Comparison of Linear Discriminant Functions by K-fold Cross Validation

Shuichi Shinmura
Faculty of Economics, Seikei Univ.
Tokyo, Japan
shinmura@econ.seikei.ac.jp

***Abstract*— To discriminate two classes is essential in the science, technology, and industry. Fisher defined the linear discriminant function (Fisher's LDF) based on the variance-covariance matrices. It was applied for many applications. After Fisher's LDF, several LDFs such as logistic regression and a soft margin support vector machine (S-SVM) are proposed. But, there are serious two problems of the discriminant analysis. First, the numbers of misclassifications (NMs) or error rates by these LDFs may not be correct because these LDFs cannot discriminate cases on the discriminant hyper-plane correctly. Second, these LDFs cannot recognize the linear separable data properly. Only revised optimal LDF by integer programming (Revised IP-OLDF) resolves these problems. In this paper, we compare seven LDFs by 100-fold cross validation using 104 different discriminant models. It is shown that the mean error rates of Revised IP-OLDF are better than other LDFs in the training and validation samples.**

Keywords- Fisher's linear discriminant function; logistic regreesion; soft margin SVM; Revised IP-OLDF; minimum number of misclassifications; k-fold cross validation.

I. INTRODUCTION

To discriminate two classes or objects is essential in the science, technology, and industry. Fisher defined the linear discriminant function (LDF) to maximize the variance ratio (between classes/within class) [2]. If two classes satisfy the Fisher's assumption that two classes belong to the normal distribution such as N_i (x: m_i, Σ_i) i=1, 2 and $\Sigma_1 = \Sigma_2$, the same LDF is formulated by the plug-in rule in (1).

$$\text{Log}\,(N_1(x: m_1, \Sigma_1) / N_2(x: m_2, \Sigma_2)) = 0 \qquad (1)$$

And it is defined by the variance-covariance matrices explicitly in (2).

$$f(\mathbf{x}) = \{\mathbf{x}-(\mathbf{m}_1+\mathbf{m}_2)/2\}'\textstyle\sum^{-1}(\mathbf{m}_1-\mathbf{m}_2) \qquad (2)$$

x: p-independent variables (p-features).
$\mathbf{m}_1/\mathbf{m}_2$: mean vectors in class1/class2.
$\sum$: pooled variance-covariance matrix.

Statistical software packages adapt this equation, and many useful methods such as the variable selection methods are developed. It was applied for many applications such as the medical diagnosis, genome discrimination, pattern recognition, the rating of stocks and the pass/fail determination of exams score [16] etc.

The discriminant rule is very simple: If $y_i*f(\mathbf{x}_i) > 0$, $\mathbf{x}_i$ is classified to class1/class2 correctly. If $y_i*f(\mathbf{x}_i) < 0$, $\mathbf{x}_i$ is misclassified. This simplicity may hide the following problems:

1) Problem 1: If there are cases on the discriminant hyper -plane ($f(\mathbf{x}_i)=0$), we cannot discriminate these cases correctly. This is the unresolved problem of discriminant analysis. Until now, most statistical user treats that these cases belong to class1 without any reason. Some statisticians explain that this is decided by the probability because statistics is a study, which is based on the probability. But statistical software adopt former rule. And the medical doctors who use the discriminant analysis in the medical diagnosis are surprised and disappointed by the latter explanation. They devote heart and soul to discriminate the patient near by the discriminant hyper-plane.

2) Problem 2: A hard margin SVM (H-SVM) defines the discrimination of linear separable data clearly. But there are few researches about it. First reason is that Fisher's LDF, logistic regression and soft-margin SVM (S-SVM) cannot recognize linear separable data. Second reason is there are no good research data of linear separable data. Ranges of 18 error rates of Fisher's LDF and quadratic discriminant function (QDF) are [2.2%, 16.7%] and [0.8%, 8.5%] by the pass/fail determination of exams scores [18], nevertheless those are linear separable.

These two problems are resolved by IP-OLDF and Revised IP-OLDF [19] [21].

3) Problem 3: The discriminant functions based on the variance-covariance matrices need to compute the inverse matrices. But if some variables are constant, those are not computed. The generalized inverse matrix technique may be expected to resolve this defect. But the serious problem is found in the special case in QDF [18].

In this research, problem 4 is discussed.

4) Problem 4: After Fisher's LDF, many LDFs are proposed. There are few comparisons of these LDFs. In this research, seven LDFs are compared by k-fold cross validation using 104 different discriminant models of four data such as Fisher's iris data [1], Swiss bank note data [3],

 ISBN: 978-1-61208-358-2

Cephalo Pelvic Disproportion (CPD) data [11], and the student data [14].

II. LINEAR DISCRIMINANT FUNCTIONS COMPARED IN THIS RESEARCH

After Fisher's LDF, QDF and the multi-class discrimination using the Maharanobis distance are proposed in the statistical approach. These methods are formulated by the variance-covariance matrices. In this research, only seven LDFs in this chapter are compared by 100-fold cross validation in order to approach problem 4.

A. Logistic Regression

In the medical diagnosis, the discriminant methods are very important and useful. But, real data scarcely satisfy the Fisher's assumption, especially in the epidemiological study. Therefore, the logistic regression in (3) was developed by Framingham sturdy.

$$\text{Log}\,(P_i/(1-P_i)) = b_1x_1+\ldots+b_px_p+b_0 \quad (3)$$

P_1 / P_2: the probability of the normal / ill class.
$(x_1,\ldots,x_p)$ or $\mathbf{x}$: p-features (independent variables) vector.
$(b_1,\ldots,b_p)$ or $\mathbf{b}$: p-discriminant coefficients vector.
b_0: the constant of LDF.
n: the number of cases (n_1/n_2: the normal / ill class)

B. Support Vector Machine

The regression and discriminant analyses are easily approached by the mathematical programming (MP) because MP can find the minimum / maximum (global optimal) value of function. Before SVM, there are many researches of L_p-norm discriminant functions by linear programming (LP). Stam summarized these researches and sorrowed "statistical users rarely use these functions" [22]. Statistical users use SVM because there are many evaluations of SVM by the real data. On the other hand, there are no evaluations of the MP-based discriminant functions before SVM. Vapnik proposed three kinds of SVM, such as H-SVM, S-SVM and kernel SVM [23]. H-SVM in (4) indicates the discrimination of linear separable data definitely. Cases $\mathbf{x}_i$ are classified correctly by the support vectors (SVs). The object function minimizes (1/ the distance between two SVs). This is to maximize the distance between two SVs. It has been proven that the generalization ability of H-SVM is good.

$$\text{MIN} = \|\mathbf{b}\|^2/2; \quad y_i*(\mathbf{x}_i'\mathbf{b}+b_0) \geq 1; \quad (4)$$
$$y_i = 1 / -1 \text{ for } \mathbf{x}_i \in \text{class1/class2}.$$

Real data are rarely linear-separable. Therefore, S-SVM has been defined in (5). S-SVM permits certain cases that are not discriminated by SV ($y_i*f(\mathbf{x}_i) < 1$). The second objective is to minimize the summation of distances of misclassified cases (Σe_i) from SV. These two objects are combined by defining "penalty c." The Markowitz portfolio model to minimize risk and maximize return is as same as S-SVM. However, the return is incorporated in the constraint, and the objective function minimizes only risk. The decision maker chooses a good solution on the efficient frontier. On the contrary, S-SVM does not have a rule to determine c. Nevertheless, it can be solved by an optimization solver. In this research, we try to evaluate two S-SVMs (c =10^4 and 1).

$$\text{MIN} = \|\mathbf{b}\|^2/2 + c*\Sigma e_i; \quad y_i*(\mathbf{x}_i'\mathbf{b}+b_0) \geq 1 - e_i; \quad (5)$$
e_i : non-negative decision variable.
c: penalty c to combine two objectives.

C. Heuristic-OLDF and IP-OLDF

Shinmura and Miyake [10] developed the heuristic algorithm of OLDF based on MNM criterion affected by Warmack and Gonzalez [24]. This OLDF solved only five features (5-vars) model of CDP data because of the lack of the CPU power.

SAS was introduced into Japan in 1978 [6]. LINDO was introduced into Japan in 1983. Several regression models are formulated by MP [8]. Least-squares method can be solved by QP, and Least Absolute Value (LAV) regression is solved by LP. Without a survey of previous research, the formulation of IP-OLDF [12][13] can be defined as in (6).

$$\text{MIN} = \Sigma e_i; \quad y_i*(\mathbf{x}_i'\mathbf{b}+1) >= -M*e_i; \quad (6)$$
e_i: 0/1 integer variable corresponding to $\mathbf{x}_i$.
M: 10,000 (Big M constant).

This notation is defined on p-dimensional coefficients space because the constant of LDF is fixed to 1. In pattern recognition, the constant is a free variable. In this case, the model is defined on (p+1)-coefficients space, and we cannot elicit the same deep knowledge as with IP-OLDF. This difference is very important. IP-OLDF is defined on both p-dimensional data and coefficients spaces. This is very important to find new facts of the discriminant analysis [14]. We can understand new knowledge of the discriminant analysis about the relation between the NMs and LDFs clearly. This relation tells us the following new facts and shows a clue of problem solving.

1) Fact 1: Optimal Convex Polyhedron

The linear equation $H_i(\mathbf{b}) = y_i*(\mathbf{x}_i'\mathbf{b}+1) = 0$ divides p-dimensional coefficients space into plus and minus half-planes ($H_i(\mathbf{b}) > 0$, $H_i(\mathbf{b}) < 0$). If $\mathbf{b}_j$ is in the plus half-plane, $f_j(\mathbf{x}) = y_i*(\mathbf{b}_j'\mathbf{x}+1)$ discriminates $\mathbf{x}_i$ correctly because $f_j(\mathbf{x}_i) = y_i*(\mathbf{b}_j'\mathbf{x}_i+1) = y_i*(\mathbf{x}_i'\mathbf{b}_j+1) > 0$. On the contrary, if $\mathbf{b}_j$ is included in the minus half-plane, $f_j(\mathbf{x})$ cannot discriminate $\mathbf{x}_i$ correctly because $f_j(\mathbf{x}_i) = y_i*(\mathbf{b}_j'\mathbf{x}_i+1) = y_i*(\mathbf{x}_i'\mathbf{b}_j+1) < 0$. The n linear equations $H_i(\mathbf{b})$ divide the coefficients space into a finite number of convex polyhedrons. Each interior point of a convex polyhedron has a unique NM that is equal to the number of minus half-planes of n linear equations. We define the "Optimal Convex Polyhedron (OCP)" as that for which NM is equal to MNM.

2) Fact 2: $\text{MNM}_q \geq \text{MNM}_{(q+1)}$

Let us MNM_q be MNM of q-vars model, and $\text{MNM}_{(q+1)}$ be MNM of (q+1)-vars model adding one variable to the

former. The proof is very easy. The OCP of q-vars model is concluded in (q+1)-discriminant coefficients space. At least, we know there exits the convex polyhedron in (p+1)-coefficients space, NM of which is MNM_q.

3) Fact 3: Two kinds of the discrimination

If MNM_q =0, all MNMs including these q-features are zero. IP-OLDF found Swiss bank note data is linear separable by 2-features such as (X4, X6). It consisted of two kinds of bills: 100 genuine and 100 counterfeit bills. There were six features: X1 was the length of the bill; X2 and X3 were the width of the left and right edges; X4 and X5 were the bottom and top margin widths; X6 was length of the image diagonal. A total of 63 ($=2^6$-1) models were investigated. We had better considered about two types of discriminations: 16 linearly separable discriminant models, and other 47 models. This data is adequate whether or not LDFs can discriminate linearly separable data correctly.

D. Revised IP-OLDF

The Revised IP-OLDF in (7) can find the true MNM because it can directly find the interior point of the OCP. This means there are no cases where $H_i(\mathbf{b}) = 0$. And only Revised IP-OLDF is free from problem 1. If $\mathbf{x}_i$ is discriminated correctly, $e_i = 0$ and $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= 1$. If $\mathbf{x}_i$ is misclassified, $e_i = 1$ and $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= -9999$. It is expected that all misclassified cases will be extracted to alternative SV, such as $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) = -9999$. Therefore, the discriminant scores of misclassified cases become large and negative, and there are no cases where $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) = 0$. Revised IP-OLDF can resolves first and second problems. Therefore, it is ready to be compared with other LFDs by 100-fold cross validation.

$$MIN = \Sigma e_i ; \quad y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= 1 - M* e_i ; \quad (7)$$

b_0: free decision variables.

If e_i is a non-negative real variable, we utilize Revised LP-OLDF, which is an L1-norm LDF. Its elapsed runtime is faster than that of Revised IP-OLDF. If we choose a large positive number as the penalty c of S-SVM, the result is almost the same as that given by Revised LP-OLDF because the role of the first term of the objective value in equation (5) is ignored.

Revised IPLP-OLDF is a combined model of Revised LP-OLDF and Revised IP-OLDF. In the first step, Revised LP-OLDF is applied for all cases, and e_i is fixed to 0 for cases that are discriminated correctly by Revised LP-OLDF. In the second step, Revised IP-OLDF is applied for cases that are misclassified in the first step. Therefore, Revised IPLP-OLDF can obtain an estimate of MNM faster than Revised IP-OLDF [20].

It is regretful that all LDFs except for Revised IP-OLDF are not free from problem 1.

III. THE ROLL OF DATA IN THE RESEARCH

This basic research started after 1997 and ended in 2012. There are the following reasons why it needed sixteen years.

1) IP solver requested huge computation time before 2000 [20]. Therefore, it was too earlier to start from 1997.

2) IP-OLDF may not find true MNM if data is not in general position. This is not confirmed without the survey using the student data. Ibaraki and Muroga defined the same Revised IP-OLDF already [4]. But, it is very difficult to find mechanism why it can find the true MNM without the examination by real data and previous research of IP-OLDF.

In the first stage of this basic sturdy, the iris and CPD data were used for the evaluation of IP-OLDF and comparison with Fisher's LDF and QDF. IP-OLDF finds new facts such as: 1) the relation of NMs and LDFs, and 2) OCP, 3) MNM decreases monotonously. In the second stage, IP-OLDF finds that Swiss bank note data is linear separable. The student data reveals the defect of IP-OLDF that relates to problem 1. Even now, many researchers are not aware of this problem. Revised IP-OLDF resolved to find the interior point of the OCP directly.

After 2009, we started the applied research of linear separable data. I negotiated with the National Center for Univ. Entrance Examination (NCUEE), and got research data consisting of 105 exams in 14 subjects over three years. It was confirmed that error rates of LDFs except for Revised IP-OLDF cannot definitely recognize the linear separable data. More specifically, those of Fisher's LDF and QDF are very high. Eighteen pass/fail determinations of my statistical lectures are used for the research data. Tests have 100 items with 10 choice that are categorized four testlets scores such as: T1, T2, T3 and T4. If the pass mark is 50 points, a trivial LDF such as f = T1 + T2 + T3 + T4 - 50 can discriminate the pass/fail classes completely by the rule: $f \geq 0$ or $f < 0$. Students on the discriminant hyper-plane (f=0) are classified in the pass class because the discriminant rule is defined by four features definitely. Discrimination by 100 items finds serious problem 3 about the algorithm of the generalized inverse matrices of QDF. By the discrimination using four testlets, the error rates of Fisher's LDF and QDF are very high, and this is confirmed by 100-fold cross validation [15] [17].

IV. K-FOLD CROSS-VALIDATION

Re-sampling samples are generated from 4 real data sets. These are analyzed by 100-fold cross validation. Fisher's LDF and logistic regression are analyzed by JMP [7]. JMP division of SAS Institute Japan supports us to develop the program. Other LDFs are analyzed by LINGO [9]. LINDO Systems Inc. supports us to develop the program that is showed in [18][21]. The most important interest is the mean error rates of seven LDFs in the training and validation samples.

A. 100-fold cross validation of CPD

CPD data consisted of two classes: 180 pregnant women whose babies were born by natural delivery and 60 pregnant women whose babies were born by Caesarean section. There

were 19 features such as: X1 was the pregnant woman's age, X7 was the shortest anteroposterior distance, X8 was the fetal biparietal diameter, and X9 was X7-X8, X12 was X13-X14 (small normal random noise are added to X9 and X12), X13 was the area at the pelvic inlet, X14 was the area of the fetal head, and X19 was the lateral conjugate. X9 and X12 cause the multicolinearity. About 19 models selected by the forward stepwise method from 1-vat to 19-vars, NM of QDF is as follows: 22→20→22→18→18→16→15→9→9→ 8→9→ 21→17→16 →17→21→19→17→16. From 11–features to 12-features, NM increases 9 to 21 because X14 enter the 11-vars model and 12-vars model includes (X12, X13, X14). On the other hand, MNM decreases monotonously.

TABLE I. CPD DATA

OLDF	M1	M2
1-19	0.04	3.70
1-5,7-19	0.06	3.68
1,2,4,5,7-19	0.08	3.72
1,2,4,5,7-9,11-19	0.13	3.86
1,2,4,5,7,9,11-19	0.18	3.73
1,2,4,5,7,9,11-15,17-19	0.26	3.59
1,2,4,5,7,9,12-15,17-19	0.44	3.76
1,2,5,7,9,12-15,17-19	0.56	3.88
1,2,5,7,9,12,13,15,17-19	0.57	3.74
1,2,5,7,9,12,15,17-19	0.63	3.71
1,2,5,7,9,12,15,17-18	0.82	3.70
1,2,7,9,12,15,17-18	1.66	4.81
1,2,9,12,15,17-18	1.88	4.67
2,9,12,15,17-18	2.24	4.68
9,12,15,17-18	3.24	6.12
9,12,15,18	3.54	5.56
9,12,18	4.35	6.06
9,12	4.81	5.99
12	7.80	9.15
	M1Diff.	M2Diff.
SVM4	[0.08, 2.56]	[0.08, 1.07]
	0.45	0.39
SVM1	[1.03, 2.56]	[0.28, 1.43]
	1.76	1.43
LP	[0.07, 2.56]	[0, 1.28]
	0.45	0.28
IPLP	[-0.01(1), 0.05]	[-0.25(10), 0.11]
	0	0.02
Logistic	[0.18, 2.94]	[0.23, 1.52]
	0.97	0.63
LDF	[2.95, 7.69]	[1.68, 5.92]
	7.52	5.69

We examine 19 different models of CPD data selected by the forward stepwise method because there are over 500,000 models ($=2^{19}$-1). Table I shows the results by 100-fold cross validation. 'OLDF' is the result of Revised IP-OLDF. First column of 'OLDF' shows 19 models from 19-vars model to 1-var model. "M1 and M2" are the mean error rates for the training and validation samples. Those are computed by mean of 100 error rates of 19 different models. Therefore, M1 decreases monotonously as same as MNM. M1 of the full model is always minimum. The minimum value of M2 is 3.59% of 14-vars model. Therefore, we compare Revised IP-OLDF with six LDFs by this model.

"SVM4, SVM1, LP, IPLP, Logistic and LDF" are the results of S-SVM ($c=10^4$ and 1), Revised LP-OLDF, Revised IPLP-OLDF, logistic regression and Fisher's LDF, respectively. "M1 Diff. and M2Diff." are the difference of ((M1/M2 of six LDFs) - (M1/M2 of OLDF)). First row shows the ranges of 19 models. Second row shows the "M1Diff. & M2Diff." of the 14-vars model. "M2Diff." of LDF is 5.96%, and it is too bad. "M2Diff." of SVM1 is 1.43%, and it is worse than those of SVM4, LP and logistic. If we choose large value of penalty c such as 10000, the role of $\|\mathbf{b}\|^2/2$ in (5) is less meaning, and it may be similar to Revised LP-OLDF.

Only one 'M1Diff.' of IPLP is -0.01%. This means that Revised IPLP-OLDF is not free from problem 1 because M1 of Revised IP-OLDF is the minimum M1 among all LDFs. But ten 'M2Diff.' of IPLP are less than zero. Although some results may be caused by problem 1, other results may show that some M2 of Revised IPLP-OLDF are less than those of Revised IP-OLDF.

B. 100-fold cross validation of Iris data

Iris data [1] consisted of 100 cases with 4-features. Table II shows the results of 15 models by 100-fold cross validation. First column of OLDF shows the all possible combinations of features from 4-vars model (X1, X2, X3, X4) to 1-var model (X1). M1 of the full model is always minimum because M1 of (q+1)-features is always less than equal M1 of q-features theoretically. Although M2 of full model is minimum and it is 2.55, this is no guarantee theoretical. We consider the model with minimum M2 of Revised IP-OLDF as best model. Therefore, we can compere seven LDFs on this full model.

If we focus on 'M2Diff.' of the full model, those of SVM4, SVM1, LP, IPLP, Logistic and LDF are 0.46, 0.46, 0.4, 0.17, 0.39 and 0.64% worse than Revised IP-OLDF in the second row of SVM4, SVM1, LP, IPLP, Logistic and LDF. Results of six LDFs are not so bad because this data is very famous evaluation data of Fisher's LDF that satisfy the Fisher's assumption. Fisher chosen the best data for the validation of Fisher's LDF. Six maximum values of 'M2Diff.' are almost better than those 'M1Diff.'. This may imply that Revised IP-OLDF over-fit the training sample and the mean of error rates in the validation samples are worse than the training samples.

One 'M1Diff.' of LP and two 'M1Diff.' of IPLP are minus. This means that Revised LP-OLDF and Revised IPLP-OLDF are not free from problem 1 because M1 of Revised IP-OLDF is the minimum value among all LDFs. But, several M2 of Revised IP-OLDF are worse than others. Although some

 ISBN: 978-1-61208-358-2

results depend on the unresolved problem, other results may be caused by overestimate of Revise IP-OLDF.

TABLE II. IRIS DATA

OLDF	M1	M2
1,2,3,4	0.46	2.55
2,3,4	0.82	2.96
1,3,4	1.30	3.51
1,2,4	2.49	5.12
1,2,3	1.57	3.63
3,4	2.46	4.42
2,4	3.58	5.73
1,4	4.18	5.59
2,3	4.42	6.97
1,3	2.86	4.78
1,2	22.76	27.41
4	5.39	6.16
3	6.03	7.29
2	36.03	39.01
1	25.85	28.34
	M1Diff.	M2Diff.
SVM4	[0.58, 4.85]	[-0.67(3), 1.52]
	0.58	0.46
SVM1	[0.58, 4.85]	[-0.67(3), 1.52]
	0.58	0.46
LP	[-0.38(1), 3.63]	[-1.49(4), 1.43]
	0.47	0.40
IPLP	[-0.02(2), 0.07]	[-0.07(9), 0.17]
	0.01	0.17
Logistic	[0.71. 4.8]	[-1.11(3), 1.58]
	0.71	0.39
LDF	[0.51, 4.78]	[-0.89(3), 2.49]
	2.03	0.64

C. 100-fold cross validation of Swiss Bank data

Swiss bank note data consisted of two kinds of bills: 100 genuine and 100 counterfeit bills. There were six features. A total of 63 (=2^6-1) models were investigated. We had better considered about two types of discriminations: 16 linearly separable models and other 47 models. Table Ⅲ shows only 16 linear separable models in the training sample.

Only 4 M2s of Revised IP-OLDF are zero. In this case, we had better chosen the minimum number of features such as (1, 2, 4, 6). We compare Revised IP-OLDF with six LDFs in this 4-vars model. All models of six LDFs have the minimum M2. Those values are 0.38, 0.52, 0.27, 0.41, 0.38 and 0.47%, respectively. The results of six LDFs are not so bad. But, all M1s of SVM1 and Fisher's LDF are not zero. This means that both LDFs cannot recognize linear separable data, nevertheless this data may satisfy Fisher's assumption because genuine/counterfeit bills are industry products. And the results of H-SVM are as same as SVM4.

TABLE III. SWISS BANK NOTE DATA

OLDF	M1	M2
1,2,3,4,5,6	0.00	0.00
2,3,4,5,6	0.00	0.24
1,2,4,5,6	0.00	0.00
1,2,3,4,6	0.00	0.00
1,2,3,5,6	0.00	0.10
2,4,5,6	0.00	0.21
2,3,4,6	0.00	0.16
1,2,4,6	0.00	0.00
2,3,5,6	0.00	0.03
1,2,3,6	0.00	0.08
1,2,5,6	0.00	0.10
2,4,6	0.00	0.15
2,3,6	0.00	0.02
2.5,6	0.00	0.02
1,2,6	0.00	0.03
2,6	0.00	0.01
	M1Diff.	M2Diff.
SVM4	0	[0.22, 0.68]
/ HSVM	0	0.38
SVM1	[0.24, 0.57]	[1, 2.5]
	0.27	0.52
LP	0	[0.27, 0.75]
	0	0.27
IPLP	0	[0.25, 0.75]
	0	0.41
Logistic	0	[0.22, 0.65]
	0	0.38
LDF	[0.44, 0.95]	[0.25, 1.02]
	0.44	0.47

D. 100-fold cross validation of Student data

The student data consists of two groups: 25 students who pass the exam and 15 students who fail. There were 3 features: X1 was the hours of study per day; X2 was spending money per month; X3 was number of days drinking per week. If we analyze 2-features (X1, X2) by IP-OLDF, IP-OLDF chosen X2=5 as the discriminant hyper-plane. Four pass students and four fail students spent 50,000 yen/month. These eight students are on the discriminant hyper-plane. Only three fail students who spent less than 50,000 yen are misclassified by IP-OLDF. Revised IP-OLDF finds three true MNMs are 5 using by LINGO k-best option [21].

We examine 7 different models of student data by 100-fold cross validation in the Table Ⅳ. M1 decreases monotonously from 1-var to 3-vars. There are 6 passes such as : 1→(1,2)/(1,3)→(1,2,3), 2→(1, 2)/(2, 3) →(1, 2, 3), 3→

 ISBN: 978-1-61208-358-2

(1,3)/(2, 3) →(1,2,3). We compare Revised IP-OLDF with six LDFs by the model (2, 3). 'M2Diff.' of LDF, logistic, SVM4, SVM1 and LP are 7.19, 5.88, 4.1, 4.1 and 3.23%. These are very bad. 'M2Diff.' of Revised IPLP-OLDF is -0.2%. This may be the defect of Revised IP-OLDF.

TABLE IV. STUDENT DATA

OLDF	M1	M2
1,2,3	5.70	12.78
1,2	9.18	15.15
1,3	10.30	15.45
2,3	7.45	9.05
1	16.43	19.10
2	14.68	17.63
3	17.75	21.23
	M1Diff.	M2Diff.
SVM4	[1.2, 4.15]	[-0.7(2), 4.1]
	3.2	4.1
SVM1	[1.2, 4.15]	[-0.7(2), 4.1]
	3.2	4.1
LP	[-0.29(3), 3.35]	[-4.53(3), 3.23]
	2.7	3.23
IPLP	[0, 0.2]	[-0.75(5), .25]
0	0	-0.2
Logistic	[0.83, 5.43]	[-1.45(3), 5.88]
	4.8	5.88
LDF	[2.55, 6.55]	[-1.15(1), 7.19]
	6.03	7.19

V. CONCLUSION

Many statisticians believe that MNM criterion is foolish criterion because it over-fit for the training sample and it may overestimate the validation sample. On the contrary, generalization ability of LDF is best because it follows the normal distribution without examination by real data. In our paper, this claim may be wrong. In near future, this will be confirmed by the discrimination of the linear separable data using the pass/fail determination. In addition, the mean error rates of Fisher's LDF are higher than other LDFs. Past important researches using LDF should be reviewed, especially in the medical diagnosis. K-fold cross validation is very useful, compared with the leave-one-out method [5].

ACKNOWLEDGMENT

My research was achieved by LINGO of LINDO Systems Inc., and JMP of SAS Institute Inc.

REFERENCES

[1] A. Edgar, "The irises of the Gaspe Peninsula," Bulltin of the American Iris Society, vol. 59, pp. 2-5, 1945.

[2] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7, pp. 179–188, 1936.

[3] B. Flury and H. Rieduyl, Multivariate Statistics: A Practical Approach. Cambridge University Press, 1988.

[4] T. Ibaraki and S. Muroga, "Adaptive linear classifier by linear programming," IEEE transaction On systems science and cybernetics, SSC-6, pp. 53-62, 1970.

[5] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," Technometrics vol. 10, pp.1-11, 1968.

[6] J. P. Sall, SAS Regression Applications. SAS Institute Inc. 1981.

[7] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, 3rd ed. SAS Institute Inc. 2004.

[8] L. Schrage, LINDO－An Optimization Modeling System－. The Scientific Press. 1991.

[9] L. Schrage, Optimization Modeling with LINGO. LINDO Systems Inc. 2006.

[10] S. Shinmura and A. Miyake, "Optimal linear discriminant functions and their application," COMPSAC79, pp. 167-172, 1979.

[11] A. Miyake and S. Shinmura, "An Algorithm for the Optimal Linear Discriminant Function and its Application," Japanese Journal of Medical Electronics and Biological Engineering, Vol 18/1, pp. 15-20, Feb. 1980.

[12] S. Shinmura, "Optimal Linear Discrimrnant Functions using Mathematical Programming," Journal of the Japanese Society of Computer Statistics, vol. 11/2, pp. 89-101, 1998.

[13] S. Shinmura, "A new algorithm of the linear discriminant function using integer programming," New Trends in Probability and Statistics, vol. 5, pp. 133-142. 2000.

[14] S. Shinmura, The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing. 2010.

[15] S. Shinmura, "Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis -," 2011 ISI CD-ROM, pp.1-6. 2011.

[16] S. Shinmura, "Problems of Discriminant Analysis by Mark Sense Test Data," Japanese Society of Applied Statistics, vol. 40/3, pp. 157-172, 2011.

[17] S. Shinmura, "Evaluation of Optimal Linear Discriminant Function by 100-fold cross-validation," 2013 ISI CD-ROM, pp.1-6, 2013.

[18] S. Shinmura, "Evaluation of Revised IP-OLDF with S-SVM, LDF and logistic regression by K-fold cross-validation," IEICE Technical Report IBISML 2013-44 (2013-11), pp.61-68.

[19] S. Shinmura, "End of Discriminant Functions based on Variance-Covariance Matrices," ICORE2014, pp. 5-16. 2014.

[20] S. Shinmura, "Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP," Statistics, Optimization and Information Computing, vol. 2, June 2014, pp 114-129.

[21] S. Shinmura, "Three Serious Problems and New Facts of the Discriminant Analysis" Operations Research snd Enterprise Systems, ICORES 2014, Revised Selected Papers, in Press.

[22] A. Stam, "Nontraditinal approaches to statistical classification: Some perspectives on Lp-norm methods," Annals of Operations Research, vol. 74, pp. 1-36, 1997.

[23] V. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag. 1995.

[24] R. Warmack and R. C. Gonzalez, "An Algorithm for the optimal solution of linear inequalities and its application to pattern recognition," IEEE Trans. Computers, pp. 1065-1075, 1973.

 ISBN: 978-1-61208-358-2