UNIVERSIDADE FEDERAL DE SANTA CATARINA

CARACTERIZAÇÃO DA J-DIVERGÊNCIA, SUAS GENERALIZAÇÕES E APLICAÇÕES EM PROBABILIDADE DE ERRO

ORIENTADOR : PROF.(DR.) INDER JEET TANEJA

FERNANDO LUIZ TAVARES DA SILVA

MARÇO - 1983

ESTA TESE FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE

"MESTRE EM CIÊNCIAS"

ESPECIALIDADE EM MATEMÁTICA, E APROVADA EM SUA FORMA FINAL PELO CURSO DE PÓS-GRADUAÇÃO EM MATEMÁTICA DA UNIVERSIDADE FEDERAL DE SANTA CATARINA

Prof. Inder Jeet Taneja, Ph.D.
COORDENADOR

BANCA EXAMINADORA :

Prof. Inder Jeet Taneja, Ph.D.
ORIENTADOR

Prof. Gur Dial, Ph.D.

Profa. Maria Helena Oltramari, Ms.

Aos meus

pais,

esposa,

e filhos.

AGRADECIMENTOS

Ao Prof. Inder Jeet Taneja, por sua orientação segura, pelo seu apoio, compreensão e amizade durante a elaboração deste trabalho.

Aos colegas, pelo incentivo, apoio e colaboração na elaboração deste trabalho.

À Universidade Federal de Santa Catarina.

RESUMO

Na literatura existem medidas de informação e distância ligadas com probabilidade de erro. J-divergência de Kullback é uma destas medidas.

Neste trabalho, apresentamos as relações entre as medidas de informação e distância juntamente com J-divergência e suas aplicações. Também apresentamos caracterizações de J-divergência e suas generalizações.

ABSTRACT

In the liturature there exist measures of information and distance connected with probability of error. Kullback's J-divergence is one of them.

In this work, we present the relations between the measures of information and distance jointly with J-divergence and its applications. We also present characterizations of J-divergence and its generalizations.

ÍNDICE

INTRODUÇÃO

No Capítulo I, apresentamos várias medidas de informação associadas com uma e duas distribuições de probabilidades. Também, apresentamos neste capítulo, o conceito de probabilidade de erro e suas cotas inferiores e superiores ligadas com várias medidas de informação.

No Capítulo II, algumas desigualdades foram obtidas entre a J-divergência de Kullback, o coeficiente de Bhattacharrya, a distância de Matusita e a distância variacional de Kolmogorov. Também foram obtidas cotas inferiores apertadas para a J-divergência de Kullback entre duas distribuições e as probabilidade de erro de três regras de decisão. Estas relações são importantes quando alguém pretende saber que forma pode ser esperada a partir de uma decisão quando características foram selecionadas usando a J-divergência. As três regras de decisão consideradas são a regra de Bayes (optimal), a regra do vizinho mais próximo e a regra de predição proporcional com decisão aleatória.

No Capítulo III, apresentamos várias generalizações de J-divergência tais como : J-divergência de grau β , J-diver-

gência de ordem α e duas formas diferentes de J-divergência de ordem α e grau β . Também daremos uma caracterização de J-divergência de ordem α e grau β através de discriminação (ou informação relativa) de Kullback de ordem α e grau β.

CAPÍTULO I

MEDIDAS DE INFORMAÇÃO E PROBABILIDADE DE ERRO

1.1. INTRODUÇÃO

A teoria da informação teve origem em 1948, com Claude E. Shannon, que, com seus estudos, criou resultados fundametais nesta área.

Este novo ramo de estudos tem muita importância e aplicações em muitas áreas como : Ciências Físicas e Biológicas , Psicologia, Linguística, Economia, Química, Ciência de Computação, Estatística etc..

É muito importante uma reexaminação e reformulação de idéias básicas para buscar novos significados e generalizações.

Neste capítulo apresentamos várias medidas associadas com uma e duas distribuições de probabilidades.

Seja

$$\Delta_M = \{ P=(p_1,p_2,\ldots,p_M) : p_i \geq 0, \sum_{i=1}^{M} p_i = 1\} , \qquad (1.1)$$

o conjunto de distribuições de probabilidades discretas.

Definimos as seguintes medidas de informação :

1.1.1. Entropia de Shannon (SHANNON |27|)

$$H(P) = -\sum_{i=1}^{M} p_i \log p_i \ , \qquad (1.2)$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

1.1.2. Entropia de Ordem α (RÉNYI |26|)

$$H_\alpha(P) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{M} p_i^{\alpha}\right), \quad \alpha \neq 1, \alpha > 0, \qquad (1.3)$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

Esta entropia tende para a entropia de Shannon quando $\alpha \to 1$.

1.1.3. Entropia de Grau β (HAVRDA e CHARVAT |13|; DARÓCZY |8|)

$$H^\beta(P) = (2^{1-\beta}-1)^{-1}\left\{\sum_{i=1}^{M} p_i^{\beta} -1\right\} \ , \beta \neq 1, \beta > 0, \qquad (1.4)$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

Esta entropia também tende para a entropia de Shannon quando $\beta \to 1$.

1.1.4. γ - Entropia (ARIMOTO |2|)

$$_\gamma H(P) = (\gamma-1)^{-1}\left\{\left(\sum_{i=1}^{M} p_i^{\gamma}\right)^{\frac{1}{\gamma}} -1\right\}, \gamma \neq 1, \ \gamma > 0, \qquad (1.5)$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

É fácil ver que $\lim_{\gamma \to 1} {}_{\gamma}H(P) = H(P)$.

Podemos verificar as seguintes relações entre a entropia de ordem α e a γ - entropia.

$$ {}_{\gamma}H(P) \leq H_{\alpha}(P) \leq H(P) \qquad \text{para} \quad 0 \leq \gamma = \alpha^{-1} < 1, \tag{1.6} $$

$$ {}_{\gamma}H(P) \leq H_{\alpha}(P) \leq H(P) \qquad \text{para} \quad \gamma = \alpha^{-1} > 1. \tag{1.7} $$

1.1.5. Entropia de Ordem α e Grau β (SHARMA E MITTAL |29|)

$$ H_{\alpha}^{\beta}(P) = (2^{1-\beta} - 1)^{-1}\{(\sum_{i=1}^{M} p_i^{\alpha})^{\frac{\beta-1}{\alpha-1}} - 1\}, \tag{1.8} $$

$$ \alpha \neq 1, \ \beta \neq 1, \ \alpha, \beta > 0. $$

Esta entropia possui todas as entropias definidas acima como casos particulares.

1.2. MEDIDAS ASSOCIADAS A DUAS DISTRIBUIÇÕES DE PROBABILIDADES

A entropia de Shannon associa uma medida de informação com uma distribuição de probabilidade. Duas medidas semelhantes introduzidas por KULLBACK |21| e KERRIDGE |20| associam uma medida com um par de distribuições de probabilidades de uma incerteza variável. Estas medidas são mais gerais do que a entropia de Shannon.

1.2.1. Informação Discriminação de Kullback

Seja X uma variável aleatória com número finito de valores $x_1, x_2, \ldots, x_M$ com probabilidades $Q = (q_1, q_2, \ldots, q_M) \in \Delta_M$, de um experimento E. Seja a frequência relativa definida como $P=(p_1, p_2, \ldots, p_M) \in \Delta_M$ então, a informação discriminação de Kullback de um experimento é definida por

$$I(P:Q) = \sum_{i=1}^{M} p_i \log\left(\frac{p_i}{q_i}\right). \qquad (1.9)$$

Se algum q_i for zero, então o correspondente p_i também será zero. Tomamos $0 \log\left(\frac{0}{0}\right) = 0 \log 0 - 0 \log 0 = 0$.

Um estudo mais detalhado dessa medida com suas aplicações em estatística, encontra-se em Kullback |21|.

Algumas aplicações dessa medida em Economia, foram apresentadas for THEIL |40|.

1.2.2. Informação de Kerridge

Suponha um dado experimento no qual as probabilidades dos M eventos distintos $x_1, x_2, \ldots, x_M$ são $Q=(q_1, q_2, \ldots, q_M) \in \Delta_M$. Então suas reais probabilidades são $P=(p_1, p_2, \ldots, p_M)$. Logo, a transmissão pode ser menos precisa em dois casos :

a) A transmissão pode ser indefinida por falta de informação ;

b) A transmissão pode ser incorreta porque a informção é incorreta.

KERRIDGE |20| introduziu uma medida que considera estes dois aspectos. Sua medida é dada por

$$H(P:Q) = - \sum_{i=1}^{M} p_i \log q_i . \qquad (1.10)$$

Se qualquer q_i for zero então o correpondente p_i também será zero, e é adotada a convenção $0 \log 0 = 0$. Nessa entropia temos :

$$H(P:Q) = - \sum_{i=1}^{M} p_i \log p_i + \sum_{i=1}^{M} p_i \log \frac{p_i}{q_i}$$

$$= H(P) + E(P:Q).$$

Kerridge chamou $E(P:Q)$ de erro condicional.

As generalizações destas duas medidas (1.9) e (1.10) juntamente com suas caracterizações foram estudadas por RÉNYI |26|, TANEJA |33|, SHARMA e TANEJA |31|, SHARMA e AUTAR |28| e outros.

1.2.3. J-Divergência de Kullback

Também foi estudada por KULLBACK |21| uma medida associada a duas distribuições de probabilidades $P \in \Delta_M$ e $Q \in \Delta_M$ dada por

$$J(P:Q) = \sum_{i=1}^{M} p_i \log\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{M} q_i \log\left(\frac{q_i}{p_i}\right) \qquad (1.11)$$

$$= I(P:Q) + I(Q:P),$$

onde $I(P:Q)$ é a informação discriminação de Kullback dada em (1.9).

Este medida tem tido muitas aplicações em reconhecimento de modelos para obter cotas superiores e inferiores da probabilidade de erro. Mais detalhes do estudo desta medida, veremos no Capítulo II.

As propriedades desta medida podem ser encontradas em KULLBACK |21|, MATHAI e RATHIE |23|.

1.2.4. GENERALIZAÇÕES DE J-DIVERGÊNCIA

1.2.4.1. J- Divergência de Grau β

RATHIE e LILLIAN |35| estudaram uma generalização de J-divergência, chamada J-divergência de grau β , dada por

$$J^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1} \{ \sum_{i=1}^{M} p_i^{\beta} q_i^{1-\beta} + \sum_{i=1}^{M} p_i^{1-\beta} q_i^{\beta} - 2\}, \quad (1.12)$$

$$\beta \neq 1, \beta > 0$$

para todo $P, Q \in \Delta_M$.

Podemos escrever

$$J^{\beta}(P:Q) = I^{\beta}(P:Q) + I^{\beta}(Q:P) , \quad (1.13)$$

onde

$$I^{\beta}(P:Q) = (2^{\beta-1} - 1)^{-1} \{ \sum_{i=1}^{M} p_i^{\beta} \, q_i^{1-\beta} - 1 \}, \qquad (1.14)$$

é uma generalização de informação discriminação de grau β .

A medida (1.10) satisfaz as seguintes propriedades :

i) Não-Negatividade: $J^{\beta}(P:Q) \geq 0$ com a igualdade se, e somente se, $p_i = q_i$ para qualquer $i = 1,2,\ldots,M$.

ii) Simetria :

$$J^{\beta}(p_1,p_2,\ldots,p_M:q_1,q_2,\ldots,q_M) = J^{\beta}(p_{a_1},p_{a_2},\ldots,p_{a_M} : q_{a_1},q_{a_2},\ldots,q_{a_M})$$

onde $\{a_1,\ldots,a_M\}$ é uma permutação arbitrária de $\{1,2,\ldots,M\}$.

iii) Expansibilidade :

$$J^{\beta}(p_1,\ldots,p_M,0:q_1,\ldots,q_M,0) = J^{\beta}(p_1,\ldots,p_M:q_1,\ldots,q_M).$$

iv) Recursividade :

$$J^{\beta}(p_1,\ldots,p_M:q_1,\ldots,q_M) = J^{\beta}(p_1+p_2,p_3,\ldots,p_M: q_1+q_2,q_3,$$

$$\ldots,q_M) + (p_1+p_2)^{\beta}(q_1+q_2)^{1-\beta} \, I^{\beta}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} : \frac{q_1}{q_1+q_2}, \right.$$

$$\left. \frac{q_2}{q_1+q_2}\right) + (q_1+q_2)^{\beta}(p_1+p_2)^{1-\beta} I^{\beta}\left(\frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2} : \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$$

para $p_1 + p_2 > 0$, $q_1 + q_2 > 0$, onde I^β é a informação generalizada de Kullback dada em (1.14).

v) Não-Aditividade :

Para todo $P, Q \in \Delta_M$ e $R,S \in \Delta_M$,

$$J^\beta(P*R:Q*S) = J^\beta(P:Q) + J^\beta(R:S) + (2^{\beta-1} - 1)\{I^\beta(P:Q)\ I^\beta(R:S) + I^\beta(Q:P)\ I^\beta(R:S)\}$$

vi) Não-Aditividade Forte :

Para $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$, $Q=(q_1,q_2,\ldots,q_M) \in \Delta_M$, $P_i=(p_{1i},p_{2i},\ldots,p_{Ni}) \in \Delta_N$ e $Q_i=(q_{1i},q_{2i},\ldots,q_{Ni}) \in \Delta_N$, para $i=1,2,\ldots,M$, temos

$$J^\beta(p_1p_{11},p_2p_{21},\ldots,p_1p_{N1},\ldots,p_Mp_{1M},\ldots,p_Mp_{NM}:$$
$$q_1q_{11},q_2q_{21},\ldots,q_1q_{N1},\ldots,q_Mq_{1M},\ldots,q_Mq_{NM})$$
$$= J^\beta(P:Q) + \sum_{i=1}^{M} p_i^\beta\, q_i^{1-\beta}\, I^\beta(P_i:Q_i) + \sum_{i=1}^{M} q_i^\beta\, p_i^{1-\beta}\, I^\beta(Q_i:P_i) ,$$

vii) Continuidade :

$J^\beta(P:Q)$ é uma função contínua nas $2M$ - variáveis.

viii) Troca das Distribuições :

Para todo $P, Q \in \Delta_M$, temos

$$J^{\beta}(P:Q) = J^{\beta}(Q:P).$$

ix) Nulidade :

Seja $g(x,y) = J^{\beta}(x,1-x:y,1-y)$, $x, y \in [0,1]$,

$g(x,x) = 0$ para $x \in [0,1]$.

x) Normalização :

$$g(0,1) = 2(1-2^{\beta-1})^{-1} .$$

Baseada em algumas propriedades citadas acima foi caracterizada por RATHIE e LILLIAN |25|, a medida (1.12) no seguinte Teorema.

TEOREMA :

As propriedades : a) simetria (ii), quando M=3, b) recursividade (iv), c) Nulidade (ix), d) normalização (x) e e) a exitência de uma derivada de $F(x,y)$ com respeito a x e y em $(0,1)$, determinam unicamente a J-divergência de grau β , $J^{\beta}(P:Q)$ para $\beta \neq 1$, $\beta > 0$.

1.2.4.2. J-Divergência de Ordem α

No Capítulo III apresentamos a seguinte medida de informação chamada de J-divergência de ordem α.

$$J_\alpha(P:Q) = \frac{1}{1-\alpha} \log\{ \sum_{i=1}^{M} p_i^\alpha \, q_i^{1-\alpha} + \sum_{i=1}^{M} p_i^{1-\alpha} \, q_i^\alpha \}, \qquad (1.15)$$

para todo $P \in \Delta_M$, $Q \in \Delta_M$.

1.2.4.3. J-Divergência de Ordem α e Grau β

Apresentamos também no Capítulo III, as seguintes medidas de informação chamadas de J-divergência de ordem α e grau β.

$$'J_\alpha^\beta(P:Q) = (2^{\beta-1}-1)^{-1} \{ (\sum_{i=1}^{M} p_i^\alpha \, q_i^{1-\alpha})^{\frac{\beta-1}{\alpha-1}} + (\sum_{i=1}^{M} p_i^{1-\alpha} q_i^\alpha)^{\frac{\beta-1}{\alpha-1}} - 2 \}$$

.... (1.16)

$$''J_\alpha^\beta(P:Q) = (2^{\beta-1}-1)^{-1} \{ (\sum_{i=1}^{M} p_i^\alpha \, q_i^{1-\alpha} + \sum_{i=1}^{M} p_i^{1-\alpha} q_i^\alpha)^{\frac{\beta-1}{\alpha-1}} - 2 \},$$

.... (1.17)

$$\alpha \neq 1, \beta \neq 1, \quad \alpha, \beta > 0.$$

Estas duas medidas tem como casos particulares, J-divergência (1.11), J-divergência de grau β (1.12) e J-divergência de ordem α (1.15).

Para o estudo desta medida em obtenção de cotas superiores e inferiores de probabilidade de erro veja TANEJA |38| .

1.3. PROBABILIDADE DE ERRO

Consideremos o problema de teoria de decisão, de classificar uma observação X proveniente de M possíveis classes (hipóteses) $C = (C_1, C_2, \ldots, C_M)$.

Denotemos $P_i = Pr\{C=C_i\}$, $i=1,2,\ldots,M$ probabilidade de classe $C=C_i$, $i=1,2,\ldots,M$ e denotemos $f_1(x)$, $f_2(x)$, $\ldots, f_M(x)$ a função densidade condicional dada a verdadeira classe ou hipótese, i.é.,

$$f_i(x) = Pr\{X=x/C=C_i\} \ , \ i=1,2,\ldots,M.$$

Suponhamos que $f_i(x)$ são completamente conhecidas. Dada qualquer observação X=x, podemos calcular a probabilidade condicional de C, pela regra de Bayes :

$$P(C_i/x) = Pr\{C=C_i/X=x\} = \frac{P_i \ f_i(x)}{\sum_{i=1}^{M} P_i \ f_i(x)} \ , \ i=1,2,\ldots,M.$$

É bastante conhecido que a regra de decisão que minimiza a probabilidade de erro, é a regra de decisão de Bayes, a qual escolhe a hipótese com a maior probabilidade posterior. Usando esta regra, a probabilidade de erro para um dado X=x é expressa por

$$P(e/x) = 1 - \max \{P(C_1/x), P(C_2/x), \ldots, P(C_M/x)\} \ .$$

Antes de observar x, a probabilidade de erro $P(e)$ associada a X, é definida como a probabilidade de erro esperada após observa-lo, i.é.,

$$P(e) = E_X \{1 - \max [P(C_1/x), \ldots, P(C_M/x)]\}$$

$$= 1 - E_X\{ \max [P(C_1/x), \ldots, P(C_M/x)]\}.$$

1.3.1. Cotas Superiores

i) Em |11|, |14| são obtidas cotas superiores da probabilidade de erro em termos de entropia de Shannon.

$$P(e) \leq \frac{1}{2} I \ , \tag{1.18}$$

onde I é a expectativa dada por

$$I = E_X[H(C/x)] \ , \tag{1.19}$$

$H(C/x)$ é a entropia condicional de Shannon de C dado $X=x$, e é dada por

$$H(C/x) = - \sum_{i=1}^{M} P(C_i/x) \log P(C_i/x) \ . \tag{1.20}$$

ii) Em |3| uma cota superior da probabilidade de erro foi obtida em termos de entropia de Rényi (entropia de ordem α) e é dada por

$$P(e) \leq \frac{1}{2} I_\alpha \ , \quad \alpha > 0 \tag{1.21}$$

onde

$$I_\alpha = E_X\{H_\alpha(C/x)\} \ , \tag{1.22}$$

e $$H_\alpha(C/x) = \frac{1}{1-\alpha} \log\{\sum_{i=1}^{M} P(C_i/x)^\alpha\} \quad \alpha\neq 1, \alpha > 0. \tag{1.23}$$

A cota (1.16) é mais próxima da desigualdade que (1.14) desde que H seja uma função decrescente de α. Suponhamos que $P(e/x) \geq \frac{1}{2}$. Então, é obtida uma cota mais geral dada por

$$P(e) \leq \frac{1}{2} I_\alpha \ , \ \alpha > 0. \tag{1.24}$$

iii) Em |36| uma cota superior da probabilidade de erro em termos de γ - entropia (ARIMOTO |2|) é dada por :

$$P(e) \leq \frac{1}{2} \ {}_\gamma I \ , \ \gamma > 0, \tag{1.25}$$

onde

$${}_\gamma I = E_X \ \{{}_\gamma H(C/x)\} \tag{1.26}$$

e

$${}_\gamma H(C/x) = (\gamma-1)^{-1} \ \{(\sum_{i=1}^{M} P(C_i/x)^{\frac{1}{\gamma}})^\gamma - 1\} \ , \tag{1.27}$$
$$\gamma \neq 1, \ \gamma > 0.$$

Isto é , a entropia condicional de C dado X=x.

Usando a desigualdade (1.7) podemos concluir que o resultado (1.25) é mais próximo da probabilidade de erro do que (1.21).

iv) Em |39| uma cota superior da probabilidade de erro foi obtida em termos de entropia de ordem α e grau β e é dada por

$$P(e) \leq \frac{1}{2} I_{\alpha}^{\beta} \quad , \quad 0 < \alpha \leq \frac{1}{2-\beta} \quad , \tag{1.28}$$

onde

$$I_{\alpha}^{\beta} = E_X\{ H_{\alpha}^{\beta}(C/x)\} \quad ,$$

é a equivocação generalizada e $H_{\alpha}^{\beta}(C/x)$ é a entropia condicional de ordem α e grau β dada por

$$H_{\alpha}^{\beta}(C/x) = (2^{1-\beta}-1)^{-1}\{ [\sum_{i=1}^{M} P(C_i/x)^{\alpha}]^{\frac{\beta-1}{\alpha-1}} -1\}, \tag{1.29}$$

$$\alpha \neq 1, \quad \beta \neq 1, \quad \alpha,\beta > 0.$$

Suponhamos que para todos os valores possíveis de x, $P(C/x) \geq \frac{1}{2}$. Obtemos uma cota superior mais geral como

$$P(e) \leq \frac{1}{2} \quad I_{\alpha}^{\beta} \quad , \quad \alpha > 0, \quad 0 < \beta \leq 2 \tag{1.30}$$

Também foi mostrada em |39| que a cota (1.28) é bem melhor que (1.18), (1.21) e (1.25).

1.3.2. Cotas Inferiores

Em |39| foi obtida uma cota inferior da probabilidade de erro em termos de entropia de ordem α e grau β que é dada por

$$H_{\alpha}^{\beta}(C/X) \leq (2^{1-\beta}-1)^{-1}\{[(1-P(e))^{\alpha}+(M-1)(\frac{P(e)}{M-1})^{\alpha}]^{\frac{\beta-1}{\alpha-1}} -1\}, \quad \ldots\ldots \quad (1.31)$$

$$\alpha > 0, \beta \geq 2 - \frac{1}{\alpha} .$$

Casos Particulares :

i) Quando $\alpha = \beta$, $\beta > 0$ temos

$$H^{\beta}(C/X) \leq (2^{1-\beta} -1)^{-1}\{(1-P(e))^{\beta} + (M-1)((\frac{P(e)}{M-1})^{\beta}-1\}, \quad (1.32)$$

o qual é cota inferior para entropia de grau β estudada por DEVIJVER |10| e TANEJA |34|.

ii) Quando $\gamma = \frac{1}{\alpha} = 2 - \beta$, temos

$$_{\gamma}H(C/X) \leq (2^{-1}-1)^{-1}\{[(1-P(e))^{\frac{1}{\gamma}} +(M-1)(\frac{P(e)}{M-1})^{\frac{1}{\gamma}}] -1\}, \quad (1.33)$$

a qual é cota inferior para γ - entropia estudada por BOEKEE e LUBBE |4|.

iii) Quando $\beta \to 1$, temos

$$\lim_{\beta \to 1} H_{\alpha}^{\beta}(C/X) = H_{\alpha}(C/X) \text{ e}$$

$$H_{\alpha}(C/X) \leq (1-\alpha)^{-1} \log\{(1-P(e))^{\alpha}+ (M-1)(\frac{P(e)}{M-1})^{\alpha}\}, \quad (1.34)$$

$$\alpha > 0.$$

a qual é cota inferior para entropia de ordem α estudada por TOUSSAINT |43| e TANEJA |35|.

iv) Quando $\alpha = 1$ e $\beta \to 1$, temos

$$H(C/X) \leq - (1-P(e)) \log (1-P(e)) - P(e) \log \frac{P(e)}{M-1}, \quad (1.35)$$

a qual é bem conhecida como cota de Fano.

v) Quando $\alpha = 2$, de (1.31) temos

$$P(e) \geq \frac{M-1}{M} \left\{ 1 - \sqrt{\frac{M\, B(C/X) - 1}{M-1}} \right\}, \quad (1.36)$$

onde

$$B(C/X) = E_X \{ \sum_{i=1}^{M} P(C_i/x) \} ,$$

é a distância Bayesiana definida por DEVIJVER |9|.

A cota inferior (1.36) foi estudada pela primeira vez por DEVIJVER |9|.

b) KAILATH |18| obteve a cota inferior da probabilidade de erro em termos de J-divergência de classe dois e mostrou que

$$P(e) \geq \frac{1}{2} \exp \left(- \frac{J}{2}\right) , \quad (1.37)$$

onde

$$J = \int [P(X/C_1) - P(X/C_2)] \log \left[\frac{P(X/C_1)}{P(X/C_2)}\right] dx$$

O resultado (1.37) foi bastante superado por TOUSSAINT |42| e é dado por

$$P(e) > \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4 \exp\left[-2 H(\pi) - J(C/X)\right]} , \qquad (1.38)$$

onde $H(\pi)$ é a entropia de Shannon para $M=2$ dada por

$$H(\pi) = - \pi_1 \log \pi_1 - \pi_2 \log \pi_2 .$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, torna-se

$$P(e) > \frac{1}{2} - \frac{1}{2}\sqrt{1 - \exp\left(-\frac{J}{2}\right)} . \qquad (1.39)$$

Um estudo mais geral sobre estes resultados, juntamente com suas aplicações, é apresentado no Capítulo II.

CAPÍTULO II

DESIGUALDADES ENTRE MEDIDAS DE DISTÂNCIA, J-DIVERGÊNCIA E PROBABILIDADE DE ERRO

Neste Capítulo, algumas desigualdades foram obtidas entre a divergência de Kullback, o coeficiente de Bhattacharraya, a distância de Matusita e a distância variacional de Kolmogorov. Também foram obtidas cotas inferiores apertadas para a divergência de Kullback entre duas distribuições e as probabilidades de erro de tres regras de decisão. As tres regras de decisão consideradas são a regra de Bayes (optimal), a regra do vizinho mais próximo e a regra de predição proporcional com decisão aleatória. Para mais detalhes deste estudo veja referências |5| , |24| , |41| e |42|.

2.1. DESIGUALDADES ENTRE MEDIDAS DE DISTÂNCIA

Recentemente, houve um grande interesse em aplicar medidas de distância entre duas funções de densidade de probabilidade (FDP) ao problema de reconhecimento de modelos. Uma das primeiras medidas a ser usada, a qual tem recebido considerável atenção é a J-divergência de Kullback.

Seja $P(X/C_i)$ a classe condicional FDP, onde X representa o vector característica que pode assumir algum valor na linha real. A divergência entre os dois FDP condicionais, representando as duas classes (modelos) C_1 e C_2, é dada por (utilizamos logarítmos naturais)

$$J = \int [P(X/C_1) - P(X/C_2)] \log [\frac{P(X/C_1)}{P(X/C_2)}] \, dx \, . \qquad (2.1)$$

Recentemente outras medidas muito relacionadas a probabilidade de erro tem sido aplicadas a avaliação tais como : coeficiente Bhattacharya, a distância variacional Kolmogorov e a distância Matusita, dadas repectivamente por

$$\rho = \int [P(X/C_1) \; P(X/C_2)]^{\frac{1}{2}} \, dX \, , \qquad (2.2)$$

$$V = \int [P(X/C_1) - P(X/C_2)] \, dX \, , \qquad (2.3)$$

e

$$M = \left\{ \int \{ [P(X/C_1)]^{\frac{1}{2}} - [P(X/C_2)]^{\frac{1}{2}} \}^2 \, dX \right\}^{\frac{1}{2}} . \qquad (2.4)$$

Tentativas estão sendo feitas para encontrar relações entre as medidas definidas acima e a probabilidade de erro, assim como, entre as próprias medidas, com o objetivo de entendê-las melhor. Neste Capítulo algumas relações entre as próprias medidas são deduzidas nos tres teoremas que seguem :

TEOREMA 2.1.

A divergência está relacionada ao coeficiente de Bhatta-

charrya pela seguinte desigualdade :

$$J \geq 4(1-\rho). \tag{2.5}$$

Demonstração :

De (2.1) obtemos

$$J = 2 \sum_{i=1}^{2} \int P(X/C_i) \log\{ \frac{P(X/C_i}{[P(X/C_1)P(X/C_2)]^{\frac{1}{2}}} \} \; dX \; , \tag{2.6}$$

ou, alternadamente,

$$J = 2 \sum_{i=1}^{2} \int P(X/C_i) \log P(X/C_i) dX - 2 \sum_{i=1}^{2} K(X,i), \tag{2.7}$$

onde $$K(X,i) = \int P(X/C_i) \log G(X) \; dX \; , \tag{2.8}$$

e $$G(X) = [P(X/C_1) \; P(X/C_2)]^{\frac{1}{2}} \; . \tag{2.9}$$

Sejam duas funções de correção $C(X,i)$, $i=1,2$ definidas como $G(X)$, e escritas como :

$$G(X) = P(X/C_i) + C(X,i) \tag{2.10}$$

$$= P(X/C_i) \; [1 + \frac{C(X,i)}{P(X/C_i)}], \; i=1,2.$$

Integrando (2.10) e usando (2.2) temos

$$\int C(X,i) \; dX = \rho - 1 \; , \quad i=1,2. \tag{2.11}$$

Substituindo (2.10) em (2.8) temos :

$$K(X,i) = \int P(X/C_i) \log P(X/C_i)\, dX + \int P(X/C_i) \log\left[1 + \frac{C(X,i)}{P(X/C_i)}\right] dX \quad (2.12)$$

Agora, pode fàcilmente ser verificado que, para $i=1,2$, para qualquer valor de X

$$\frac{C(X,i)}{P(X/C_i)} \geq \log\left[1 + \frac{C(X,i)}{P(X/C_i)}\right], \quad (2.13)$$

Substituindo (2.13) em (2.12) temos

$$K(X,i) \leq \int P(X/C_i) \log P(X/C_i)\, dX + \int C(X,i)\, dX. \quad (2.14)$$

Substituindo (2.8) e (2.11) em (2.14), temos

$$\int P(X/C_i) \log \frac{P(X/C_i)}{G(X)}\, dX \geq 1 - \rho. \quad (2.15)$$

Finalmente, substituindo (2.15) em (2.6), obtemos (2.5).

TEOREMA 2.2.

A divergência está relacionada a distância de Matusita pela seguinte desigualdade

$$J \geq 2\, M^2 . \quad (2.16)$$

Demonstração :

Considere

$$d = \int \{ [P(X/C_1)]^{\frac{1}{2}} - [P(X/C_2)]^{\frac{1}{2}} \}^2 \, dX = M^2 \qquad (2.17)$$

(por (2.4))

Kailath |18| mostrou que

$$d = 2 \, (1 - \rho). \qquad (2.18)$$

Em outras palavras, a distância de Matusita é unicamente relacionada ao coeficiente de Bhattcharrya. Substituindo (2.18) e (2.17) em (2.5), temos (2.16).

TEOREMA 2.3.

A divergência está relacionada à distância variacional de Kolmogorov pela seguinte desigualdade :

$$J \geq 4 \{1 - [1 - (\frac{V}{2})^2]^{\frac{1}{2}}\}. \qquad (2.19)$$

Demonstração :

Para o problema de dois modelos de classe, Kailath apresenta a seguinte desigualdade

$$K \leq \frac{1}{2} [1 - 4 \, P(C_1) \, P(C_2) \, \rho^2]^{\frac{1}{2}} , \qquad (2.20)$$

onde

$$K = \frac{1}{2} \int [P(X/C_1) \, P(C_1) - P(X/C_2) \, P(C_2)] \, dX$$

e $P(C_i)$ é "a priori" a probabilidade do i-ésima modelo de classe.

Quando $P(C_1) = P(C_2) = \frac{1}{2}$ e (2.3) é substituído em (2.20), obtemos

$$\rho \leq [1 - (\frac{V}{2})^2]^{\frac{1}{2}} . \qquad (2.21)$$

Finalmente, substituindo (2.21) em (2.5) obtemos (2.19).

2.2. PROBABILIDADE DE ERRO DE VÁRIAS REGRAS DE DECISÃO E J-DIVERGÊNCIA DE KULLBACK

Seja a classe C_i, a priori, com a probabilidade π_i, $i=1,2$, $\pi_1 + \pi_2 = 1$. Isto é útil para definir medidas mais gerais que (2.1) e (2.2), como segue :

$$\rho(\pi_1,\pi_2) = \sqrt{\pi_1\pi_2} \int \sqrt{P(X/C_1)\ P(X/C_2)}\ dX$$

$$= \int P(X) \sqrt{P(C_1/X)\ P(C_2/X)}\ dX$$

$$= \int P(X)\ \rho(X)\ dX , \qquad (2.22)$$

onde $\rho(X) = \{P(C_1/X)\ P(C_2/X)\}^{\frac{1}{2}}$.

$$J(\pi_1,\pi_2) = \int [\pi_1 P(X/C_1) - \pi_2 P(X/C_2)] \log \frac{\pi_1 P(X/C_1)}{\pi_2 P(X/C_2)}\ dX$$

$$= \int P(X)\ [P(C_1/X) - P(C_2/X)] \log \frac{P(C_1/X)}{P(C_2/X)}\ dX$$

$$= \int P(X)\ J(X)\ dX , \qquad (2.23)$$

$$J(X) = [P(C_1/X) - P(C_2/X)] \log \frac{P(C_1/X)}{P(C_2/X)} .$$

Segue que $\rho(\frac{1}{2},\frac{1}{2}) = \frac{\rho}{2}$ e $J(\frac{1}{2},\frac{1}{2}) = \frac{J}{2}$ em (2.22) e (2.23) respectivamente, $P(X)$ é a distribuição mistura dada por

$$P(X) = P(X/C_i)\ \pi_1 + P(X/C_2)\ \pi_2 .$$

Apresentamos agora as tres regra de decisão :

i) A primeira regra de decisão é a regra (optimal) de Bayes. Dado um vetor X característica, a partir de algum modelo desconhecido P, P é classificado como pertencente a classe C_i se $P(C_i/X) > P(C_j/X)$, $i=1,2$, $i \neq j$. Esta regra dá a mínima probabilidade de erro possível a qual é dada por

$$P(e) = \int \min_i \left[P(X/C_i)\ \pi_i\right] dX \quad , \ i=1,2.$$

$$= \int P(X) \min_i [P(C_1/X),\ P(C_2/X)] dX$$

$$= \int P(X)\ P_e(X)\ dX . \qquad (2.25)$$

ii) A segunda regra de decisão considerada aqui, é a regra do vizinho mais próximo. Seja $\{X,\theta\} = \{X_1,\theta_1; X_2,\theta_2;\ldots;X_M,\theta_M\}$ o conjunto de M espécies de modelos disponíveis, onde X_i e θ_i denotam, respectivamente,

o vetor característica e a informação de classificação do espécie modelo i-ésimo. É certo que cada θ_i associada com X_i é o procedimento correto, isto é, as espécies modelos tem sido corretamente pré-classificadas. Seja $(X_M,\theta_M)\ \varepsilon\{X,\theta\}$ a espécie mais próxima do X desconhecido. P é então classificado como pertencente a classe associada com θ_M. COVER e HART |7| mostraram que, quando $M \to \infty$, a percentagem de erro do vizinho mais próximo, denotada por R, é dada por

$$R = \int [2\ P(X/C_1)\pi_1\ P(X/C_2)\pi_2/P(X)]dX$$

$$= \int P(X)[\ 2\ P(C_1/X)\ P(C_2/X)]\ dX$$

$$= \int P(X)\ R(X)\ dX\ ,$$

onde $R(X) = 2\ P(C_1/X)\ P(C_2/X)$.

iii) A terceira regra de decisão sob investigação é a regra de decisão aleatória. Sejam as distribuições da classe condicional conhecidas como na regra determinística de Bayes. Dado o vetor X característica de algum modelo P desconhecido. P é classificado como pertencente a classe C_i, i=1,2, pelo lançamento de uma moeda, a qual indica C_i, com probabilidade $P(C_i/X)$. Este tipo de regra de decisão, tende a produzir uma distribuição de classificação mais similar a distribuição que a regra determinística de Bayes.

A probabilidade de erro, usando esta regra, denotada por R, por razões que não se tornam aparentes, pode ser deduzida como se segue :

Para algum valor dado de X, C_1 ocorre com probabilidade $P(C_1/X)$ e é decidida que pertence a classe C_2 com probabilidade $1 - P(C_1/X)$. Similarmente, C_2 ocorre com probabilidade $P(C_2/X)$ e é decidida que pertence a classe C_1 com probabilidade $1 - P(C_2/X)$. Portanto, para um valor dado de X, a probabilidade de erro é dada por :

$$R(X) = P(C_2/X)\left[1-P(C_2/X)\right] + P(C_1/X)\left[1-P(C_1/X)\right]$$

$$= 2\, P(C_1/X)\, P(C_2/X). \qquad (2.27)$$

Tomando o valor esperado de (2.27) com relação a $P(X)$ temos

$$R = \int P(X)\, R(X)\, dX\ , \qquad (2.28)$$

o qual é mesmo que a percentagem de erro do vizinho mais próximo .

2.2.1. Uma Desigualdade Entre $J(\pi_1,\pi_2)$ e $\rho(\pi_1,\pi_2)$

A divergência entre duas distribuições, ocorrendo "a priori" com probabilidades π_1 e π_2, podem ser escritas como:

$$J(\pi_1,\pi_2) = -\pi_1 E_1\{\log[\frac{P(X/C_2)\pi_2}{P(X/C_1)\pi_1}]\}$$

$$-\pi_2 E_2\{\log[\frac{P(X/C_1)\pi_1}{P(X/C_2)\pi_2}]\},$$

onde E_i, $i=1,2$, representa valor esperado com relação a $P(X/C_i)$. Como $\log x$ é uma função convexa crescente, a desigualdade de Jensen aplicada (2.29) fornece

$$J(\pi_1,\pi_2) \geqq -2\pi_1 \log E_1\{\frac{P(X/C_2)\pi_2}{P(X/C_1)\pi_1}\}$$

$$-2\pi_2 \log E_2\{\frac{P(X/C_1)\pi_1}{P(X/C_2)\pi_2}\},$$

ou, alternadamente,

$$J(\pi_1,\pi_2) \geqq -2\pi_1 \log[\rho(\pi_1,\pi_2)/\pi_1]$$

$$-2\pi_2 \log[\rho(\pi_1,\pi_2)/\pi_2]. \qquad (2.30)$$

Expandindo (2.30) e recombinando termos, temos o seguinte resultado :

$$J(\pi_1,\pi_2) \geqq -2\left[H(\pi) + \log \rho(\pi_1,\pi_2)\right], \qquad (2.31)$$

onde $H(\pi)$ é a função de entropia dada por

$$H(\pi) = -\pi_1 \log \pi_1 - \pi_2 \log \pi_2. \qquad (2.32)$$

Quando $\pi_1 = \pi_2 = \frac{1}{2}$, $H(\pi) = \log 2$, (2.31) reduz para

$$J \geq - 4 \log \rho \ .$$

2.2.2. Cotas Inferiores Para R e J

Apesar de existência de um cota inferior apertada para R(X) em termos de J(X), nenhuma cota é disponível na literatura entre R e J. HORIBE |16| mostrou que

$$R \geq 2\left[\rho(\pi_1,\pi_2)\right]^2 \ ,$$

a partir do qual segue que

$$\log \rho(\pi_1,\pi_2) \leq \log \sqrt{R/2} \ . \qquad (2.34)$$

Substituindo (2.34) em (2.31) temos

$$R \geq \exp\left[- 2\, H(\pi) - J(\pi_1,\pi_2)\right] , \qquad (2.35)$$

para $\pi_1 = \pi_2 = \frac{1}{2}$, (2.35) se reduz para

$$R \geq \frac{1}{2} \exp\left[-\frac{J}{2}\right], \qquad (2.36)$$

onde a igualdade se mantém quando $J = 0$ e $J = \infty$.

CHITTI BABU |6| mostrou que

$$R(X) \geq \frac{1}{2}\left[1 - J(X)/2\right] . \qquad (2.37)$$

Obtendo valores esperados de ambos os lados de (2.37) temos uma segunda cota inferior para R em termos de J, como é

mostrada abaixo :

$$R \geq (\tfrac{1}{2}) \left[1 - J(\pi_1,\pi_2)/2 \right], \tag{2.38}$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, (2.38) se reduz para

$$R \geq (\tfrac{1}{2}) - \frac{J}{8} , \tag{2.39}$$

onde a igualdade é valida quando $J = 0$.

Uma cota inferior para J em termos de R, que é mais exata que (2.36) e (2.39), é respectivamente dada por

$$J(\pi_1,\pi_2) \geq \sqrt{1 - 2R}\, \log\{ \frac{1+\sqrt{1+2R}}{1-\sqrt{1-2R}} \}. \tag{2.40}$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, (2.40) se reduz para

$$J \geq 2\sqrt{1-2R}\, \log \{\frac{1 +\sqrt{1 + 2R}}{1 -\sqrt{1 - 2R}} \} , \tag{2.41}$$

pois $J(\frac{1}{2},\frac{1}{2}) = \frac{J}{2}$.

Observação :

Apesar de (2.41) dar a desigualdade mais exata, ele tem a desvantagem de não poder ser solucionado para R, como uma função de J, a qual é a forma mais útil, desde que J está para ser computado explìcitamente mais que R. Para demonstração de que (2.41) é mais exato que (2.36) e (2.39), veja Apêndice A.

2.2.3. Cotas Inferiores Para P(e) e J

Existe na literatura, uma cota inferior para P(e) em termos de J. Isto foi primeiramente desenvolvido por KAILATH |29| e é dada por

$$P(e) \geq (\frac{1}{2}) \quad \exp\ (-\ \frac{J}{2}\ )\ . \tag{2.42}$$

onde a igualdade é válida quando $J = \infty$.

COVER e HART |7| , mostraram que

$$R \leq 2\ P(e)\ (1-\ P(e))\,, \tag{2.43}$$

e, portanto,

$$R \leq 2\ P(e)\,. \tag{2.44}$$

Substituindo (2.44) em (2.35) temos

$$P(e) \geq \exp\ [-\ 2\ H(\pi)\ -\ J(\pi_1,\pi_2)]\ . \tag{2.45}$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, (2.45) se reduz a (2.42) e portanto (2.45) pode ser considerado como uma generalização da cota de KAILATH.

Da mesma forma, substituindo (2.44) em (2.38), temos

$$P(e) \geq (\frac{1}{4})\ [1\ -\ J(\pi_1,\pi_2)/2]\ . \tag{2.46}$$

Para $\pi_1=\pi_2= \frac{1}{2}$, (2.46) se reduz a

$$P(e) \geq \frac{1}{4}\ -\ \frac{J}{16}\ .$$

Cotas mais apertadas que (2.45) e (2.46) podem ser obtidas usando (2.33) do que (2.44).

Substituindo (2.43) em (2.35) temos :

$$P(e)(1-P(e)) \geq \exp\left[- 2\, H(\pi) - J(\pi_1,\pi_2)\right] . \qquad (2.48)$$

Resolvendo (2.48) para P(e), temos

$$P(e) \geq \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4 \exp\left[- 2\, H(\pi) - J(\pi_1,\pi_2)\right]} . \qquad (2.49)$$

Para $\pi_1 = \pi_2 = \frac{1}{2}$, (2.49) se reduz a:

$$P(e) \geq \frac{1}{2} - \frac{1}{2}\sqrt{1 - \exp\left(- \frac{J}{2}\right)} , \qquad (2.50)$$

onde a igualdade vale para $J = 0$ e $J = \infty$.

Do mesmo modo, substituindo (2.43) em (2.38), e solucionando para P(e), temos

$$P(e) \geq \frac{1}{2}\sqrt{- J(\pi_1,\pi_2)/8} , \qquad (2.51)$$

o qual para probabilidades iguais, se reduz a

$$P(e) \geq \frac{1}{2} - \frac{1}{4}\sqrt{J} , \qquad (2.52)$$

onde a igualdade vale quando $J = 0$.

Uma cota inferior para J em termos de P(e), a qual é mais apertada que todas os cotas inferiores acima, pode ser obtida como segue :

De (2.23), temos

$$J(X) = \left[P(C_1/X) - P(C_2/X)\right] \log\left[\frac{P(C_1/X)}{P(C_2/X)} \right]. \qquad (2.53)$$

Também de (2.24) temos que

$$P_e(X) = \min[\, P(C_1/X), P(C_2/X \,]. \qquad (2.54)$$

Como $J(X)$ é simétrica com relação a $P(C_1/X)$ e $P(C_2/X)$, pode ser expressa em termos de $P_e(X)$ como :

$$J(X) = [2\, P_e(X) - 1] \log\left[\frac{P_e(X)}{1 - P_e(X)}\right] . \qquad (2.55)$$

Consideremos, agora, a função

$$f(x) = (2x-1) \log\left(\frac{x}{1-x}\right) , \quad 0 \leq x \leq \frac{1}{2} .$$

A primeira derivada de $f(x)$ com relação a x é dada por

$$\frac{df(x)}{dx} = \frac{2x-1}{x-1} + \frac{2x-1}{x} + 2 \log\left(\frac{x}{1-x}\right) . \qquad (2.56)$$

A segunda derivada é dada por :

$$\frac{d^2 f(x)}{dx^2} = \frac{1-2x}{x^2} + \frac{4}{x} + \frac{4}{1-x} - \frac{1-2x}{(1-x)^2} . \qquad (2.57)$$

Pode ser fàcilmente mostrado que (2.57) é não-negativo. Portanto, $f(x)$ e (2.55) são funções convexas $\cup$. Para uma função convexa $\cup$, a desigualdade de Jensen é dada por :

$$E\{f(x)\} \geq f(E\{x\}) , \qquad (2.58)$$

onde E denota o valor esperado.

Obtendo o valor esperado de ambos os lados de (2.55) com relação a $P(X)$ e usando (2.58), temos o resultado dese-

jado dado por :

$$J(\pi_1,\pi_2) \geq (2P(e)-1)\ \log\left[\frac{P(e)}{1-P(e)}\right]. \qquad (2.59)$$

Para, "a priori", probabilidades iguais, (2.59) se reduz para

$$J \geq 2(2P(e)-1)\ \log\left[\frac{P(e)}{1-P(e)}\right], \qquad (2.60)$$

onde a igualdade é válida para $P(e) = 0$ e $P(e) = \frac{1}{2}$.

(2.60) é a desigualdade mais exata entre J e $P(e)$, mas tem a desvantagem de que não pode ser solucionada para $P(e)$ como uma função de J. Para uma prova de que (2.60) é mais exata que (2.50) e ((2.42), veja Apêndice B .

De (2.25) segue-se que

$$R(X) = 2\ P(C_1/X)\ P(C_2/X)\ , \qquad (2.61)$$

o qual pode ser escrito como

$$R(X) = 2\ P_e(X)(1-P_e(X))\ , \qquad (2.62)$$

onde $P_e(X)$ é dado por (2.54).

Resolvendo $P_e(X)$ temos :

$$P_e(X) = \frac{1}{2} - \sqrt{(\frac{1}{4}) - \frac{R(X)}{2}}. \qquad (2.63)$$

Pode ser fàcilmente mostrado que (2.63) é uma função de $R(X)$ convexa $\cup$ obtendo os valores esperados, com relação

a $P(X)$, de ambos os lados de (2.63); usando a desigualdade de Jensen, temos

$$P(e) \geq \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{R}{2}} \quad . \tag{2.64}$$

Substituindo (2.64) em (2.59) temos (2.40).

2.2.4. Cotas de Kullback

Os números de Kullback |21| são dados por

$$I(1,2) = \int P(X/C_1) \log\left[\frac{P(X/C_1)}{P(X/C_2)}\right] dX$$

$$I(2,1) = \int P(X/C_2) \log\left[\frac{P(X/C_2)}{P(X/C_1)}\right] dX \quad . \tag{2.66}$$

Seja $P(e_i)$ a probabilidade de erro dada a classe C_i, $i=1,2$, onde $P(e) = \pi_1 P(e_1) + \pi_2 P(e_2)$.

As cotas de Kullback são dadas por :

$$I(1,2) \geq P(e_1)\log[P(e_1)/(1-P(e_2))] + (1-P(e_1))\log[(1-P(e_1))/P(e_2)]. \quad .... \tag{2.68}$$

$$I(2,1) \geq P(e_2) \log[P(e_2)/(1-P(e_1))] + (1-P(e_2))\log[(1-P(e_2))/P(e_1)]. \quad .... \tag{2.69}$$

onde as distribuições são tais como :

$$\int_{\Omega_{X/C_1}} P(X/C_2)\, dX = \int_{\Omega_{X/C_2}} P(X/C_1)\, dX \quad , \qquad (2.69)$$

onde $\Omega_{X/C_1} \in \{ \Omega_X : P(X/C_i) > P(X/C_j) \}$, $i,j=1,2$, $i \neq j$ e Ω_X é todo o espaço característica. Então $P(e_1) = P(e_2) = P(e)$.

Somando (2.67) e (2.68), substituindo $P(e_1)=P(e_2)=P(e)$ e usando o fato de que $I(1,2) + I(2,1) = J$, temos

$$J \geq 2(2P(e)-1) \log\left[\frac{P(e)}{1-P(e)}\right] \quad , \qquad (2.70)$$

o qual é um caso especial de (2.59).

É bom saber que a condição (2.69) não é necessária para valer (2.70) em geral.

2.3. APLICAÇÕES

2.3.1. Aplicação à Medida de Equivocação de Shannon

A medida de equivocação de Shannon é a mais conhecida, e, para o problema de duas classes, é dada por :

$$H(C/X) = -\int P(X) \sum_{i=1}^{2} P(C_i/X) \log P(C_i/X)\, dX. \qquad (2.71)$$

Menos conhecido é a equivocação quadrática de VAJDA |44| dada por :

$$Q(C,X) = -\int P(X) \sum_{i=1}^{2} P(C_i/X) \left[P(C_i/X)-1\right] dX$$

$$= 1 - \int P(X) \sum_{i=1}^{2} P(C_i/X)^2 \, dX \ . \qquad (2.72)$$

Recentemente, TOUSSAINT |42|, propôs uma família de medidas de equivocação dadas por :

$$M_k(C/X) = \int P(X) \sum_{i=1}^{2} \left[P(C_i/X) - \frac{1}{2}\right]^{k^*} dX \ , \qquad (2.73)$$

onde $k^* = 2(k+1)/2k+1)$ é $k=0,1,2,...$

O que há de muito interessante aqui, é $M_o(C/X)$ dada por

$$M_o(C/X) = P(X) \sum_{i=1}^{2} \left[P(C_i/X) - \frac{1}{2}\right]^2 dX$$

$M_o(C/X)$ é relacionada a R por

$$R = \frac{1}{2} - M_o(C/X) \ .$$

O que segue :

$$M_o(C/X) = 1 - Q(C/X) \ , \qquad (2.74)$$

e

$$R = Q(C/X) \ . \qquad (2.75)$$

Assim, a medida de informação $Q(C/X)$, a qual é obtida aproximando $\log x$ por $(x-1)$ na equivocação de Shannon (2.71), e também a medida de distância (a medida harmonia entre $P(X,C_1)$ e $P(X,C_2)$) e probabilidade de erro de decisão aleatória.

Como $\log x \leq x-1$, podemos concluir que R é limitada superiormente pela equivocação de Shannon, i.e.,

$$R \leq H(C/X). \tag{2.76}$$

Substituindo (2.74) e (2.75) em cotas de R na seção 2.2.2., temos as desigualdades apertadas entre J-divergência e várias medidas de equivocação.

$$Q(C/X) \geq 2 \exp[-2H(\pi) - J(\pi_1,\pi_2)] , \tag{2.77}$$

$$Q(C/X) \geq \frac{1}{2}[1 - J(\pi_1,\pi_2)/2] , \tag{2.78}$$

e $$J(\pi_1,\pi_2) \quad 1 - Q(C/X) \log\left[\frac{1+\sqrt{1-2Q(C/X)}}{1-\sqrt{1-2Q(C/X)}}\right] \tag{2.79}$$

2.3.2. Aplicação Para Medida de Distância

Ito |17| propôs uma família de medidas de distância, chamada função Q, dada por

$$Q_n = \frac{1}{2} - \frac{1}{2}\int P(X)[P(C_1/X) - P(C_2/X)]\cdot$$

$$\cdot[P(C_1/X) - P(C_2/X)]^{n^*} dX , \tag{2.80}$$

onde $n^* = \frac{1}{2n+1}$ e n é um número natural. De grande interes se é Q_o dado por

$$Q_o = \frac{1}{2} - \frac{d}{2} , \tag{2.81}$$

onde

$$d = P(X)\left[P(C_1/X) - P(C_2/X)\right]^2 dX \ . \qquad (2.82)$$

ITO |17| mostrou que

$$Q_o = R \ , \qquad (2.83)$$

$$Q = P(e) \ , \qquad (2.84)$$

e $$Q_{n+1} \leq Q_n \ . \qquad (2.85)$$

Substituindo estes resultados pelas cotas inferiores para R , relaciona-se a função Q à divergência .

LISSACK e FU |22| investigaram a seleção de características e a estimativa probabilidade usando a medida de seperabilidade :

$$J_\alpha = \int P(X) \left| P(C_1/X) - P(C_2/X) \right|^\alpha dX \ , \ \alpha > 0 \ . \qquad (2.86)$$

Pode ser fàcilmente mostrado que

$$P(e) = \frac{1}{2} - \frac{J_1}{2} \ , \qquad (2.87)$$

e $$R = \frac{1}{2} - \frac{J_2}{2} \ . \qquad (2.88)$$

Portanto, substituindo (2.87) e (2.88) nos resultados de seções 2.2.2 e 2.2.3, relaciona-se J_α a J-divergência.

DEVIJVER |19| também fez muitos trabalhos conhecidos

como distância Bayesiana dada por

$$B(C/X) = \int P(X) \sum_{i=1}^{2} P(C_i/X)^2 \, dX \, . \qquad (2.89)$$

É óbvio que

$$B(C/X) = 1 - R \, . \qquad (2.90)$$

Portanto, usando os resultados da seção 2.2.2., temos desigualdades precisas entre a distância Bayesiana e a divergência. Por exemplo, deixando B indicar B(C/X), para simplificar a notação e substituindo (2.90) em (2.41) temos

$$J \geq 2\sqrt{2B-1} \quad \log\left[\frac{1+\sqrt{2B-1}}{1-\sqrt{2B-1}}\right]. \qquad (2.91)$$

2.4. COMENTÁRIOS FINAIS

Foi mostrado que a probabilidade de erro da regra de decisão aleatória de predição proporcional é equivalente a taxa de erro da regra determinística do vizinho mais próximo assintòticamente. Prèviamente, nenhum dos valores foram disponíveis para R e a divergência J. Nesta seção, melhores cotas inferiores foram dadas para R e J. A cota mais apertada é dada por (2.41). Para a avaliação característica usando J, (2.36) e (2.39) são mais úteis.

Sejam

$$R_1 = \frac{1}{2} \exp\ (-\frac{J}{2})$$

e $$R_2 = \frac{1}{2} - \frac{J}{8} \quad .$$

A melhor cota recomendada para uso futuro é

$$R \geq \max\ \{R_1,\ R_2\} \quad .$$

Comentários idênticos são válidos para P(e).

$$P(e) \leq \frac{1}{2}\ (\frac{J}{4})^{\frac{1}{4}} \quad . \qquad (2.92)$$

Sejam

$$L_1 = \frac{1}{2} - \frac{1}{2}\sqrt{1 - \exp(-\frac{J}{2})} \quad ,$$

e $$L_2 = \frac{1}{2} - \frac{1}{4}\ J$$

de (2.50) e (2.52).

A melhor cota inferior disponível para complementar (2.92) acima, é dada por

$$P(e) \geq \max\ \{L_1,\ L_2\} \quad ,$$

a qual é muito superior à cota prévia disponível (2.42).

APÊNDICE A

A equação (2.39) pode ser descrita na forma

$$J \geq 4(1-2R). \qquad (A1)$$

Para mostar que (2.41) é mais precisa que (2.39), deve ser provado que

$$2\sqrt{1-2R}\ \log\left[\frac{1+\sqrt{1-2R}}{1-\sqrt{1-2R}}\right] \geq 4(1-2R). \qquad (A2)$$

Fazendo-se uso da transformação

$$|1 - 2R|^{\frac{1}{2}} = x\ , \quad 0 \leq x \leq 1,$$

em (A2), temos

$$2x \log \frac{1+x}{1-x} \geq 4x^2\ ,$$

ou

$$\log \frac{1+x}{1-x} \geq 2x\ . \qquad (A3)$$

Sabemos que

$$\log\left[\frac{1+x}{1-x}\right] = 2\sum_{i=1}^{\infty}\left[\frac{1}{2k-1}\right]x^{2k-1}\ , \qquad (A4)$$

para $x^2 > 1$. Desde que $x \geq 0$, todos os termos em (A4) são não negativos e segue-se que, para $k=1$, (A4) reduz-se para (A3), provando o resultado.

A equação (2.36) pode ser escrita na forma

$$J \geq -2 \log(2R), \qquad (A5)$$

Para mostrar que (2.41) é melhor que (2.36). Deve valer que

$$2x \log \frac{1+x}{1-x} \geq -2 \log(1-x^2) , \qquad (A6)$$

onde x é como acima.

Usando-se a transformação $x = 2y-1$, $\frac{1}{2} \leq y \leq 1$. Deve valer que

$$2(2y-1) \log \frac{y}{1-y} \geq -2 \log \frac{4y}{1-y} . \qquad (A7)$$

Expandindo-se (A7) e recombinando-se os termos, resulta em

$$H(y,1-y) \leq \log 2 ,$$

onde $H(y,1-y)$ é a função entropia.

O máximo de $H(y,1-y)$ ocorre para $y=\frac{1}{2}$ e é dado pelo $\log 2$, provando então, o resultado desejado.

APÊNDICE B

A equação (2.50) pode ser escrita na forma

$$J \geq -2 \log [4 P(e)(1-P(e))]. \qquad (B1)$$

Para provar que (2.60) é melhor que (2.50), deve ser mostrado que

$$2(2P(e)-1) \log\left[\frac{P(e)}{1-P(e)}\right] \geq -2 \log\left[\frac{4P(e)}{1-P(e)}\right],$$

é da mesma forma que (A7), provando então o resultado.

A equação (2.52) pode ser escrita na forma

$$J \geq 4(1-P(e))^2 . \qquad (B2)$$

Para provar que (2.60) é melhor que (2.52), deve ser mostrado que

$$\log \frac{1-x}{x} \geq 2(1-2x) , \qquad (B3)$$

para $0 \leq x \leq \frac{1}{2}$.

Usando a trasformação $x = \frac{1}{z+1}$ com $1 \leq z \leq \infty$, deve ser mostrado que

$$\log z \geq 2\left(\frac{z-1}{z+1}\right). \qquad (B4)$$

É sabido que, para $z > 0$,

$$\log z = 2 \sum_{k=1}^{\infty} \left[\frac{1}{2k-1}\right] \left[\frac{z-1}{z+1}\right]^{2k+1} . \qquad (B5)$$

Para $z \geq 1$, todos os termos em (B5) são não-negativos.

Portanto, para $k=1$, (B5) se reduz a (B4), provando o resultado.

CAPÍTULO III

CARACTERIZAÇÃO DE GENERALIZAÇÕES DE J-DIVERGÊNCIA

3.1. DISCRIMINAÇÃO DE KULLBACK E SUAS GENERALIZAÇÕES

KULLBACK |21| introduziu uma medida de informação ou informação discriminação ou informação relativa entre duas distribuições de probabilidades, dada por

$$I(P:Q) = \sum_{i=1}^{M} p_i \log_2 \frac{p_i}{q_i} \quad , \tag{3.1}$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

Esta medida é uma generalização de medida de Shannon, isto é, a entropia de Shannon dada por

$$H(P) = - \sum_{i=1}^{M} p_i \log_2 p_i \quad , \tag{3.2}$$

para todo $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$.

Uma comparação entre estas duas medidas (3.1) e (3.2) pode ser encontrada em HOBSON e CHENG |15| onde uma importância lógica da medida (3.1) foi discutida.

A medida de Kullback (3.1) satisfaz a aditividade

$$I(P*R:Q*S) = I(P:Q) + I(R:S) \quad , \tag{3.3}$$

para todo $P, Q \in \Delta_M$, $R, S \in \Delta_N$ e $P*R, Q*S \in \Delta_{MN}$.

Dependente desta propriedade básica, uma caracterização de medida (3.1) foi estudada por KANNAPPAN |30| e dada por

$$I(P:Q) = \sum_{i=1}^{M} f(p_i, q_i), \qquad (3.4)$$

onde $f(p_i, q_i)$ é uma solução contínua da seguinte equação funcional

$$\sum_{i=1}^{M} \sum_{j=1}^{N} f(p_i r_j, q_i s_j) = \sum_{i=1}^{M} f(p_i, q_i) + \sum_{j=1}^{N} f(q_j, s_j). \qquad (3.5)$$

SHARMA e AUTAR |28| , TANEJA |33| através da equação funcional e SHARMA e TANEJA |32| através de axiomas, incluindo a propriedade recursiva, apresentaram a seguinte medida de grau β :

$$I^{\beta}(P:Q) = (2^{\beta-1} - 1)^{-1} \{ \sum_{i=1}^{M} p_i^{\beta} q_i^{1-\beta} - 1 \}, \quad \beta \neq 1, \beta > 0,$$

... (3.6)

a qual, em lugar de (3.3), satisfaz a não aditividade

$$I^{\beta}(P*R:Q*S) = I^{\beta}(P:Q) + I^{\beta}(R:S) + (2^{\beta-1}-1)\, I^{\beta}(P:Q) I^{\beta}(R:S).$$

... (3.7)

É óbvio que, quando $\beta \to 1$, $I^{\beta}(P:Q)$ se reduz a (3.1).

Por outro lado RÉNYI |26|, considerando identidade (3.3) e a properiedade do valor médio, mostrou que

$$I_{\phi}(P:Q) = \phi^{-1} \left(\sum_{i=1}^{M} p_i \, \phi(I(p_i:q)) \right) , \qquad (3.8)$$

onde ϕ é uma função estritamente monótona e contínua e $I(p_i:q_i)$ é a informação dos eventos únicos. RÉNYI |26| apresentou a seguinte medida de ordem α :

$$I_{\alpha}(P:Q) = \frac{1}{\alpha-1} \log_2 \left(\sum_{i=1}^{M} p_i^{\alpha} \, q_i^{1-\alpha} \right), \; \alpha \neq 1, \alpha > 0, \qquad (3.9)$$

onde $I_{\alpha}(P:Q)$ é uma generalização aditiva do valor médio sobre a medida de Kullback (3.1).

SHARMA e MITTAL |30| estudaram uma outra generalização não-aditiva do valor médio em cima das medidas (3.6) e (3.9), usando não-aditividade (3.7) e a propriedade do valor médio (3.8) e apresentaram a seguinte medida

$$I_{\alpha}^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1} \left\{ \left(\sum_{i=1}^{M} p_i^{\alpha} \, q_i^{1-\alpha} \right)^{\frac{\beta-1}{\alpha-1}} - 1 \right\} , \qquad (3.10)$$

$$\alpha \neq 1, \; \beta \neq 1, \; \alpha, \beta > 0.$$

Esta medida (3.10) contém todas as medidas apresentadas acima, como casos particulares.

3.2. J-DIVERGÊNCIA E SUAS GENERALIZAÇÕES

Kullback também apresentou a seguinte generalização da informação relativa (3.1) chamada como J-divergência :

$$J(P:Q) = \sum_{i=1}^{M} p_i \log_2 \frac{p_i}{q_i} + \sum_{i=1}^{M} q_i \log_2 \frac{q_i}{p_i} \quad . \qquad (3.11)$$

Esta medida tem várias aplicações em estatística (ref. KULLBACK |21|) e em reconhecimento de modelos (Cap.II),

(3.11) também pode ser escrita como

$$J(P:Q) = I(P:Q) + I(Q:P) \quad , \qquad (3.12)$$

onde $I(P:Q)$ é como foi dado em (3.1).

RATHIE e SHENG |25| estudaram a J-divergência de grau β

$$J^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\{\sum_{i=1}^{M} p_i^{\beta} q_i^{1-\beta} + \sum_{i=1}^{M} p_i^{1-\beta} q_i^{\beta} - 2\}, \qquad (3.13)$$

$$= I^{\beta}(P:Q) + I^{\beta}(Q:P) \quad , \ \beta \neq 1, \beta > 0. \qquad (3.14)$$

Neste Capítulo, apresentamos a seguinte generalização não-aditiva de (3.11) e (3.13):

$$J_{\alpha}^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\{(\sum_{i=1}^{M} p_i^{\alpha} q_i^{1-\alpha})^{\frac{\beta-1}{\alpha-1}} + (\sum_{i=1}^{M} p_i^{1-\alpha} q_i^{\alpha})^{\frac{\beta-1}{\alpha-1}} - 2\} \quad , \qquad (3.15)$$

$$\alpha \neq 1, \beta \neq 1, \ \alpha, \beta > 0.$$

A medida (3.15) pode ser escrita como :

$$J_{\alpha}^{\beta}(P:Q) = I_{\alpha}^{\beta}(P:Q) + I_{\alpha}^{\beta}(Q:P) \quad , \qquad (3.16)$$

onde $I_{\alpha}^{\beta}(P:Q)$ é dado como em (3.10).

Quando $\alpha=\beta$, (3.15) se reduz para (3.13).

Caracterizando $I^{\beta}(P:Q)$ e definindo $I^{\beta}_{\alpha}(P:Q)$ como dado em (3.16), obtemos uma caracterização de (3.15).

A medida (3.15) também satisfaz a não-aditividade (3.7).

Apresentamos na seção 3.3 uma caracterização da medida (3.10).

Também existem as seguintes generalizações de $J(P:Q)$ e $J^{\beta}(P:Q)$:

$$J_{\alpha}(P:Q) = \frac{1}{\alpha-1} \log_2 \sum_{i=1}^{M} (p_i^{\alpha} q_i^{1-\alpha} + p_i^{1-\alpha} q_i^{\alpha}), \quad \alpha \neq 1, \ \alpha > 0, \qquad \ldots \quad (3.17)$$

e

$$J^{\beta}_{\alpha}(P:Q) = (2^{\beta-1}-1)^{-1} \left\{ \sum_{i=1}^{M} (p_i^{\alpha} q_i^{1-\alpha} + p_i^{1-\alpha} q_i^{\alpha})^{\frac{\beta-1}{\alpha-1}} - 2 \right\},$$

$$\alpha \neq 1, \ \beta \neq 1, \ \alpha, \beta > 0. \quad (3.18)$$

É fácil provar que

$$\lim_{\beta \to 1} J^{\beta}_{\alpha}(P:Q) = J_{\alpha}(P:Q) \ ,$$

e, quando $\alpha=\beta$, $J^{\beta}_{\alpha}(P:Q)$ reduz a $J^{\beta}(P:Q)$.

Vamos denotar

$$W_{\alpha}(P:Q) = \sum_{i=1}^{M} (p_i^{\alpha} q_i^{1-\alpha} + p_i^{1-\alpha} q_i^{\alpha}) \ . \qquad (3.19)$$

Então (3.13), (3.17) e (3.18) respectivamente, podem ser escritos como

$$J^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\{ W_{\beta}(P:Q) - 2\}, \beta \neq 1, \beta > 0, \quad (3.20)$$

$$J_{\alpha}(P:Q) = \frac{1}{\alpha-1} \log W_{\alpha}(P:Q), \ \alpha \neq 1, \ \alpha > 0 \quad (3.21)$$

e $$J_{\alpha}^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\{ W_{\alpha}(P:Q)^{\frac{\beta-1}{\alpha-1}} - 2\}, \quad (3.22)$$

$$\alpha \neq 1, \beta \neq 1, \quad \alpha, \beta > 0.$$

É importante notar que $W_{\alpha}(P:Q)$ é uma chave principal nas três medidas citadas acima.

Mais detalhes destas medidas podem ser encontrados em |37|.

Apresentamos agora, uma caracterização da medida não-aditiva da informação relativa, $I_{\alpha}^{\beta}(P:Q)$, que resulta numa caracterização da medida $J_{\alpha}^{\beta}(P:Q)$ (3.15).

3.3. MEDIDA NÃO-ADITIVA DA INFORMAÇÃO RELATIVA

Seja a ocorrência de evento x, associado a duas probabilidades p e q. No teorema seguinte, caracterizamos I(P:Q), como a informação relativa entre p e q sobre a não-aditividade e propriedade do valor médio.

TEOREMA 3.1.

Se a função contínua $I(P:Q)$ satisfaz os seguintes axiomas :

A_1: $I(pr:qs) = I(P:q) + I(r:s) + (2^{\beta-1}-1)\, I(p:q)I(r:s)$, ... (3.23)

A_2: $I(1:\frac{1}{2}) = 1$; (3.24)

A_3: $I(p:p) = 0$; (3.25)

então

$$I(p:q) = I^{\beta}(p:q) = (2^{\beta-1}-1)^{-1}\{(\frac{p}{q})^{\beta-1} - 1\},\ \beta \neq 1. \quad (3.26)$$

Para demonstrar este teorema, basta ver a referência SHARMA e MITTAL |30|.

O teorema seguinte dá uma caracterização das medidas $I_1^{\beta}(P:Q)$ e $I_{\alpha}^{\beta}(P:Q)$.

TEOREMA 3.2.

A informação relativa $I_{\phi}(P:Q)$ contínua e não-aditiva, dada em (3.8) da distribuição $P=(p_1,p_2,\ldots,p_M) \in \Delta_M$ com respeito a $Q=(q_1,q_2,\ldots,q_N) \in \Delta_N$, satisfazendo a não aditividade (3.7) e $I(P:P) = 0$, pode ser escrita apenas das duas formas seguintes :

$$I_1^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\left[\, 2^{(\beta-1)\sum_{i=1}^{M} p_i \log_2(\frac{p_i}{q_i})} - 1 \right],\ \beta \neq 1, \quad (3.27)$$

e

$$I_{\alpha}^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\left[\, (\sum_{i=1}^{M} p_i^{\alpha}\, q_i^{1-\alpha})^{\frac{\beta-1}{\alpha-1}} - 1 \right],\ \beta\neq 1,\ \alpha\neq 1. \quad (3.28)$$

Demonstração :

Para $P = \{p\}$, $Q = \{q\}$ $R = \{r\}$ e $S = \{s\}$, (3.7) se reduz para (3.23) e portanto pelo Teorema 3.1, temos

$$I(p:q) = \left[\left(\frac{p}{q}\right)^{\beta-1} - 1\right] \left(2^{\beta-1}-1\right)^{-1}, \quad \beta \neq 1, \ \beta > 0. \qquad (3.29)$$

Vamos considerar agora $R = \{r\}$, $0 < r \leq 1$ e $S = \{s\}$, $0 < s \leq 1$. Então, (3.7) torna-se

$$I(P*r:Q*s) = I(P:Q) + I(r:s) + \left(2^{\beta-1}-1\right) I(P:Q)I(r:s), \qquad (3.30)$$

onde

$$P*r = (p_1 r, p_2 r, \ldots p_M r)$$

e $$Q*s = (q_1 s, q_2 s, \ldots, q_N s).$$

Usando (3.29) e (3.8) em (3.30), obtemos

$$\phi^{-1}\left\{\sum_{i=1}^{M} p_i \phi\left[\left(\frac{p_i r}{q_i s}\right)^{\beta-1} - 1\right] \left(2^{\beta-1}-1\right)^{-1}\right.$$

$$= \phi^{-1}\left\{\sum_{i=1}^{M} p_i \phi\left[\left(\frac{p_i}{q_i}\right)^{\beta-1} - 1\right]\left(2^{\beta-1} - 1\right)^{-1}\right\}\left(\frac{r}{s}\right)^{\beta-1}$$

$$\left. + \left[\left(\frac{r}{s}\right)^{\beta-1} - 1\right]\left(2^{\beta-1} - 1\right)^{-1}\right.,$$

ou $$\psi_{r/s}^{-1}\left\{\sum_{i=1}^{M} p_i \psi_{r/s}\left(\left[\left(\frac{p_i}{q_i}\right)^{\beta-1} - 1\right] \left(2^{\beta-1}-1\right)^{-1}\right)\right\}$$

$$= \phi^{-1}\left\{\sum_{i=1}^{M} p_i \phi\left(\left[\left(\frac{p_i}{q_i}\right)^{-1} - 1\right] \left(2^{\beta-1} - 1\right)^{-1}\right)\right\}, \qquad (3.31)$$

Por HARDY, LITTLEWOOD e PÓLYA |12|, existe uma relação linear entre $\psi_{r/s}$ e ϕ,

$$\psi_{r/s}((x^{\beta-1}-1)(2^{\beta-1}-1)^{-1}) = A(y)\ \phi((x^{\beta-1}-1)(2^{\beta-1}-1)^{-1}) + B(y)\ , \qquad (3.33)$$

onde $x = \frac{p_i}{q_i}$ e $y = \frac{p}{q}$ isto é,

$$\phi\left(\frac{x^{\beta-1}\ y^{\beta-1}\ -1}{2^{\beta-1}-1}\right) = A(y)\ \phi\left(\frac{x^{\beta-1}-1}{2^{\beta-1}-1}\right) + B(y)\ , \qquad (3.34)$$

isto é,

$$g(xy) = A(y)\ g(x) + B(y)\ , \qquad (3.35)$$

onde

$$g(x) = \phi\left(\frac{x^{\beta-1}\ -1}{2^{\beta-1}-1}\right)\ , \qquad (3.36)$$

ou $$G(xy) = A(y)\ G(x) + G(y), \qquad (3.37)$$

onde

$$G(x) = g(x) - g(1), \qquad (3.38)$$

para qualquer x real.

Da simetria de $G(x,y)$ em x e y, temos

$$G(xy) = G(yx),$$

Isto é,

$$A(y)\ G(x) + G(y) = A(x)\ G(y) + G(x)\ , \qquad (3.39)$$

Isto é,

$$G(x)\left[\ A(y) - 1\right] = G(y)\left[\ A(x) - 1\right]. \qquad (3.40)$$

Observe que existem dois casos :

i) Quando $A(x) - 1 = 0$ e

ii) Quando $A(x) - 1 \neq 0$

i) Agora, quando $A(x) - 1 = 0$, $A(x) = 1$, (3.40) se reduz a

$$G(xy) = G(x) + G(y). \qquad (3.41)$$

A solução geral da equação (3.41) (ref. ACZÉL e DARÓCZY [1]) é dada por

$$G(x) = A \log x, \quad x > 0$$

onde A é uma constante arbitrária.

Considerando $g(1) = a$ e usando (3.36), obtemos

$$\phi(x) = a + \frac{A}{\beta-1} \log \left[1 + x(2^{\beta-1} - 1)\right]\ , \ \beta \neq 1. \qquad (3.42)$$

ii) Se $A(x) - 1 \neq 0$, se (3.39) temos

$$\frac{G(x)}{A(x)-1} = \frac{G(y)}{A(y)-1} = \frac{1}{k}\ , \qquad (3.43)$$

ou $A(x) - 1 = k\, G(x)$, (3.44)

para todo x real.

A aplicação desta relação em (3.37) dá

$$A(xy) = A(x)\, A(y) \ . \qquad (3.45)$$

A solução geral da equação (3.44) é dada por

$$A(x) = x^{\alpha-1} \ ,$$

para todo x real.

Com este valor de $A(x)$ e usando (3.43), (3.13) e (3.11) obtemos

$$\phi(x) = a + \frac{\left[(2^{\beta-1} - 1)x + 1\right]^{\frac{\alpha-1}{\beta-1}} - 1}{k} \ , \quad \alpha \neq 1, \beta \neq 1. \qquad (3.46)$$

Com estes valores de $\phi(x)$ dado em (3.43) e (3.46), obtemos (3.27) e (3.28), respectivamente.

3.3.1. J-Divergência de Ordem α e Grau β

Definimos

$$J_\alpha^\beta(P:Q) = I_\alpha^\beta(P:Q) + I_\alpha^\beta(Q:P), \qquad (3.47)$$

e usando (3.28) em (3.46), obtemos a seguinte medida de informação chamada J-divergência de ordem α e grau β :

$$J_{\alpha}^{\beta}(P:Q) = (2^{\beta-1}-1)^{-1}\left\{\left(\sum_{i=1}^{M} p_i^{\alpha} q_i^{1-\alpha}\right)^{\frac{\beta-1}{\alpha-1}} + \left(\sum_{i=1}^{M} p_i^{1-\alpha} q_i^{\alpha}\right)^{\frac{\beta-1}{\alpha-1}} - 2\right\},$$

a qual é dada como em (3.18).

BIBLIOGRAFIA

|1|. ACZÉL, J e Z. DARÓCZY : On Measures of Information and Their Characterizations - Academic Press, New York, 1975.

|2|. ARIMOTO, S. : Information-Theoretic Considerations on Estimation Problems - Information and Control, Vol. 19(1971), 181-194.

|3|. BEN-BESSAT e J. RAVIV : Rényi's Entropy and the Probability of Error - IEEE Trans. on Inform. Theory, Vol.IT-24(1978), 324-331.

|4|. BOEKEE, D.E. e J.C.A. VAN DER LUBBE : The R-Norm Information Measure - Information and Control, Vol.45 (1980), 136-155.

|5|. CHEN, C.H. : Statistical Pattern Recognition - Hayden Book Co., Inc., 1973.

|6|. CHITTI, BABU C. : On Divergence and Probability of Error in Pattern Recognition - Proc. IEEE, June 1973, 798-799.

|7|. COVER, T.M. e P.E. HART : Nearest Neighbour Pattern Classification - IEEE Trans. on Inform. Theory, Vol. IT-13(1967), 21-27.

|8|. DARÓCZY, Z. : Generalized Information Functions - Information and Control, 16(1970), 36-51.

|9|. DEVIJVER, P.A. : On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition -

IEEE Trans. on Computers, Vol. C-23(1974), 70-80.

|10|. DEVIJVER, P.A. : Entropies of Degree and Lower Bounds for the Average Error Rate - Information and Control, Vol. 34, Nº 3(1977), 222-226.

|11|. GALLAGER, R.G. : Information Theory and Reliable Communications , New York, Wiley, 1968.

|12|. HARDY, G.H., J.E. LITTLEWOOD e G. PÓLYA : Inequalities , Cambridge University Press, 1952.

|13|. HAVRDA, J. e F. CHARVAT : Quantification Method of Classification Process : The Concept of Strutural a-Entropy - Kybernetika, 3(1967), 30-35.

|14|. HELLMAN, M.E. e J. RAVIV : Probability of Error, Equivocation and Chernoff Bound - IEEE Trans. on Information Theory, Vol.IT-16(1970), 368-372.

|15|. HOBSON, A. e B.K. CHENG : A Comparison of Shannon and Kullback Information Measures - J. Statist. Phys. Vol, 7(1973), 301-310.

|16|. HORIBE, Y. : On Zero Error Probability of Binary Decisions - IEEE Trans. on Inform. Theory, Vol. IT-16(1970), 347-348.

|17|. ITO, T. : On Approximate Error Bounds in Pattern Recognition - Systems, Computers, Controls, Vol.4(1973), 85-92.

|18|. KAILATH, T. : The Divergence and Bhattacharya Distance Measure in Signal Selection - IEEE Trans. on Communication Technology, Vol.COM-15(1967), 52-60.

|19|. KANNAPPAN, PL. : On Shannon's Entropy, Directed-Divergence, and Inaccuracy - Z. Wahrs. und Verw Geb., 22(1972), 95-100.

|20|. KERRIDGE, D.F. : Inaccuracy and Inference - J. Royal Stat. Soc., Ser. B, 23(1961), 184-194.

|21|. KULLBACK, S. : Information Theory and Statistics - Dover Publications, Inc., New York, 1968.

|22|. LISSAK, T. e K.S. FU : A Separability Measure for Feature Selection and Error Estimation in Pattern Recognition - Tech. Rep. Nº TR-EE, 72-15(1972), School of Electrical Engineering , Perdue University.

|23|. MATHAI, A.M. e P.N. RATHIE : Basic Concepts in Information Theory and Statistics, Wiley Eastern Limited, New Delhi, 1975.

|24|. MATUSITA, K. : A Distance and Related Statistics in Multivariate Analysis - Multivariate Analysis, P.R. Kirishnaiah Ed., New York, Academic Press, 1966, 187-200.

|25|. RATHIE, P.N. e LILLIAN T. SHENG : The J-Divergence of Order - J. Comb. Inform. and System Sci., Vol.6(1981), 197-205.

|26|. RÉNYI, A. : On Measures of Entropy and Information - Proc. 4th Berkeley Symp. Math. Statist. Prob., 1(1961), 547-561.

|27|. SHANNON, C.E. : A Mathematical Theory of Communication, Bell System Tech. J. ,27(1948), 379-423, 623-658

|28|. SHARMA, B.D. e R. AUTAR : Relative Information Functions and Their Type (α,β) Generalizations - METRIKA, 21(1974), 41-50.

|29|. SHARMA, B.D. e D.P. MITTAL : New Non Additive Measures of Entropy for Discrete Probability Distributions - J. Math. Sci., 10(1975), 28-40.

|30|. SHARMA, B.D. e D.P. MITTAL : New Non Additive Measures of Relative Information - J. Comb. Inform. and Syst. Sci., Vol.2, Nº 4(1977), 122-132.

|31|. SHARMA, B.D. e I.J. TANEJA : On Axiomatic Characterization of Information-Theoretic Measures - J. Stat. Phys., Vol.10, Nº4(1974), 337-346.

|32|. SHARMA, B.D. e I.J. TANEJA : Entropy of Type (α,β) and other Generalized Measures in Information Theory - METRIKA, 22(1975), 205-215.

|33|. TANEJA, I.J. : On Measures of Information and Inaccuracy - J. Statist. Phys., Vol.14, Nº 3(1976), 263-270.

|34|. TANEJA, I.J. : Comentário sobre o paper "Entropies of Degree β and Lower Bounds for the Average Error Rate - IEEE Trans. Syst. Man and Cybernetics, Feb/March, 1983.

|35|. TANEJA, I.J. : Comentário sobre o paper "A Generalization of Shannon's Equivocation and the Fano Bound - IEEE Trans. Syst. Man and Cybernetics, May/June, 1983

|36|. TANEJA, I.J. : Generalized η - Entropy and Probability of Error - Proc. International Conference on Cybernetics and Society, Seattle, Washington, Oct. 28-30, 1982, 463-466.

|37|. TANEJA, I.J. : Characterization of J-Divergence and Their Generalizations - Comunicada.

|38|. TANEJA, I.J. : Generalizations of J-Divergence and Their Applications to Pattern Recognition - Comunicada.

|39|. TANEJA, I.J. e ALBERTINA ZATELLI : Generalized Parametric Entropy and Probability of Error - International Symp. on Information Theory, Les Arcs, França, 1982.

|40|. THEIL, H. : Economics and Information Theory - North Holland Pub. Co., Amsterdam, 1967.

|41|. TOUSSAINT, G.T. : Some Inequalities Between Distance Measures for Feature Evaluation - IEEE Trans.on Computers, April 1972, 409-410.

|42|. TOUSSAINT, G.T. : On the Divergence Between Two Distributions and the Probability of Misclassification of Several Decision Rules - Proc. 2nd International Joint Conference on Pattern Recognition , Copenhagen, August 1974, 1-8.

|43|. TOUSSAINT, G.T. : A Generalization of Shannon's Equivocation and the Fano Bound - IEEE Trans. on System, Man and Cybernetics, 1977.

|44|. VAJDA, I. : A Contribution to the Informational Analysis of Pattern in Methodologies of Pattern Recognition, Ed., S. Watanabe, Academic Press, 1969.