A Unification of Determined and Probabilistic Methods in Pattern Recognition

Geo Kunev[1], Georgi Varbanov[2] and Christo Nenov[3]

***Abstract* – Main goal of the present task is to find possibilities for achieving uniform rules for pattern recognition, regardless of significant difference in initial positions and different methods for achieving the end form of algorithms. We will try to present most used rules for decision making in aggregate of:**

- **uniform procedure for estimate of state, making a calculation of linear or quadratic form,**
- **comparing of results with some threshold value.**

We can apply this results in KDD (data mining) software.

***Keywords* – Classification, Bayes, Pattern recognition, KDD**

I. INTRODUCTION

By non-concave location of features by class's areas we search dividing function in full quadratic form or in a part:

$$g(x) = c_0 + \sum_{i=1}^{d} c_i x_i + \sum_{k=1}^{d}\sum_{j=1}^{d} c_{ij} x_i x_j (x, \bar{c}) \quad (1)$$

or:

$$g(x) = X^T AX + a^T X + a_0 \quad (2)$$

We search one-type pattern recognition rules for:

1. Algorithms, based on optimal statistical decision making theory,
2. Algorithms, based on dispersion of probabilistic recognition features,
3. Algorithms, based on minimizing of geometric mean recognition error.

II. OPTIMAL STATISTICAL DECISION MAKING

In this well known case [1], pattern recognition task is treat as common statistical task with predefined optimum criteria.

$$R(A/x) = M[I(x\hat{x}/x] \quad (3)$$

Most used criterion is a loss function $I(x\hat{x})$ for calculation of conditional mean risk, searching the best estimation [4]:

$$R(A) = M[R(A/x)] = M[I(x\hat{x})] \Rightarrow \min \quad (4)$$

Average risk by repeatedly recognition of *M* classes is:

$$R = \sum_{k=1}^{M}\sum_{m=1}^{M}\int_{X_M} c_{km} P_k . f(\bar{y}/x_k) dy \quad (5)$$

Decision rule for *M* classes *(j=1,M)* is:

$$\bar{y}_i \in X_j, \_if\ \_ - c_j . P(x_j) f(y/x_j) \Rightarrow \max \quad (6)$$

Decision rule for two classes is:

$$\bar{y}_i \in X_1, \_if\ \_\frac{f(y_i/x_1)}{f(y_i/x_2)} = \lambda > \lambda_0 = \frac{P_2(c_{21} - c_{22})}{P_1(c_{12} - c_{11})}$$
$$\bar{y}_i \in X_2, \_if\ \_\lambda < \lambda_0 \quad (7)$$

By normal feature dispersion on *M* classes:

$$\bar{y}_i \in X_1, \_if\ \_g_j(\bar{y}) = -\frac{1}{2} y^T \sum_j y + (\sum_j^{-1} \mu_j)^T -$$
$$-\frac{1}{2}\mu_j{}^T \sum_j^{-1} \mu_j - \frac{1}{2}\ln|\sum_j \mu_j| + \ln P(x_j) + \ln|c_j| = \quad (8)$$
$$= Y^T A_j Y + a_j{}^T y + a_{j0} \Rightarrow \max$$

By two classes and conditions:

$$P(x_1)=P(x_2),$$
$$(c_{12}- c_{11}= c_{21}- c_{22}), \quad (9a)$$
$$\Sigma_1=\Sigma_2,$$

we have linear recognition rule:

$$\bar{y}_i \in X_1, \_if\ \_y^T \sum^{-1} (\mu_1 - \mu_2)$$
$$-\frac{1}{2}(\mu_1 - \mu_2)^T \sum^{-1} (\mu_1 - \mu_2) < 0 \quad (9b)$$

]

Bayesian strategies compare projection of recognized observation (vector) on directions eq.(10), with the mean values vector on the same direction:

$$a = \sum^{-1} (\mu_1 - \mu_2) \text{ - by two classes}$$
$$a = \sum_j^{-1} \mu_j \text{ - by } M \text{ classes.} \quad (10)$$

[1]Geo Kunev is with the Faculty of Computer Sciences, TU Varna, Studentska 1, 9000 Varna, Bulgaria, E-mail: geo_qnew@hotmail.com

[1]Georgi Varbanov is with the Faculty of Computer Sciences, TU Varna, Studentska 1, 9000 Varna, Bulgaria, E-mail: varbanov@mbox.digsys.bg

[1]Christo Nenov is with the Faculty of Computer Sciences, TU Varna, Studentska 1, 9000 Varna, Bulgaria, E-mail: chr_nenov@gmail.com

III. Recognition Matrices Study

By limited a priory statistic we search recognition rules, based on sub-areas metrics for different classes. For two classes we can use Fisher's nonzero hypothesis criterion [2]:

$$F = \frac{S_\Omega{}^2}{S_R{}^2}, \qquad : \tag{11a}$$

where:
$S_\Omega{}^2$ - is dispersion estimation (by groups),
$S_R{}^2$ – is dispersion estimation (in group).

As multidimensional analysis we can use Mahalanobis distance between groups with means μ_i and μ_j and common covariance matrix *V:*

$$V^2=(\mu_i-\mu_j)^T V^{-1}(\mu_i-\mu_j) \tag{11b}$$

An estimation of dispersion in groups for *M* classes is:

$$S_w = \sum_{k=1}^{M} P(x_k) E\left\{(y-\mu_k)(y-\mu_k)^T / x_k\right\} = \\ = \sum_{k=1}^{M} P(x_k) \sum\nolimits_k \tag{12}$$

where:
$P(x_k)$ – every class a priory probability,
Σ_κ – classes covariance matrices,
y, μ_k – observations and means by classes vectors.

An estimation of dispersion by groups for *M* classes is:

$$S_B = \sum_{k=1}^{M} P(x_k)(\mu_k-\mu_0)(\mu_k-\mu_0)^T \tag{13}$$

An estimation based only on statistic:

$$S_{B\,k,m} = (\mu_k-\mu_m)(\mu_k-\mu_m)^T \tag{14}$$

We search best transformation in form [1]:

$$Z=a^T Y+a_0 \tag{15}$$

like:

$$J = \frac{|S_B{}^*|}{|S_W{}^*|} = \frac{\sum_{k=1}^{M} m_k(\mu_k{}^*-\mu_0{}^*)(\mu_k{}^*-\mu_0{}^*)^T}{\sum_{k=1}^{M}\sum_{z\in Z_k}(Z-\mu_0{}^*)(Z-\mu_0{}^*)^T} = \\ = \frac{|A^T(\mu_k-\mu_0)(\mu_k-\mu_0)^T A|}{|\sum_{y\in Y_k}(y-\mu_0)(y-\mu_0)^T|} = \frac{|A^T S_B A|}{|A^T S_W A|} \tag{16}$$

where:
y, μ_k, μ_0, S_B, S_W – are observations, means and dispersion matrixes in initial areas,
and Z, $\mu_k{}^*$, $\mu_0{}^*$, $S_B{}^*$, $S_W{}^*$ - the same after linear transformation.

By two classes and

$$a^T\Sigma_\kappa a = \sigma_\kappa{}^{*2}, \quad \text{и} \quad \mu_k{}^*,=a^T\mu_k + a_0 \tag{17}$$

we have:

$$J = \frac{(\mu_k{}^*-\mu_j{}^*)^2}{\sigma_k^2-\sigma_j^2} \tag{18}$$

and extremum conditions can be defined:

$$\frac{dJ}{da_i} = \frac{dJ}{d\sigma_k^{*2}}\cdot\frac{d\sigma_k^{*2}}{da_i} + \frac{dJ}{d\sigma_j^{*2}}\cdot\frac{d\sigma_j^{*2}}{da_i} + \\ + \frac{dJ}{d\mu_k^*}\cdot\frac{d\mu_k^*}{da_i} + \frac{dJ}{d\mu_j^*}\cdot\frac{d\mu_j^*}{da_i} = 0 \\ \frac{dJ}{da_0} = \frac{dJ}{d\sigma_k^{*2}}\cdot\frac{d\sigma_k^{*2}}{da_0} + \frac{dJ}{d\sigma_j^{*2}}\cdot\frac{d\sigma_j^{*2}}{da_0} + \\ + \frac{dJ}{d\mu_k^*}\cdot\frac{d\mu_k^*}{da_0} + \frac{dJ}{d\mu_j^*}\cdot\frac{d\mu_j^*}{da_0} = 0 \tag{19a}$$

where:

$$\frac{dJ}{d\sigma_k^{*2}} = -\frac{(\mu_k^*-\mu_j^*)^2}{\sigma_k^{*2}+\sigma_j^{*2}}; \frac{dJ}{d\sigma_j^{*2}} = -\frac{(\mu_k^*-\mu_j^*)^2}{\sigma_k^{*2}+\sigma_j^{*2}} \\ \frac{dJ}{d\mu_k^*} = -\frac{dJ}{d\mu_j^*} = -\frac{2(\mu_k^*-\mu_j^*)}{\sigma_k^{*2}+\sigma_j^{*2}} \\ \frac{d\sigma_k^{*2}}{da_i} = -2\Sigma_k a; \frac{d\mu_k^*}{da_i} = \mu_k; \frac{d\sigma_j^{*2}}{da_i} = -2\Sigma_j a; \frac{d\mu_i^*}{da_i} = \mu_j \\ \frac{d\sigma_k^2}{da_0} = 0; \frac{d\mu_k^*}{da_0} = 1; \frac{d\sigma_j^2}{da_0} = 0; \frac{d\mu_j^*}{da_0} = 1; \tag{19б}$$

then:

$$2\left[\frac{(\mu_k^*-\mu_j^*)^2}{\sigma_k^{*2}+\sigma_j^{*2}}\right]\left(\frac{1}{2}\Sigma_k + \frac{1}{2}\Sigma_j\right)a = \mu_k - \mu_j \tag{20}$$

or:

$$a = \left(\frac{1}{2}\Sigma_k + \frac{1}{2}\Sigma_j\right)^{-1}(\mu_k-\mu_j) \tag{21}$$

By equal covariance matrix:

$$a = \Sigma^{-1}(\mu_k-\mu_j) \tag{22}$$

In case of a linear transformation by two classes with equal covariance matrix, the decision is known as Fisher's linear discriminant [2],[3]:

$$S_B.\ W_i = \lambda.\ S_w.\ W_i \tag{23}$$

If S_w is a non-degenerate matrix:

$$S_B.\ W_i = \lambda.\ S_w.\ W_i \tag{24}$$

By dichotomy of two classes and S_B=S_B‘

$$W = S_w^{-1}.(\mu_\kappa - \mu_i) \quad (25)$$

and this is same as eq.(22).

In multidimensional case:

$$a_\kappa = S_{W\kappa}^{-1}.(\mu_\kappa - \mu_0) \quad (26)$$

After centering:

$$a_\kappa = S_{W\kappa}^{-1}.\mu_\kappa \quad (27)$$

IV. Minimizing of Geometric Mean Recognition Error

This group of methods has goal to develop procedures for finding dividing functions coefficients.

After:

$$y = \begin{vmatrix} 1 \\ x \\ \varphi(x) \end{vmatrix}; a = \begin{vmatrix} c_0 \\ c_i \\ c_{ij} \end{vmatrix} \quad \text{or:} \quad y' = \begin{vmatrix} x \\ \varphi(x) \end{vmatrix}; a' = \begin{vmatrix} c_i \\ c_{ij} \end{vmatrix} \quad (28)$$

we can have eq. (1) in form of linear dividing function [3]:

$$g(x) = a^T Y \quad (29)$$

or:
$$g(x) = a'^T Y' + c_0 \quad (30)$$

Presuming linear class divider:

$$a^T(y_m - y_n) = 0 \quad (31)$$

vector a^T is normal to every vector on dividing surface.

So, every pair of classes (W_i, W_j):

$$a^T y_m > 0, \ if \ y_m \in W_i$$
$$a^T y_m < 0, \ if \ y_m \in W_i \quad (32)$$

After transformation:

$$y_m = -y_m;$$
$$\forall y_m \in W_j,$$

decision rule become:

$$a^T y_m < 0; \ \forall (y_m \in W_i, W_j) \quad (33)$$

Having a decision area supposes non-single decision, therefore it can be used additional limitations, for example, searching minimum weight vector:

$$a^T y_m \geq b > 0; \ \forall (y_m \in W_i, W_j) \quad (34)$$

In this case a learning task is to find weight vector ā, matching the best possible equation in form:

$$Ya = b; \ Y[n,d]; \ b[n,1]; \ n > d \quad (35)$$

This system in common case has no exact decision, therefore after setting error vector:

$$e = Ya - b \quad (36)$$

the task can be treated as classical case of minimizing of:

$$J(a) = \| Ya - b \|^2 = \sum_{i=1}^{n} (a^T y_i - b_i)^2 \quad (37)$$

or:
$$\nabla J(a) = 2Y^T(Ya - b) = \sum_{i=1}^{n} 2(a^T y_i - b_i) y_i = 0 \quad (38)$$

and:

$$Y^T Ya = Y^T b \quad (39)$$

If the matrix $Y^T Y$ (d.d) is non-degenerate, the vector is:

$$a = (Y^T Y)^{-1} Y^T b = Y^{\#} \quad (40)$$

where $Y^{\#} = (Y^T Y)^{-1} Y^T$; $[d_x n]$ is well known in theory pseudo-reversed matrix.

In this procedure it can be problems with pseudo-reversing. There are different concrete schemes for applying the least squares method and the best is the Ho-Kashap procedure. This procedure moves in steps to the minimum of eq.(37), keeping gradient directions:

$$\nabla_a J = 2Y^T(Ya - b) \quad (41)$$
$$\nabla_b J = -2(Ya - b) \quad (42)$$

It begins with statistic, grouped in two classes as a generalized normalized observation matrix in form:

$$y = \begin{bmatrix} u_i & x_i \\ -u_j & -x_j \end{bmatrix} \quad (43)$$

where:

x_i and x_j are observations of W_i и W_j, classes, and vector-columns u_i include n_i threshold values, equal to one.

If weight and limitation vectors are:

$$A = \begin{bmatrix} a_0 \\ a \end{bmatrix}; b = \begin{vmatrix} \frac{n}{n_i} & u_i \\ \frac{n}{n_j} & u_j \end{vmatrix}; n = n_1 + n_2 \quad (44)$$

and according to eq.(39):

$$\begin{bmatrix} u_i^T & -u_j^T \\ x_i^T & -x_j^T \end{bmatrix} \begin{bmatrix} u_i & x_i \\ -u_j & -x_j \end{bmatrix} \begin{bmatrix} a_0 \\ a \end{bmatrix} =$$

$$= \begin{bmatrix} u_i^T & -u_j^T \\ x_i^T & -x_j^T \end{bmatrix} \begin{bmatrix} \frac{n}{n_i} & u_i \\ \frac{n}{n_j} & u_j \end{bmatrix} \quad (45)$$

After including:

$$\frac{1}{n_i} \sum_{x \in x_i} X = \mu_i$$

$$\sum_{i-1}^{2} \sum_{x \in x_i} (X - \mu_i)(X - \mu_i)^T = S_W \quad (46)$$

we have:

$$a_0 = -\mu^T a$$

$$\left[\frac{1}{n} S_W + \frac{n_i n_j}{n^2} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \right] a = \mu_i - \mu_j \quad (47)$$

and:

$$a \frac{1}{n} S_W = (\mu_i - \mu_j) - a \frac{n_i n_j}{n^2} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (48a)$$

As vector direction:

$$a \frac{n_i n_j}{n^2} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (48b)$$

by every a coincide with vector direction $(\mu_i - \mu_j)$:

$$a \frac{1}{n} S_W = \alpha (\mu_i - \mu_j) \quad (49)$$

whence:

$$a = n\alpha . S_W^{-1} (\mu_i - \mu_j) \quad (50)$$

After excluding inessential scalar coefficient $n\alpha$ we have direction that minimizes sum of squares in form:

$$a = S_W^{-1} (\mu_i - \mu_j) \quad (51)$$

V. Conclusion

Comparison of eq. (21), (22), (25), (26), (27) gives the conditions, where recognition procedures, based on optimal linear feature area transformation – eq.(15), and linear Fisher's discriminant has got equal mathematical sense.

In both cases we have projection of observation vector on direction $S_w^{-1}.(\mu_\kappa - \mu_i)$, and comparing the result with some threshold value. So, according to eq.(8), (9), (10), we can assert about algorithmic equality of recognition procedures.

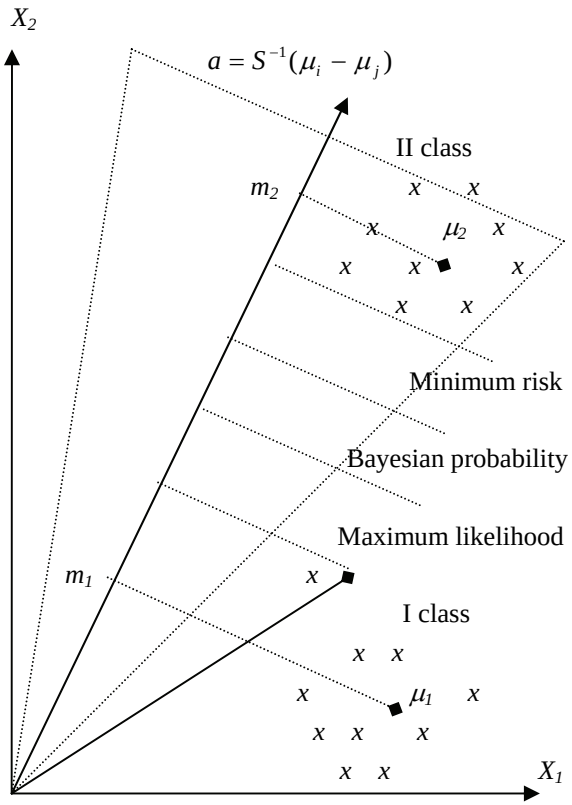

Fig. 1 Classes and classification methods

After comparison of eq.(10), (22), (26), (51) we can see relation between linear parametric and non-parametric, probabilistic and determined methods, and also equivalence conditions of respective recognition procedures.

The fact, that least squares procedure and maximum likelihood procedure approximate by probability to the linear Fisher's discriminant shows that we have reason to speak not about different, but asimptotic approximate procedures with common computing scheme. The computed value (projection) is compared with threshold valued, which define optimum as decision making in sense of minimizing of risk, maximum confitional or a posteriori probability (Fig. 1).

Acknowledgement

This paper is a result of a study, with the kind guidance of professor dr Asen Nedev in TU Varna.

References

[1] V. Vapnik, Statistical Learning Theory, J.Wiley, N.Y. 1998
[2] R.A. Fisher, Contibushions to Mathematical Statistics. J.Wiley, N.Y. 1994
[3] Fukunaga. K, Introduction to statistical pattern recognition. Academ press, N.Y, 1990
[4] A. Nedev, K. Tenekedjiev, Technical diagnostics and pattern recognition, TU Varna, 1997