Mathematica Balkanica

Mathematical Society of South-Eastern Europe
A quarterly published by
the Bulgarian Academy of Sciences – National Committee for Mathematics

Mathematica Balkanica

New Series Vol.10, 1996, Fasc.1

g–Divergences on Image Measure

V. Boscaiu

Presented by P. Kenderov

A necessary and sufficient condition for a variable transformation to preserve a multi–class discriminant–information–measure is defined.

1. Introduction

The transformation of variables seldom appears in pattern–recognition problems as a preliminary step of discrimination between classes. The aim should be feature–selection, feature–extraction, obtaining a linear decision function, etc. This kind of problems were extensively studied: J. M. Weiner and O. J. Dunn (1966), M. A. Morgan (1975), J. W. Van Ness and C. Simpson (1976), W. Schaafsma (1982), etc.

A class of transformations which preserve the discriminant information will be studdied below. The concept of discriminant information will be generally defined by some properties especially selected with a view of describing the capacity of a variable to discriminante between two or more classes.

A straightforward description of discriminant capacity should be correlated with the best possible (minimal) probability of misclassification in the classes set. But: 1) the best classification procedure is not known a priory; 2) even for a given procedure, the probability of misclassification is not known if (as usual) the decision is only based on a training sample and it must be estimated (bibliography on subject: G. Toussaint (1974); B. Efron's contributions must be specially mentioned). Thus, the misclassification probability should be replaced by a measure of discriminant information which must have at most two features: 1) to be estimated before discrimination; 2) maximal discriminant information of a probability structure must imply a near–optimal (near–maximal) probability of correct classification.

A well–known example of the informational equivalence of the discriminant information and the probability of misclassification is the classification in two equiprobable, normally distributed classes, $N(\mu_i, \sum)$, $i = 1, 2$. In the case, T. W. Anderson (1958) proved that the minimal misclassification probability is $2\Phi(-\triangle/2)$ where Φ is the Laplace function and $\triangle$ is the Mahalanobis distance of two classes, $\triangle = (\mu_1 - \mu_2)' \sum^{-1}(\mu_1 - \mu_2)$. On the other hand, S. Kullback (1968) proved that the divergence J of the normally distributed classes is $J = \triangle$. Therefore in this case, the divergence – as a measure of discriminant information – completely characterises the misclassification probability. But this is the very special case.

The study of some real and Monte–Carlo–simulated data supports the conclusion that the replacement of minimal misclassification probability criterion may be made in many classification models. T. L. Boullion, P. L. Odell, B. S. Duran (1975) concluded that the Kullback divergence and the Mahalanobis distance offer useful aproaches in the problem of variable selection, especially for small samples.

Generally, S. Watanabe (1981) considers many classical problems of pattern recognition in the framework of the information theory as an attempt of finding minimal entropy. Below, entropy will be replaced by the class of the g–divergence function, which are defined for two or more classes. F. Liese and I. Vajda (1987) presented a related approach for two classes. Other approaches may be found in G. Zbaganu (1993).

2. Statement of the problem

The following definitions will be used.

a. The classes set $\pi_1, \ldots, \pi_n$ is a partition of the set π.

b. X is a finite set, $card(X) = c$. Alternately, X will also be used to denote a random variable defined on π.

c. $p : (X, \mathcal{P}(X)) \to [0, 1]$ is a probability.

d. $p_i : (X, \mathcal{P}(X)) \to [0, 1]$ is the conditional probability corresponding to π_i, $i = \overline{1, n}$. The p_i probability can be interpreted as a vector: $p_i =: (p_{i1}, \ldots, p_{ic})$, where $p_{ij} =: p_i(x_j) = p(x_j | \pi_i)$, $x_j \in X$.

e. The g–divergence function $J : ([0, 1]^c)^n \to [0, \infty)$ is defined by (1) – (3) as follows:

$$(1) \qquad J(p_1, \ldots, p_n) =: J(X) =: \sum_{x \in X} g(p_1(x), \ldots, p_n(x)) = \sum_{k=1}^{c} g(p_{1k}, \ldots, p_{nk}).$$

(2) $\quad g : [0, 1]^n \to [0, \infty)$ is a convex function.

(3) $\quad g(rt_1, \ldots, rt_n) = rg(t_1, \ldots, t_n) \qquad \forall r \geq 0.$

f. q_i is the conditional probability induced on Y by $T : X \to Y$ and p_i, $i = 1, n$ (the set $S(y)$ will be only defined for editing reasons):

$$q_i(y) =: p_i(T^{-1}(Y)) = \sum_{x \in S(y)} p_i(x) \tag{4}$$

$$S(y) =: T^{-1}(y) =: \{x \in X | T(x) = y\} \tag{5}$$

g. **The problem definition** is: determination of the transformations–class preserving the discriminant information, namely the g–divergence value.

Kullback (1968) completely solved this problem for two classes and Kullback divergence: the necessary and sufficient condition for preserving the J–value in the T–image space is

$$_1(x)/f_2(x) = h_1(x)/h_2(x) \qquad \mu\text{-}a.e. \tag{6}$$

where f_i is the conditional probability density function of X with respect to the measure μ and h_i is the density of $T \circ X$ with respect to image measure.

In the finite case, $card(X) = c < \infty$, the above mentioned result will be formulated in a more direct manner for g–divergence family, including also global divergences for more then two classes.

Examining the relations (4) and (5), it may be noted that the probability q_i is depending on a partition of the X set. In this perspective, the transformation problem is meant to determine a partition of the X set, so that the g–divergence of the new probability structure remains invariable.

3. Main result

Theorem 1. *X is a finite set; $T : X \to Y$ is a function; $p_i : (X, \mathcal{P}(X)) \to [0, 1]$ is the probability and q_i is the probability on $T(X) =: Im(T)$ generated by the relation (4), $i = 1, n$; $n > 1$; J is a g–divergence (J has the properties (1) – (3)).*

The followin statments hold:

I. $J(X) \geq J(T(X))$.

II. If T is a bijective function, then

$$(X) = J(T(X)). \tag{7}$$

III. If $p_1(x) \neq 0$ for all $x \in X$ and the function $T_0 : X \to Y$ is

$$_0(x) =: t(p_2(x)/p_1(x), \ldots, p_n(x)/p_1(x)), \tag{8}$$

$t : [0,\infty)^{n-1} \to Y$ beeing an injective function, then

$$(X) = J(T_0(X)). \tag{9}$$

IV. If $g(1,\cdot,\ldots,\cdot)$ is a strictly convex function defined on R^{n-1} and $J(X) = J(T(X))$ then

$$ardT(X) \geq cardT_0(X) \tag{10}$$

Also, the following implication holds, $\forall u, v \in X$:

$$(u) = T(v) \Rightarrow p_i(u)/p_1(u) = p_i(v)/p_1(v), \qquad i = 2, n. \tag{11}$$

The proof of Theorem 1 will use Lemma 2.

Lemma 2. *The function $g : (0,\infty) \times [0,\infty)^{n-1} \to R$ is defined by*

$$(v) =: g(v_1,\ldots,v_n) =: v_1 h(v_2/v_1,\ldots,v_n/v_1) \tag{12}$$

where $h : R^{n-1} \to R$ is a convex function and $v =: (v_1,\ldots,v_n)$.

I. g is a convex function.

II. If h is strictly convex, $v_1 z_1 > 0$ and $z_i > 0$ for some $i > 1$, then the relation (13) implies the relation (14), where:

$$(rv + (1-r)z) = rg(v) + (1-r)g(z), \qquad r \in (0,1); \tag{13}$$

$$_1/z_1 = v_2/z_2 = \ldots = v_n/z_n \tag{14}$$

(convention: if (14), then for $i \geq 2$, $z_i = 0$ iff $v_i = 0$).

III. If h is a strictly convex function and

$$g((v^1 + \cdots + v^s)/s) = g(v^1) + \cdots + g(v^s), \qquad s > 0 \tag{15}$$

then (using the above convention):

$$_{1i}/v_{11} = \ldots = v_{si}/v_{s1}, \qquad i = 2, n, \qquad v^j =: (v_{j1},\ldots,v_{jn}) \in R^n. \tag{16}$$

P r o o f of Lemma 2.

I. If $\mu =: rv_1/(rv_1 + (1-r)z_1)$, we have for $i = 2, n$:

$$A_i =: (rv_i + (1-r)z_i)/(rv_1 + (1-r)z_1) = \mu(v_i/v_1) + (1-\mu)(z_i/z_1),$$

$$g(rv+(1-r)z) = g(rv_1+(1-r)z_1, \ldots, rv_n+(1-r)z_n) = (rv_1+(1-r)z_1)h(A_2, \ldots, A_n).$$

Now, the convexity of h–function will be used:

$$\begin{aligned} &g(rv + (1-r)z) = \\ &= (rv_1 + (1-r)z_1)h(\mu(v_2/v_1, \ldots, v_n/v_1) + (1-\mu)(z_2/z_1, \ldots, z_n/z_1)) \leq \\ &\leq (rv_1 + (1-r)z_1)[\mu(h(v_2/v_1, \ldots, v_n/v_1) + (1-\mu)h(z_2/z_1, \ldots, z_n/z_1)] = \\ &= rv_1(h(v_2/v_1, \ldots, v_n/v_1) + (1-r)z_1 h(z_2/z_1, \ldots, z_n/z_1)] = rg(v) + (1-r)g(z). \end{aligned}$$

(The last equality was derived using (12).)

II. The equality (13) implies that the last inequality from part I. of the proof must be equality. Because of strict–convexity of h–function, we have $v_i/v_1 = z_i/z_1$, $i = 2, n$ and the relation (14) hold for all "i" so that $z_i \neq 0$ (such "i" exists, by hypothesis).

III. Using relation (12) in the relation (15), we obtain

$$(17) \quad (v_{11} + \cdots + v_{s1})h(B_2, \ldots, B_n) = \sum_{j=1}^{s} g(v^j) = \sum_{j=1}^{s} v_{j1}h(v_{j2}/v_{j1}, \ldots, v_{jn}/v_{j1}),$$

where $B_i =: (v_{1i} + \cdots + v_{si})/(v_{11} + \cdots + v_{s1})$, $i = 2, n$. Defining $\lambda^j =: v_{j1}/(v_{11} + \cdots + v_{s1})$ for $j = 1, s$, we have the relation $B_i = \sum_{j=1}^{s} \lambda^j (v_{ji}/v_{j1})$. Using the λ's coeficients, relation (17) becomes after dividing by $v_{11} + \cdots + v_{s1}$:

$$h\Big(\sum_{j=1}^{s} \lambda^j (v_{j2}/v_{j1}), \ldots, \sum_{j=1}^{s} \lambda^j (v_{jn}/v_{j1})\Big) = \sum_{j=1}^{s} \lambda^j h(v_{j2}/v_{j1}, \ldots, v_{jn}/v_{j1}).$$

Due to the strict convexity of h, the last relation is possibible if and only if (16) holds. ■

P r o o f of Theorem 1.

Using (1), the following relations holds:

$$(18) \qquad (T(X)) = \sum_{y \in T(X)} g(q_1(y), \ldots, q_n(y)).$$

II. If T is a bijective function and $T(x) = y$, then $x = T^{-1}(y)$ and (4) means $q_i(y) = p_i(x)$. The relation (18) becomes

$$J(T(X)) = \sum_{x \in X} g(p_1(x), \dots, p_n(x)) = J(X).$$

III. The relation (3), (4), (5), (18) and $E(y) =: \sum_{x\in X} p_1(x)$ imply:

$$J(q_1, \dots, q_n) = \sum_{y \in T(X)} E(y) g(1, E^{-1}(y) \sum_{x \in S(y)} p_2(x), \dots, E^{-1}(y) \sum_{x \in S(y)} p_n(x)).$$

Considering (8) and the injectivity of t–function we have

$$\begin{aligned} S(y) &=: T_0^{-1}(y) = \{x \in X | (p_2(x)/p_1(x), \dots, p_n(x)/p_1(x)) = t^{-1}(y)\} = \\ &= \{x \in X | p_i(x)/p_1(x) = pr_i(t^{-1}(y)) =: k_i(y), \quad i = 2, n\} \end{aligned}$$

(pr_i is a projection function; because t is an injective function and $y \in T_0(X)$, k_i is function, $k_i : T_0(X) \to R$).

Considering the last relation for $S(y)$, it is evident that

$$(19) \qquad {}_i(x)/p_1(x) = k_i(y) \qquad \forall x \in S(y), \qquad i = 2, n.$$

Therefore:

$$q_i(y) = \sum_{x \in S(y)} p_i(x) = k_i(x) \sum_{x \in S(y)} p_1(x) = k_i(y) E(y), \qquad i = 2, n;$$

$$\begin{aligned} J(T_0(X)) &= J(q_1, \dots, q_n) = \sum_{y \in T_0(X)} E(y) g(1, k_2(y), \dots, k_n(y)) = \\ &= \sum_{y \in T_0(X)} \sum_{x \in S(y)} p_1(x) g(1, k_2(y), \dots, k_n(y)) = \\ &= \sum_{y \in T_0(X)} \sum_{x \in S(y)} g(p_1(x), p_1(x) k_2(y), \dots, p_1(x) k_n(y)) = \\ &= \sum_{y \in T_0(X)} \sum_{x \in S(y)} g(p_1(x), p_2(x), \dots, p_n(x)(y)) = J(p_1, \dots, p_n). \end{aligned}$$

(The last but one equality derives from (19).)

I. If $y \in T(X)$ and $m =: card(T^{-1}(y))$, (2) and (3) imply:

$$(20)(1/m \sum_{x \in S(y)} p_1(x), \dots, 1/m \sum_{x \in S(y)} p_n(x)) \le 1/m \sum_{x \in S(y)} g(p_1(x), \dots, p_n(x));$$

(21) $$(\sum_{x\in X} p_1(x),\ldots,\sum_{x\in X} p_n(x)) \le \sum_{x\in X} g(p_1(x),\ldots,p_n(x)).$$

Adding the inequalities (21) for all $y \in T(X)$ and taking into account (4) and (1), the required statement is reached:

$$\begin{aligned} J(q_1,\ldots,q_n) &= \\ &= \sum_{y\in T(X)} g(q_1(y),\ldots,q_n(y)) = \sum_{y\in T(X)} g(\sum_{x\in S(y)} p_1(x),\ldots,\sum_{x\in S(y)} p_n(x)) \le \\ &\le \sum_{y\in T(X)}\sum_{x\in S(y)} g(p_1(x),\ldots,p_n(x)) = \sum_{x\in X} g(p_1(x),\ldots,p_n(x)) = \\ &= J(p_1,\ldots,p_n). \end{aligned}$$

IV. If $J(X) = J(T(X))$, then the relations (21) and (20) must be equalities for all $y \in T(X)$. But a relation (20) with "=" has form (15) if $m > 1$. Also g verifies (3) and therefore

$$g(p_1(x),\ldots,p_n(x)) = p_1(x)g(1,p_2(x)/p_1(x),\ldots,p_n(x)/p_1(x)).$$

Moreover, by hypothesis, $g(1,\cdot,\ldots,\cdot)$ is strictly convex. Now, Lemma 2 III, applied for $h = g(1,\cdot,\ldots,\cdot)$, guarantees for each $y \in T(X)$ the following type – (16) statement: there exists $k_j(y)$, a unic number depending on y, such that

(22) $$_j(x)/p_1(x) = k_j(y), \qquad j = 2,n \qquad \forall x \in T^{-1}(y).$$

Namely, for each $y \in T(X)$ and all $x \in T^{-1}(y)$:

(23) $$p_2(x)/p_1(x),\ldots,p_n(x)/p_1(x)) = (k_2(y),\ldots,k_n(y)).$$

The relation (23) is a definition of a function $k =: (k_2,\ldots,k_n)$.

$$k : T(X) \to U, \qquad U = \{(p_2(x)/p_1(x),\ldots,p_n(x)/p_1(x))|x \in X\}.$$

The relation (23) also show that k is surjective, therefore

(24) $$ard(T(X)) \ge card(U).$$

■

Considering the definition (8) of T_0, we have $t^{-1}(T_0(X)) = U$. The injectivity of t implies $card(T_0(X)) = card(U)$ and (24) becomes (10). The implication (11) can be immediately reached from (22).

4. Comments

a) Statements I and II of Theorem 1 are natural: 1) the effect of any variable– transformation cannot bring about an increase of information; 2) a bijective transformation preserves information.

b) Statement III of Theorem 1 represents sufficient conditions for preserving the information. Relation (11) of the statement IV represents a necessary conditions.

c) Relation (10) describes the minimality condition of the transformation T_0 defined by (8): T_0 is the function which preserves information and has the least numerous set of values.

d) The structure of T_0 is explicitly described, unlike the above mentioned Kullback result (relation (6)) which involves the definitions of the c.p.d.f.'s with respect to an image probability which is not explicitly defined.

e) Theorem 1 may consider more then two classes. As it is to be seen, this is possible as sums of pairs–discriminant–information functions or globaly.

f) It must be mentioned that properties (1), (2), (3) are sufficient for proving Theorem 1, but in the framework of information theory, other characteristics of g–divergence family should be useful.

5. Examples

The g–divergence family may be choosen in different ways (but with respect to (1) - (3)). The aim of this paper was not to study and compare the properties of different g–divergences. Some cases will be mentioned in the sequel, informally only.

Corollary 3. *If $n = 2$ and $p_i(x) \neq 0$ $\forall x \in X$, $i = 1,2$ then Theorem 1 is valid for the case of Kullback divergence,*

$$(25)\qquad {}_1(p_1, p_2) =: \sum_{x \in X} (p_1(x) - p_2(x)) ln(p_1(x)/p_2(x)).$$

Proof.

$$g(y,z) =: (y - z)ln(y/z) = y(1 - z/y)ln(y/z) = yg(1, z/y) = yh(z/y),$$

where $h(t) =: (t-1)lnt = g(1,t)$, $t > 0$. For applying Theorem 1 it is necessary to prove that g is convex and $g(1,\cdot)$ is strictly convex. Taking into account

Lemma 2, it is sufficient to verify that h is strictly convex for $t > 0$. But this is true because $h''(t) > 0$. ■

Corollary 4. *If $n = 2$ and $p_1(x) + p_2(x) \neq 0$ $\forall x \in X$, then Theorem 1 is valid for the g–divergence J_2,*

$$_2(p_1, p_2) =: \sum_{x \in X} (p_1(x) - p_2(x))^2 / (p_1(x) + p_2(x)). \tag{26}$$

P r o o f. $g(y, z) =: (y - z)^2/(y - z) = yh(z/y)$, where $h(t) =: (1 - t)^2/(1 + t)$. but $h''(t) > 0$ for $t > 0$ and therefore h is strictly convex. Theorem 1 is valid, as in Corollary 3. ■

Corollary 5. *If $n = 3$ and $p_1(x) + p_2(x) + p_3(x) \neq 0$ $\forall x \in X$, then Theorem 1 is valid for the g–divergence J_3:*

$$_3(p_1, p_2, p_3) =: \sum_{x \in X} E_1(x)/E_2(x), \tag{27}$$

$$\begin{aligned} E_1(x) &=: (p_1(x) - p_2(x))^2 + (p_2(x) - p_3(x))^2 + (p_3(x) - p_1(x))^2, \\ E_2(x) &=: p_1(x) + p_2(x) + p_3(x). \end{aligned}$$

P r o o f. In the line of the proof of Corollary 1, the strict convexity of function $g(1, \cdot, \cdot)$ will be verified, where:

$$\begin{aligned} g(t, y, z) &=: ((t - y)^2 + (y - z)^2 + (z - t)^2)/(t + y + z) &= th(y/x, z/x), \\ h(u, v) &=: ((1 - u)^2 + (1 - v)^2 + (u - v)^2)/(1 + u + v) &= g(1, u, v). \end{aligned}$$

It is easy to verify that for $u, v > 0$ the Hesse matrix H of h is positively defined and $detH \neq 0$. ■

Corollary 6. *Let $n > 2$ and g–divergence function J_4 be*

$$_4(p_1, \ldots, p_n) =: \sum_{x \in X} g(x) \tag{28}$$

$$(x) =: \sum_{1 \leq i < j \leq n} f(p_i(x), p_j(x)). \tag{29}$$

The function $f : [0, 1]^2 \to [0, \infty)$ verifies (3) and the function $h : [0, 1] \to [0, \infty)$, $h(t) =: f(1, t)$ is strictly convex. Then:

I. $g(1,\cdot,\ldots,\cdot):[0,1]^{n-1}\to[0,\infty)$ *is strictly convex.*
II. Theorem 1 is valid for the g–divergence function J_4.

Proof. We have $g(t_1,\ldots,t_n)=t_1F(t_2/t_1,\ldots,t_n/t_1)$, where

$$(u_2,\ldots,u_n)=\sum_{j=2}^{n}f(1,u_j)+\sum_{2\le i<j\le n}f(u_i,u_j). \tag{30}$$

The strict convexity of h implies (31) and Lemma 2 implies the convexity of f, namely (32). For $\lambda\in[0,1]$ and $2\le i<j\le n$ we have:

$$(1,\lambda u_i+(1-\lambda)w_i)\le\lambda f(1,u_i)+(1-\lambda)f(1,w_i), \tag{31}$$

$$(\lambda u_i+(1-\lambda)w_i,\lambda u_j+(1-\lambda)w_j)\ge\lambda f(u_i,u_j)+(1-\lambda)f(w_i,w_j). \tag{32}$$

Adding all the relations (31) and (32) and considering (30), we obtain the convexity of F:

$$(u+(1-\lambda)w)\le\lambda F(u)+(1-\lambda)F(w). \tag{33}$$

If relation (33) is an equality, so must be relations (31) and (32). But the strict convexity of $f(1,\cdot)$ and the equalities (31) imply $u_i=w_i$, $i=2,n$, namely the strict convexity of $F=g(1,\cdot,\ldots,\cdot)$ and the convexity of g (Lemma 2). It is easy to prove that because f verifies (3), so does g. All necessary conditions for applying Theorem 1 are fulfiled. ■

6. Conclusions

Theorem 1 offers relation (11) as a necessary and sufficient condition for a transformation T to preserve the multi–class discriminant information. Moreover, if t is an injective function then the function $T_0(x)=:t(p_2(x)/p_1(x),\ldots,p_n(x)/p_1(x))$ defines the special case of T which verifies (11) and has the minimal cardinal of the image space. The result is valid for a general family of multi–class discriminant information: the family of g–divergences which are depending on the vector $(p_1(x),\ldots,p_n(x))$, are homogeneous and additive.

A remarcable similarity between the g–divergences family and Bayes classification procedures (which minimise the expectation of an appropriate loss function) may be noted: both are depending on the likelihood ratio $p_i(x)/p_1(x)$, $i=2,n$.

References

[1] T. W. Anderson. An Introduction to Multivariate Statisticval Analysis. Wiley, 1958.

[2] T. L. Boullion, P. L. Odell, B. S. Duran. Estimating the Probability of Misclassification and Variate Selection. *Pattern Rec.*, **7**, 1975, 139–145.

[3] B. Efron. Estimating the Error Rate of a Prediction Rule: Improvement on Cross–validation. *JASA 78*, 1983, 316–331.

[4] S. Kullback. Information Theory and Statistics. Dover Publ. Inc., 1968.

[5] F. Liese, I. Vajda. Convex Statistical Distances. Teubner, Leipzig, 1987.

[6] M. A. Morgan. The Effects of Selecting Variables for Use in the Linear Discriminant Function. *EDV in Medizin und Biologie*, Band **6**, Heft 1/2, 1975, 24–29.

[7] J. W. Van Ness, C. Simpson. On the Effects of Dimension in Discriminant Analysis. *Technometrics*, **18**, 175–187.

[8] W. Schaafsma. Selecting Variables in Discriminant Analysis for Improving upon Classical Procedures. In Handbook of Statistics, **2**, P.R. Krishnaiah, L.N. Kanal eds, North–Holand Publ. Comp., 1982.

[9] G. Toussaint. Bibliography on Estimation of Misclassification. IEEE Trans. on Inf. Th., **IT–20**(4), 1974.

[10] S. Watanabe. Pattern Recognition as a Quest for Minimum Entropy. *Patt. Rec.*, **13**(5), 1981, 381–387.

[11] J. M. Weiners, O. J. Dunn. Elimination of Variates in Linear Discriminat problems. *Biometrics*, **22**, 1966, 268–275.

[12] G. Zbaganu. Divergence and Contraction Coefficients. 1993, to apper.

Centre of Mathematical Statistics
Str. Magheru 22,
70158 Bucharest
ROMANIA

Received 26.08.1993
Revised version 11.10.1994