discrimination, and MT (Mahalanobis - Taguchi) theory in quality control, logistic regression etc.
2) In the 1950s, Pattern Recognition starts to identify the character.
3) Since the 1970s, there are many papers in regression and discriminant analysis using MP. Stam (1997) seriously asks us "Why have statisticians rarely used Lp-norm method". This answer is very easy, because these researches are not evaluated by real data.
4) Vapnik (1995) proposes fantastic methods named support vector machine (**SVM**), such as hard margin SVM, soft margin SVM, kernel SVM. These methods are evaluated by real data in many areas. Hard margin SVM propose new idea such as maximization of margin. It ascertains the generalization ability as same as the examination by the evaluation data in statistics.
5) After 1997, Shinmura (1998, 2000, 2004, 2007, 2009) and Shinmura & Tarumi(1999) develop several new linear discriminant functions based on minimum number of misclassifications (**MNM**) criterion (Shinmura & Miyake, 1979) named the optimal linear discriminant function (**OLDF**).

There are several problems about the discriminant analysis. **IP-OLDF** resolves these problems. And it is concluded that **Revised IPLP-OLDF** and **Revised IP-OLDF** are superior to LDF, QDF, logistic regression, hard margin SVM and soft margin SVM by many experimental studies.

2. Problems of Discriminant Analysis

There are six problems about discriminant analysis.

1) It is very strange that number of misclassifications (**NM**) has been neglected by evaluation of discrimination. NM of LDF and QDF are regressed by MNM, and these results reveal that QDF is very weak for multicollinearity (Shinmura, 1998).
2) Nobody explains how to treat the cases on discriminant hyper-plane that are indefinite in class1/class2. If we can't count NM correctly, we can't evaluate the results of discrimination.

3) The linear discriminant functions defined by MP have special features. Some cases are fixed on the discriminant hyper-plane or support vector (**SV**). This feature causes trouble to count NM correctly except for hard margin SVM and Revised IP-OLDF.
4) Most of data doesn't satisfy **Fisher's assumption**. Therefore, QDF, logistic regression and QT are developed. It is unfortunately there is no approach by MNM criterion.
5) Influence statistics is very important in statistics. But there are no 95% significant intervals of NM and discriminant coefficients. Therefore, there is no reason why discriminant methods must satisfy **Fisher's assumption**.
6) Discriminant methods except for Revised IP-OLDF and hard margin SVM can't find the minimum dimension of linear separable data space (MNM=0).

Above problems are completely resolved by OLDF using LINGO that is MP solver developed by Schrage (2006).

3. New methods

After 1997, several new methods are developed. Important methods are **IP-OLDF** by integer programming (**IP**), **LP-OLDF** by linear programming (**LP**), **Revised IP-OLDF** and **Revised IPLP-OLDF**.

3.1 IP-OLDF

IP-OLDF minimizes NM in formula (3.1). If $\mathbf{x_i}$ is classified correctly, e_i=0 and y_i $*f_i(\mathbf{b})= y_i *(\mathbf{x_i}'\mathbf{b}+1)>=0$. If $\mathbf{x_i}$ is misclassified, e_i =1 and $y_i *f_i(\mathbf{b})>=$ -1000000. This means that IP-OLDF choose the discriminant hyper-plane $f_i(\mathbf{b})=0$ for classified cases, and $f_i(\mathbf{b})= -1000000$ for misclassified cases by 0/1 decision variable.

$MIN = \Sigma\ e_i$

$y_i *(\mathbf{x_i}'\mathbf{b}+1) >= - M*e_i$ (3.1)

$\mathbf{x_i} =(x_{i1}, x_{i2}, \dots, x_{ip})$: p-independent variables, i=1,…,n

y_i = 1 for $\mathbf{x_i}$ ∈class1, y_i = -1 for $\mathbf{x_i}$ ∈class2, $\mathbf{b}$: p-discriminant coefficients

e_i : 0/1 decision variable corresponding to each $\mathbf{x_i}$, M: 1000,000 (Big M constant)

But, this notation has weakness. We must solve formula (3.2), because we can't decide y_i=1 for class1 and y_i= -1 for class2. Only data decides it. Therefore, both models must be solved.

$MIN=\Sigma e_i$

$y_i *(\mathbf{x_i}'\mathbf{b}$ - 1) >= - $c*e_i$ (3.2)

The constant of f_i $(\mathbf{b})$ is fixed to +1(/-1). This is very important, because we can exchange $\mathbf{x_i}$ and $\boldsymbol{b}$ such as $\mathbf{x_i}'\boldsymbol{b}$+1=$\boldsymbol{b}'\mathbf{x_i}$+1. We can consider IP-OLDF on both p-dimensional data space ($\boldsymbol{b}'\mathbf{x_i}$+1) and discriminant coefficient space (**xi'b**+1).

$f_i(\mathbf{b}) = 0$ becomes a linear hyper-plane that divides p-dimensional discriminant coefficients space into two subspaces. The dicriminant coefficients space are divided in finite convex polyhedron by n linear hyper-planes of case $\mathbf{x_i}$. Every interior point $\boldsymbol{b}$ of one of this convex polyhedron discriminate some cases correctly, and misclassify the other. If $y_i* f_i(\mathbf{b}) = y_i*(\mathbf{x_i}'\ \mathbf{b} +1) = y_i*(\mathbf{b}'\ \mathbf{x_i}$ +1) > 0, $\mathbf{x_i}$ is discriminated correctly in data space. If $y_i* f_i(\mathbf{b}) = y_i*(\mathbf{x_i}'\ \mathbf{b} +1) = y_i*(\mathbf{b}'\ \mathbf{x_i} +1) <$ 0, $\mathbf{x_i}$ is discriminated incorrectly in data space. The vertex of convex polyhedron consists of over p-cases. These cases are on the discriminant hyper-plane in data space.

Therefore, interior point of convex polyhedron discriminates same cases correctly/incorrectly. Every interior point has unique NM. And there is at least one Optimal Convex Polyhedron with MNM. If we choose the linear discriminant function corresponding to the interior point, there is no cases on the discriminant hyper-plane, because $f_i(\mathbf{b}) \neq 0$. This resolves one of the serious problems.

On the other hand, IP-OLDF finds the solution on the vertex of Optimal Convex Polyhedron and *p* cases lie on the discriminant hyper-plane in the data space if data satisfies Haar's condition. And (p+1) cases lies on the hyper-plane if data doesn't satisfy Haar's condition. In this case, MNM of IP-OLDF may not be true MNM. Some e_i on the discriminant hyper-plane may be 1, nevertheless all ones on it are counted to zero.

3.2 Revised IP-OLDF

Revise IP-OLDF in formula (3.3) resolves two defects of IP-OLDF, because it finds the interior points of **Optimal Convex Polyhedron** directly. The misclassified cases are attracted to $(\mathbf{x_i}'\mathbf{b}+b_0) = -999999$ and discriminant scores $(\mathbf{x_i}'\mathbf{b}+b_0)$ are less than equal -1. No cases are on $(\mathbf{x_i}'\mathbf{b}+b_0) = 0$. This means the solution is an interior point of Optimal Convex Polyhedron defined by IP-OLDF.

$$MIN = \Sigma e_i$$

$$y_i * (\mathbf{x_i}'\mathbf{b}+b_0) >= 1 - M*e_i \qquad (3.3)$$

b_0 : free decision variables

3.3 Revised LP-OLDF

If e_i is changed from 0/1 decision variable to non-negative real value in (3.3), Revised LP-OLDF is defined. Revised LP-OLDF minimizes the summation of distances of misclassified cases from the discriminant hyper-plane. Revised LP-OLDF is faster than Revised IP-OLDF, because it is solved by LP.

3.4 Revised IPLP-OLDF

Revised IP-OLDF has the defect that it needs enormous calculation time, especially for 100 fold cross-validations by 135 different discriminat models. Therefore, faster algorithm named Revised IPLP-OLDF combined with Revised IP-OLDF and Revised LP-OLDF is developed.

In first step, Revised LP-OLDF is applied for the sample data. If e_i is zero, x_i is classified correctly by SV. And these cases are excluded from the optimization by fixing e_i to zero in second step.

In second step, Revised IP-OLDF is applied for the misclassified cases in first step. Therefore, computation of Revised IP-OLDF is restricted for sub cases. This is reason why Revised IPLP-OLDF is faster than Revised IP-OLDF. It finds the approximation of MNM.

The 95% significant intervals of NM and discriminant coefficients of Revised IPLP-OLDF are obtained by 100 fold cross validations.

4. Solving Problems of Discriminant Analysis

New knowledge is obtained by IP-OLDF. Revised IP-OLDF obtains true MNM. .

4.1 Monotonous decrease of MNM

MNM has a remarkable feature such as $MNM_k >= MNM_{(k+1)}$. This means that $\mathrm{MNM_k}$ of Revised IP-OLDF having k-independent variables is greater than $MNM_{(k+1)}$ of Revised IP-OLDF added one variable to the former.

If $\mathrm{MNM_k} = 0$, then $\mathrm{MNM_{(k+1)}} = 0$. This feature reveals that only Revised IP-OLDF and hard margin SVM can find the minimum dimension of independent variables space that is linear separable. Other discriminant functions can't always find it. Until now, most statisticians believe the discrimination for linear separable data is easy, nevertheless LDF, QDF and logistic regression have different troubles for it. One of troubles is that stepwise methods, AIC and Cp statistics can't work correctly for linear separable data. Swiss bank note data collected by Flury and Rieduyl (1988) is a famous data for discrimination. There are 100 genuine and 100 counterfeit bills having six measurements. Revised IP-OLDF reveals that this data is linear separable in two dimensions (X4, X9) by the examination of all possible models. On the other hand, AIC and stepwise methods choose 5-independent variables, and Cp statistics choose 6-independent variables. In addition to these facts, NMs of these methods are not zero. This fact is confirmed by the modified student and CPD data. From these two kinds of data, linear separable data are generated by expanding the range of average of two classes. The same results are observed as same as Swiss bank data (Shinmura,2007).

4.2 Fisher's assumption and Inferential Statistics

LDF assumes **Fisher's assumption**. Many statisticians and statistical users doubt this one. Therefore, QDF and logistic regression are developed. And there are many researches in the fields of pattern recognition and MP. Nevertheless LDF assume **Fisher's assumption**, confidence intervals / standard errors of NM and the discriminant coefficients are not known. LDF is irrelevant from inferential statistics. There is no need to assume that the two classes are normally distributed.

Revised IPLP-OLDF can compute confidence intervals by 100 fold cross validation.

5. Experimental Study

In this study, four kinds of raw data are used for evaluation. In first stage, these data are used to examine the validity of new methods. These methods are compared with LDF, QDF, decision tree and logistic regression. In second stage, 20,000 resampling data sets are generated from raw data by Speakeasy. Raw data are used as training sample, and resampling data are used as evaluation data. NM by Revised IPLP-OLDF equal to MNM by 149 different discriminant models (Shinmura, 2009). In third stage, 100 resampling data sets having same size of raw data are generated from raw data. And 135 different discriminant models of LDF, logistic regression and Revised IPLP-OLDF is evaluated by 100 fold cross-validation.

5.1 Four kinds of Real data

Student data consists of 40 students having five independent variables. Object variable consists of two groups such as 25 students who pass the examination and 15 students who don't pass. All combinations of independent variables (31= 2^5-1) are investigated.

Iris data (Edgar,1935) consists of 100 cases having four independent variables. Object variable consists of two species such as 50 versicolor and 50 virginica. All combinations of independent variables (15= 2^4-1) are investigated.

CPD data (Shinmura & Miyake, 1979) consists of 240 patients having 19 independent variables. Object variable consists of two groups such as 180 pregnant women whose babies are born by the natural delivery and 60 pregnant women whose babies are born by Caesarian operation. Forty models selected by forward and backward stepwise methods are analyzed, because we can't examine (2^{19}-1) models by all combinations of independent variables. There are three multicollinearities in this data.

Swiss bank notes data (Flury & Rieduyl, 1988) consists of 200 cases having six independent variables.

Object variable consists of two kinds of bills such as 100 genuine and 100 counterfeit bills. Sixty three (= 2^6-1) models are investigated.

Therefore, there are 149 different models for experimental sturdy.

Two different types of resampling data are generated by Speakeasy. In second stage, four resampling data sets having 20,000 cases are generated from raw data. These data sets are used by the evaluation data. In third stage, 100 data sets having the same size of raw data sets are generated. Those are used as 100 fold cross-validations.

5.2 Results in first stage (1997-2006)

Results in first stage are summarized in Shinmura (2007).

5.3 Results in second stage (2007-2009)

Revised IP-OLDF can finds MNM of training data, but it requires large computation (CPU) time. Therefore, if Revised IPLP-OLDF gives us good approximations of MNM, it is used instead of Revised IP-OLDF. In order to confirm this, four resampling data sets having 20,000 cases are generated from raw data by Speakeasy. Raw data are used as training data, and resampling data sets are used as evaluation data. Revised IPLP-OLDF is compared with Revised IP-OLDF in 149 different discriminant models. Our second concern is how Revised IPLP-OLDF reduces CPU time compared to Revised IP-OLDF. The following results are obtained (Shinmura, 2009).

1) Revised IPLP-OLDF significantly improves CPU time.
2) In the training data, all 149- NM of Revised IPLP-OLDF equal to the MNM of Revised IP-OLDF.
3) In the evaluation data, most of NM of Revised IPLP-OLDF equal to NM of Revised IP-OLDF.
4) The generalization abilities of both methods are concluded to be high, because the difference between the error rates of training and evaluation data are almost within 2%.

Therefore, it is concluded that Revised IPLP-OLDF is useful to analyze experimental sturdy of 100 fold cross-validation in third stage on behalf of Revised IP-OLDF.

5.4 Results in third stage (2010)

One hundred resampling data sets are generated by uniform random numbers. These data sets have the same size (cases and variables) of raw data. One hundred thirty five different discriminant models of LDF, logistic regression and Revised IPLP-OLDF are done by 100 fold cross-validations. Fourteen models are dropped from CPD data. There are 100-NM and discriminant functions for 135 different discriminant models. One hundred thirty five mean error rates, and 95% confidence intervals of error rates and discriminant coefficients are calculated.

Mean error rates of Revised IPLP-OLDF are compared with LDF. All results of LDF are bad for the training samples. We obtain same results about CPD data for the evaluation samples. Only 15 (2 in Iris, 10 in Bank, 3 in student data) out of 109 models of LDF are good for the evaluation samples.

Mean error rates of Revised IPLP-OLDF are compared with logistic regression. Only 2 out of 15 models of logistic regression are good for iris data in evaluation data. Only 24 out of 63 models of logistic regression are good for Bank data in evaluation data. Only 3 and 7 out of 63 models of logistic regression are good for student data in training and evaluation data respectively. Other results of logistic regressions are bad.

6. Conclusion

IP-OLDF based on MNM criterion finds new knowledge about discriminant analysis. Optimal

Convex Polyhedron reveals the relation of NM and discriminant coefficients. MNM finds the defects of model selections in Swiss Bank note data. It reveals the difficult problems about discrimination of the linear separable data. LDF, QDF and logistic regression can't find the minimum dimension of discriminant coefficients space. Model selections by LDF choose higher dimension. QDF is very we ak for CPD data having multicollinearities. It frequently misclassifies the one group to the other. Estimations of logistic regression coefficients become unstable for the linear separable data. On the other hands, IP-OLDF may not find true MNM if the data doesn't satisfy Haar's condition. Therefore, **Revised IP-OLDF** is proposed. It can find the interior points of Optimal Convex Polyhedron directly, and avoids cases on the discriminant hyper-plane.

At last, NM of Revised IPLP-OLDF are compared with those of LDF and logistic regression by 100 fold cross-validations. The mean error rates of LDF are better than Revised IPLP-OLDF in only 15 out of 135 different discriminant models for evaluation samples. The mean error rates of logistic regressions are better than **Revised IPLP-OLDF** in only 3 out of 135 models for training samples and 33 out of 135 models for evaluation samples.

It is concluded that MNM criterion is robust, and Revised IP-OLDF gives the lower limit of NM of all linear discriminant functions such as LDF, hard margin SVM and soft margin SVM.

REFERENCES (RÉFERENCES)

Edgar, A. (1935). The irises of the Gaspé Peninsula. Bulletin of the American Iris Society, **59**, 2–5.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, **7**, 179–188.

Flury, B. & Rieduyl, H. (1988). Multivariate Statistics: A Practical Approach. Cambridge University Press.

Schrage, L. (2006). Optimization Modeling with LINGO. LINDO Systems Inc.

Shinmura, S. & Miyake, A. (1979). Optimal linear discriminant functions and their application, COMPSAC 79, 167-172.

Shinmura, S. (1998). Optimal Linear Discrimrnant Functions using Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **11 /** 2 , 89-101.

Shinmura, S. & Tarumi, T. (1999). Evaluation of the Optimal Linear Diseriminant Functions using Integer Programming for the Normal Random Data. Journal of the Japanese Society of Computer Statistics, **12 / 2**, 107-123.

Shinmura, S. (2000). A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics, **5**, 133-142.

Shinmura, S. (2004). New Algorithm of Discriminant Analysis using Integer Programming. IPSI 2004 Pescara VIP Conference CD-ROM, 1-18.

Shinmura, S. (2007). Overviews of Discriminant Function by Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **20**/1-2, 59-94.

Shinmura, S. (2009). Improvement of CPU time of Revised IPLP-OLDF using Linear Programming. Journal of the Japanese Society of Computer Statistics, **22**/1, 37-57.

Stam, A. (1997). Nontraditinal approaches to statistical classification: Some perspectives on Lp-norm methods. Annals of Operations Research, 74, 1-36.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, 1995.