ENTROPIAS DE SHANNON E RÉNYI APLICADAS AO RECONHECIMENTO DE PADRÕES

SHANNON AND RÉNYI'S ENTROPY APPLIED TO PATTERN RECOGNIZE

Alysson Ramos Artuso
Universidade Federal do Paraná – PPGMNE/UFPR
Avenida Coronel Francisco Heráclito dos Santos, 210, Jardim das Américas
Centro de Estudos de Engenharia Civil - Curitiba - PR, 81531-970
E-mail: alysson.artuso@gmail.com

RESUMO

Esse artigo teórico apresenta e discute os conceitos de entropia e informação mútua da Teoria da Informação aplicados ao reconhecimento de padrões partindo, principalmente, das ideias de Shannon e Rényi. Tendo por base a diversidade de aplicação da teoria, como nas áreas de engenharia, estatística, economia e informática e a escassez de bibliografia em português, expõe-se uma maneira de utilizar a entropia aplicada à pesquisa, com uma aplicação mostrada para a construção de árvore de decisão. Também são debatidas similaridade e diferenças com o conceito físico de entropia.
Palavras-chave: Teoria da Informação, Entropia de Shannon, Entropia de Rényi, Árvores de Decisão.

ABSTRACT

This paper presents and discusses the concepts of entropy and mutual information in Information Theory applied to pattern recognize, based mainly on the ideas of Shannon and Rényi. Based on the diversity of application of the theory, as in engineering, statistics, economics and informatics and the few literature in Portuguese, is exposed a way of using the entropy applied to the academic researches, with an application showed for the construction of decision tree. Are also discussed similarities and differences with the physical concept of entropy.
Keywords: Information Theory, Shannon's Entropy, Rényi's Entropy, Decision Tree.

1. INTRODUÇÃO

Claude E. Shannon (1948) foi um pioneiro ao considerar a comunicação como um problema matemático rigorosamente embasado na estatística, criando um ramo da teoria da probabilidade e da estatística chamado Teoria da Informação. Apesar de ser originalmente desenvolvida para informações perdidas na compressão e transmissão de mensagens com ruídos em um canal de comunicação, sua aplicabilidade se expandiu para outros domínios da engenharia, informática, estatística e economia.

Entretanto, sua similaridade com o conceito físico de entropia e seu uso em áreas diversas das quais foi pensada pode ocasionar alguns equívocos teóricos e metodológicos. Soma-se a isso a escassa literatura existente em português, em especial sobre a entropia de Rényi. Por isso, apresenta-se um breve apanhado sobre a teoria da informação, partindo do conceito de Shannon e comparando-o com a grandeza física entropia. A seguir, mostra-se a generalização feita por Rényi e os desenvolvimentos até se obter os estimadores das medidas de entropia e informação mútua. Posteriormente, os conceitos básicos da Teoria da Informação são exemplificados numa aplicação de Árvores de Decisão com o objetivo de reconhecimento de padrões.

2. TEORIA DA INFORMAÇÃO

Na área de reconhecimento de padrões, o interesse se volta para a Teoria da Informação pela sua capacidade de identificação de variáveis relevantes e utilização em métodos classificatórios. Isso se dá, a princípio, por meio de dois conceitos nomeados por Shannon: entropia e informação mútua. Nesse contexto, a entropia funciona como uma medida de incerteza de variáveis aleatórias isoladas ou combinadas. A informação mútua refere-se à dependência estocástica entre variáveis aleatórias, quantificando a informação comum entre elas.

2.1. Origem da entropia

A entropia é um termo que originalmente se refere a um conceito físico termodinâmico. Num primeiro momento, remete aos trabalhos do físico alemão Rudolf Clausius na segunda metade do século XIX. Em termos modernos, a entropia (S), como desenvolvida por Clausius, é uma função de estado relacionada com a passagem de calor (Q) e a temperatura (T), sendo expressa por:

$$dS = \frac{dQ}{T} \tag{1}$$

A opção pelo nome entropia, que anteriormente havia sido chamada de "calor perdido irreversivelmente" e "valor equivalente", se deve a origem grega da palavra e sua similaridade com energia. Mesmo não se referindo à mesma ideia, as duas estão profundamente relacionadas, de forma que Clausius achou ser apropriado ter também certa similaridade nos nomes (LAIDER, 1995).

Em 1862, Clausius enunciou o teorema que diz que a soma algébrica de todas as variações de entropia das transformações ocorridas em um processo cíclico é sempre positiva ou, em processos reversíveis, nula. Ou seja, $\Delta S \geq 0$, que corresponde à 2ª Lei da Termodinâmica.

Com base nos trabalhos de Clausius, Von Helmholtz e outros, o físico estadunidense Josiah Williard Gibbs propôs, em 1876, uma medida de "energia disponível" (ΔG – *free energy*) que seria matematicamente calculada subtraindo-se a "energia perdida" $T\Delta S$ da energia total ΔH. Isso implica que a energia disponível do Universo se reduz, visto que a variação de entropia é positiva ou, em casos extremos, nula.

Em 1877, uma definição alternativa de entropia, mas equivalente à anterior, foi formulada pelo austríaco Ludwig Boltzmann, definindo-a usualmente como

$$S = k_B \log \Omega \tag{2}$$

onde k_B é a constante de Boltzmann e Ω é o número de microestados que geram o macroestado observado.

Para compreender mais profundamente a relação da eq. 2 com probabilidades, cabe fazer um breve desenvolvimento da expressão. Primeiro, é preciso esclarecer alguns conceitos. Um macroestado é dado em função de propriedades macroscópicas da matéria (como pressão, temperatura e volume). Um microestado X é especificado por meio das 3 coordenadas de posição e das 3 coordenadas de momento de cada partícula N (indistinguível) que compõe o sistema estudado. O espaço de fase W é o espaço de 6N dimensões de todas as possibilidades de microestados e W_E é a região de W que consiste em todos os microestados com energia constante E.

Particionando o espaço de fase W_E em células ω_1, ω_1, ..., ω_k de tamanho $(\delta\omega)^N$, há uma distribuição de partículas, dada pela quantidade delas em cada célula. Mas diversas formas de arranjar as partículas nas células podem corresponder a mesma distribuição D, uma vez que as partículas são indistinguíveis. A quantidade de arranjos $G(D_i)$ compatíveis com a distribuição D_i é dada por

$$G(D_i) = \frac{N!}{n_1! n_2! \ldots n_k!} \tag{3}$$

onde n_1, n_2, ..., n_k são o número de partículas presentes nas células ω_1, ω_1, ..., ω_k.

Cada distribuição D_i corresponde a um macroestado Ω_{Di}, cujo tamanho pode ser dado pelo número de arranjos compatíveis com D_i multiplicado pelo tamanho da célula correspondente, ou seja:

$$\Omega_{Di} = G(D_i).\delta\omega^N \tag{4}$$

e da entropia de Boltzmann:

$$S\left(\Omega_{D_i}\right) = k_B \log\left(G\left(D_i\right)\delta\omega^N\right)$$

$$S\left(\Omega_{D_i}\right) = k_B \log\left(G\left(D_i\right)\right) + k_B \log\left(\delta\omega^N\right)$$

$$S\left(\Omega_{D_i}\right) = k_B \log\left(\frac{N!}{n_1!n_2!\ldots n_k!}\right) + N\log\left(\delta\omega\right)$$

$$S\left(\Omega_{D_i}\right) = k_B \log\left(N!\right) - k_B \log\left(n_1!\right) - \ldots - k_B \log\left(n_k!\right) + \text{const}$$

usando a aproximação de Stirling log n! $\cong$ n log (n) – n e o fato de que $N = n_1 + n_2 + ... + n_k$:

$$S\left(\Omega_{D_i}\right) \cong \left(Nk_B \log\left(N\right) - N\right) - \left(n_1 k_B \log\left(n_1\right) - n_1\right) \ldots - \left(n_k k_B \log\left(n_k\right) - n_k\right) + \text{const}$$

$$S\left(\Omega_{D_i}\right) \cong -k_B \sum_i n_i \log\left(n_i\right) + \text{const} \tag{5}$$

chamando de $p_j = n_j/N$ a probabilidade de encontrar aleatoriamente o microestado j na célula ω_i:

$$S\left(\Omega_{D_i}\right) = -Nk_B \sum_j p_j \log\left(p_j\right) + \text{const} \tag{6}$$

fazendo a constante nula, tem-se, usualmente, a entropia S para a distribuição D_i. Nessa equação, nota-se que a entropia é maximizada quando todos os p_i's são iguais a 1/N. A expressão p log p é considerada, por convenção, igual a zero quando p = 0. Isso se justifica porque $\lim_{p\to 0^+}\left(p\log p\right) = 0$.

Esse desenvolvimento foi ainda mais refinado por Gibbs. Para Boltzmann, cada ponto do espaço de fase W representava um possível estado do sistema. Gibbs estende esse entendimento para um possível estado de um membro de um *ensemble*[1], cujo estado é dado por uma função densidade de probabilidade ρ(x,t). Dessa forma, a entropia passa a ser calculada por:

$$S(\rho) = -k_B \int_W \rho\left(x,t\right)\log\left(\rho\left(x,t\right)\right)dx \tag{7}$$

Na mecânica estatística quântica, o conceito de entropia foi desenvolvido pelo húngaro-estadunidense John von Neumann, mas se mantém a mesma estrutura do caso clássico. Apenas ρ é dado por uma matriz densidade e utiliza-se o operador traço da matriz como substituto ao somatório:

$$S = -k_B \text{Tr}\left(\rho\ \log\rho\right) \tag{8}$$

2.2. Entropia de Shannon e relação com o conceito físico de entropia

Trabalhando nos laboratórios da Bell Telephone, Shannon desenvolveu uma medida matemática para quantificar a perda de sinal nas linhas telefônicas. Antes dele, mm estudos técnicos sobre a transmissão de dados em telégrafos, Hartley (1928) havia proposto a quantidade $Q(p_i) = -\log p_i$ como uma medida para o cálculo de pulsos transmitidos nas bandas de comunicação (Gray, 2009). Em termos mais adequados para os interesses dessa tese, é uma medida da informação produzida pela ocorrência de um evento de probabilidade p. Baseado nesse resultado, Shannon (1948) propõe a seguinte medida para quantificar a incerteza de uma transmissão, que foi conhecida como entropia de Shannon:

[1] Conjunto de vários sistemas idênticos a um sistema estatisticamente considerado.

$$H = -\sum_{i=1}^{n} p_i \log(p_i) = E(-\log(p_i)) \quad (9)$$

Dessa forma, sua entropia é uma média aritmética ponderada da informação de Hartley que assume as seguintes propriedades:

1) É contínua em p_i, i = 1, 2, ..., n;

2) Se $p_i = 1/n$, então a entropia é uma função monótona crescente em n;

3) A entropia é maximizada numa distribuição uniforme, quando $p_i = 1/n$, uma consequência da inequação de Jensen:

$$H = E\left[\log \frac{1}{p_i}\right] \leq \log\left(E\left[\frac{1}{p_i}\right]\right) = \log n \quad (10)$$

4) A entropia de um conjunto é aditiva (igual à soma da entropia dos subconjuntos);

5) A entropia é uma função de estado, isto é, dados os p_i's entre os estados inicial e final, a entropia independe do caminho percorrido para atingir esses estados.

Ou, em sua forma contínua:

$$H = \int f_x \log f_x(x)\,dx = E(-\log f_x(x)) \quad (11)$$

Originalmente, Shannon não chamou essa quantidade de entropia, mas de "informação faltante" (*missing information*). Segundo Avery (2003, p. 81), a sugestão de chamá-la de entropia foi de von Neumann, que lhe disse: "Em primeiro lugar, um desenvolvimento matemático muito próximo já existe na mecânica estatística de Boltzmann e, em segundo lugar, ninguém entende muito bem o que é entropia, então, em qualquer discussão, você estará em posição de vantagem[2]".

No âmbito da Teoria da Informação, a entropia quantifica a incerteza associada com o valor de uma variável aleatória. Nesse sentido, a entropia envolvida em um lance de dado, por exemplo, é maior do que a de um lance de moeda. Logo, conhecer o resultado de um lance de dado reduz mais a entropia (fornece um maior ganho de informação) do que conhecer o resultado de um lance de moeda.

Num sentido restrito/prático, não há muitas semelhanças entre a entropia de informação e a entropia termodinâmica. Físicos, químicos e pesquisadores da área estão interessados em estudar como um sistema evolui espontaneamente a partir de suas condições iniciais, estabelecendo sentidos de transformações de acordo com a 2ª Lei da Termodinâmica, assumindo que todos os microestados possuem a mesma probabilidade de ocorrer. Princípios e hipóteses decorrentes da entropia (aumento da entropia do Universo, sentido preferencial de processos, diminuição da energia disponível) não estão presentes na Teoria da Informação. Além disso, a Termodinâmica clássica define a entropia em termos macroscópicos, sem fazer qualquer referência a distribuições de probabilidade. Tampouco as escalas de valores de entropia de sistemas físico-químicos, mesmo quando suprimido o efeito da constante de Boltzmann, fazem sentido como entropia de informação e vice-versa.

Por outro lado, num nível multidisciplinar e num ponto de vista matemático-estatístico, algumas conexões podem ser feitas. A primeira é a clara semelhança das expressões e das consequências matemáticas derivadas disso, como suas propriedades. Ambas também estão relacionadas com o grau de incerteza de um sistema, com a entropia sendo máxima quando a desordem é máxima e uma diminuição da entropia implicando numa maior organização do objeto de estudo. Jaynes (1957) argumenta que a entropia da mecânica estatística pode, inclusive, ser vista como uma aplicação da teoria da informação de Shannon, sendo interpretada como proporcional à quantidade de informação necessária para definir o estado microscópico detalhado do sistema. Assim,

[2] No original: "In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage".

a adição de calor a um sistema aumenta a sua entropia termodinâmica porque aumenta o número de possíveis estados microscópicos do sistema, tornando mais complexa qualquer descrição do estado.

Continuando com os desenvolvimentos da entropia de Shannon, chama-se a atenção para o fato de que a entropia H (X) não é função da variável aleatória X, mas sim da distribuição de probabilidade dessa variável. Em outras palavras, não dependente dos valores que X assume, mas das suas probabilidades.

Assim, sejam X e Y dois eventos quaisquer e p (i, j) a probabilidade conjunta de ocorrência do primeiro e do segundo evento, então a entropia conjunta H (X,Y) é dada por:

$$H(X,Y) = H(Y,X) = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{12}$$

então existe uma probabilidade condicional p (i | j), onde Y assume o valor j, tal que

$$p(i \mid j) = \frac{p(i,j)}{p(j)} \tag{13}$$

Portanto, dado X, a entropia condicional de Y é:

$$H(Y \mid X) = -\sum_i p(i) \sum_j p(j \mid i) \log p(j \mid i) = -\sum_j \sum_i p(j,i) \log p(j \mid i) \tag{14}$$

onde foi usada a regra da cadeia p (i, j) = p (i) . p (j | i) e a partir da qual, com a eq. 2.13, pode-se escrever:

$$H(Y \mid X) = -\sum_j \sum_i p(j,i) \log \left(\frac{p(j,i)}{p(i)} \right)$$

$$H(Y \mid X) = -\sum_j \sum_i p(j,i) \log p(j,i) + \sum_j \sum_i p(j,i) \log p(i)$$

$$H(Y \mid X) = -\sum_j \sum_i p(j,i) \log p(j,i) + \sum_i p(i) \log p(i)$$

$$H(Y \mid X) = H(X,Y) - H(X) \tag{15}$$

onde H (X, Y) = H (Y, X) é a entropia conjunta dos eventos X e Y.

Como H (X, Y) ≤ H (X) + H (Y), tem-se H (Y | X) ≤ H (Y), com igualdade apenas se X e Y forem independentes. O que se justifica pelo fato da entropia diminuir com o conhecimento de X, pois diminui a incerteza que existe relativamente a Y, a menos que as variáveis X e Y sejam independentes. Nesse caso, qualquer informação sobre X não diminui a entropia de Y.

Caso X possa diminuir o grau de incerteza sobre Y, pode-se considerar tal diminuição como um ganho de informação, representado pela informação mútua I(Y,X):

$$I(Y, X) = H(Y) - H(Y \mid X) \tag{16}$$

de onde, usando a equação 15, deduz-se que

$$I(Y, X) = H(Y) - [H(X, Y) - H(X)]$$
$$I(Y, X) = H(X) + H(Y) - H(X, Y) \tag{17}$$

e, ainda,

$$I(Y, Y) = H(Y)$$
$$I(Y, X) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) = I(X, Y) \tag{18}$$

dessa relação, e partindo-se da equação 17, pode-se definir a informação mútua como:

$$I(X,Y) = -\sum_i p(i)\log p(i) - \sum_j p(j)\log p(j) + \sum_i \sum_j p(i,j)\log p(i,j)$$

$$I(X,Y) = \sum_i \sum_j p(i,j)\log\left(\frac{p(i,j)}{p(i).p(j)}\right) \quad (19)$$

$$I(X,Y) = E\left[\log\left(\frac{p(i,j)}{p(i).p(j)}\right)\right] \quad (20)$$

e a informação mútua condicional é definida por

$$I(X, Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) \quad (21)$$

Pela equação 19, observa-se que a informação mútua é uma medida de independência estatística, que quanto maior for, mais relacionada são as variáveis (Cover e Thomas, 1991). A Figura 1 esclarece as relações existentes entre entropia, entropia condicional e informação mútua por meio de um diagrama de Venn.

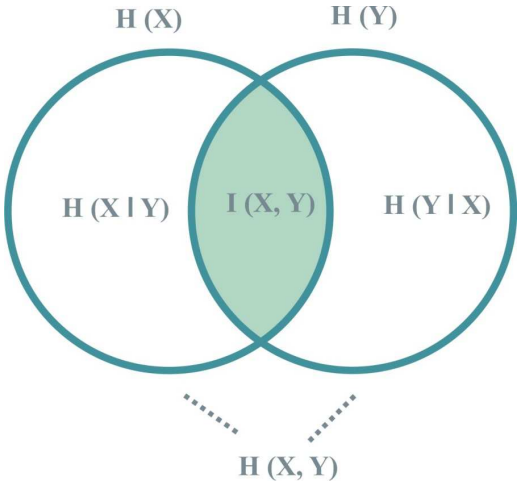

Figura 1 – Relação Entre Entropia, Entropia Condicional e Informação Mútua

Para o caso contínuo, a informação mútua passa a ser dada por

$$I(X,Y) = \int_Y \int_X f(x,y)\log\left(\frac{f_{xy}(x,y)}{f_x(x).f_y(y)}\right) dx\, dy \quad (22)$$

A informação mútua está relacionada com a divergência de Kullback-Leibler, uma medida de similaridade entre funções estritamente positivas muito utilizada para se comparar duas funções, definida por:

$$D_{KL} = \int_X f(x)\log\left(\frac{f(x)}{g(x)}\right) dx = E_f\left[\log\left(\frac{f(x)}{g(x)}\right)\right] \quad (23)$$

A divergência de Kullback-Leibler é não negativa e frequentemente referida como uma distância entre as distribuições f(x) e g(x), ainda que não seja de fato uma métrica por não ser simétrica nem satisfazer a desigualdade triangular.

Com a eq. 22, a informação mútua pode ser escrita como

$$I(X,Y) = D_{KL}\left(f_{XY}(x,y); f_X(x) f_Y(y)\right) \geq 0 \tag{24}$$

Quando a entropia de Shannon e a informação mútua precisam ser estimadas a partir de dados amostrais, pode-se recorrer à aplicação do método de histograma, utilizando o histograma de frequências relativas (Scott, 1992) com a discretização das variáveis contínuas, a entropia e a Informação Mútua de Shannon podem ser estimadas por:

$$\hat{H}(X) = -\sum_{i=1}^{N} \hat{f}_x(x_i) \log \hat{f}_x(x_i) \tag{25}$$

$$\hat{I}(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{f}_{xy}(x_i, y_j) \log \frac{\hat{f}_{xy}(x_i, y_j)}{\hat{f}_x(x_i)\hat{f}_y(y_j)} \tag{26}$$

2.3. Entropia de Rényi

A partir da equação funcional de Cauchy f(xy) = f(x) + f(y), Rényi (1961) buscou uma definição geral para medidas de informação que preservassem a aditividade de eventos independentes e fosse compatível com os axiomas da probabilidade.

A menos de uma constante de normalização, a solução é compatível com a informação de Hartley (1928) $Q(p_i) = -\log p_i$. Se fosse assumido que os eventos $x_1, x_2, ..., x_n$ pudessem ter diferentes probabilidades $p_1, p_2, ..., p_n$ e cada qual possuísse uma informação H_i, a quantidade total de informação seria dada por

$$H = -\sum_{i=1}^{n} p_i Q(p_i) \tag{27}$$

que é entropia de Shannon (eq. 9), ou seja, a média da informação de Hartley. Mas há uma suposição implícita: é utilizada a média aritmética, que não é a única possível. Afinal, para uma função g(x) com inversa g^{-1}, a média pode ser computada como

$$g^{-1}\left(\sum_{i=1}^{n} p_i g(x_i)\right) \tag{28}$$

Aplicando 28 na equação 27, tem-se:

$$H = g^{-1}\left(\sum_{i=1}^{n} p_i g(Q_i)\right) \tag{29}$$

Ao se respeitar o postulado de aditividade de eventos independentes, somente duas alternativas são possíveis para g (x):

$$g(x) = cx$$

$$g(x) = c^{-2(1-\alpha)x}$$

A primeira possibilidade fornece a entropia de Shannon, a segunda resulta em:

$$H_\alpha = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p_i^\alpha\right) \tag{30}$$

para $\alpha \geq 0$ e $\alpha \neq 1$. Esse resultado engloba uma família de medidas de informação chamadas de entropia de Rényi. É possível demonstrar[3] que a entropia de Shannon é um caso particular da entropia de Rényi quando $\alpha \to 1$.

[3] Essa demonstração é feita no Apêndice.

Ao se comparar as duas definições de entropia (eqs. 9 e 30), percebe-se que em Shannon log (p_i) é ponderado pela probabilidade, enquanto em Rényi, o logaritmo é externo à soma e α é a potência da função probabilidade. Fazendo $V_\alpha(X) = \sum_i p_i^\alpha$:

$$H_\alpha = \frac{1}{1-\alpha}\log\left(V_\alpha(X)\right) = -\log\left(\sqrt[\alpha-1]{V_\alpha(X)}\right) = -\log\left(\sqrt[\alpha-1]{E\left(V_{\alpha-1}(X)\right)}\right) \quad (31)$$

de onde se nota que $V_\alpha(X)$ é o argumento da α-norma da função probabilidade. Numa visão geométrica, as funções probabilidades compõem um espaço de n dimensões e a distância entre a origem e um ponto p (p_1, p_2, ..., p_n) é medido pela α-norma, sendo α = 2 a norma euclidiana (PRINCIPE, 2010).

Utilizando-se α = 2, tem-se a entropia quadrática de Rényi:

$$H_\alpha = \log\left(\sum_{i=1}^{n} p_i^2\right) \quad (32)$$

Essa escolha faz com que o argumento do logaritmo tenha um sentido próprio, afinal ele é E[p(X)]. Ou, de maneira alternativa, ao se fazer a mudança de variável $\varepsilon_i = p(x_i)$, o argumento é a média da variável transformada, enquanto a função probabilidade é a transformação. A entropia quadrática de Rényi é particularmente interessante por ser facilmente estimada a partir de dados amostrais.

É possível mostrar que a medida de entropia de Rényi pode ser estendida para variáveis aleatórias contínuas (Gonçalves, Macrini; 2011), porém sem a condição da entropia ser não negativa, e se torna:

$$H_\alpha = \frac{1}{1-\alpha}\log\int f^\alpha(x)\,dx \quad (33)$$

Para estimar a entropia quadrática de Rényi é possível utilizar a janela de Parzen. Para isso, atribui-se um *kernel* sobre os dados da amostra e os soma com uma normalização adequada. Uma possibilidade é:

$$\hat{f}_X(x) = \frac{1}{N\sigma}\sum_{i=1}^{N}\kappa\left(\frac{x-x_i}{\sigma}\right) \quad (34)$$

Parzen (1962) provou que esse estimador é assintoticamente não-viesado e consistente. O parâmetro σ é chamado de tamanho do *kernel*. Normalmente se utiliza uma gaussiana e σ se torna o desvio padrão. Desse modo, para a entropia quadrática utilizando um *kernel* gaussiano:

$$\hat{H}_2(X) = -\log\int_{-\infty}^{\infty}\left(\frac{1}{N}\sum_{i=1}^{N}G_\sigma(x-x_i)\right)^2 dx$$

$$\hat{H}_2(X) = -\log\frac{1}{N^2}\int_{-\infty}^{\infty}\sum_{i=1}^{N}\sum_{j=1}^{N}G_\sigma(x-x_j)\cdot G_\sigma(x-x_i)dx$$

$$\hat{H}_2(X) = -\log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}G_\sigma(x-x_j)\cdot G_\sigma(x-x_i)\,dx$$

$$\hat{H}_2(X) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}G_{\sigma\sqrt{2}}(x_j-x_i)\right) \quad (35)$$

O argumento do logaritmo é chamado de $\hat{V}_2(X)$ – Potencial de Informação (PRINCIPE, 2010). Escrevê-lo dessa maneira irá facilitar desenvolvimentos posteriores.

Com um *kernel* gaussiano, não é preciso calcular a integral explicitamente, uma vez que a integral de um produto de gaussianas é uma gaussiana com parâmetros modificados (um tamanho

maior de *kernel*). O parâmetro σ deve ser selecionado pelo usuário, normalmente com base é no método de *cross validation* ou pela regra de Silverman, (Jenssen *et al*, 2006):

$$\sigma_{opt} = \sigma_X \left(4N^{-1} (2d+1)^{-1} \right)^{\frac{1}{d+4}} \tag{36}$$

A princípio a informação mútua de Rényi não pode ser expressa em termos da entropia, como foi feito pela eq. 2.92 para a entropia de Shannon. No entanto, se for usada a divergência de Cauchy-Schwarz para se definir a informação mútua, é possível estabelecer uma relação (Gonçalves, Macrini; 2011).

A informação mútua quadrática pode ser obtida a partir da divergência de Cauchy-Schwarz. Partindo-se da desigualdade de Cauchy-Schwarz, pode-se definir a divergência:

$$\begin{aligned}
&\left|\langle u, v\rangle\right| \leq \|u\| \, \|v\| \\
&\frac{\left|\langle u, v\rangle\right|}{\sqrt{\|u\|^2 \, \|v\|^2}} \leq \frac{\|u\| \, \|v\|}{\sqrt{\|u\|^2 \, \|v\|^2}} \\
&\frac{\left|\langle u, v\rangle\right|}{\sqrt{\|u\|^2 \, \|v\|^2}} \leq 1 \\
&\log \frac{\left|\langle u, v\rangle\right|}{\sqrt{\|u\|^2 \, \|v\|^2}} \leq \log 1 \\
&\log \frac{\left|\langle u, v\rangle\right|}{\sqrt{\|u\|^2 \, \|v\|^2}} \leq 0 \\
&-\log \frac{\left|\langle u, v\rangle\right|}{\sqrt{\|u\|^2 \, \|v\|^2}} \geq 0
\end{aligned}$$

que com u e v representando funções pode ser escrita como:

$$D_{CS}(f,g) = -\log \frac{\int f(x) g(x) dx}{\sqrt{\int f^2(x) g^2(x) dx}} \tag{37}$$

A divergência D_{CS} é sempre simétrica e não negativa, sendo nula somente se f(x) = g(x). Na prática, trabalha-se com o quadrado da razão acima, fazendo com que a divergência de Cauchy-Schwarz possa ser escrita como:

$$D_{CS}(f,g) = \log\left(\int f(x)^2 dx \right) + \log\left(\int g(x)^2 dx \right) - 2\log\left(\int f(x) g(x) dx \right) \tag{38}$$

definindo-se a informação mútua quadrática por

$$I_{CS}(X,Y) = D_{CS}\left(f_{XY}(x,y); f_X(x) f_Y(y) \right) \tag{39}$$

e, em termos de entropia quadrática, seguindo a equação 38

$$I_{CS}(X,Y) = H_2\left(f_{XY} \times f_X f_Y \right) - \frac{1}{2} H_2\left(f_{XY} \right) - \frac{1}{2} H_2\left(f_X f_Y \right) \tag{40}$$

Para estimar a eq. 39, pode-se nomear os três seguintes potenciais de informação (Principe, 2010):

$$\hat{V}_{x_1} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_1(i) - x_1(j)\right)^2 \tag{41}$$

$$\hat{V}_{x_2} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_2(i) - x_2(j)\right)^2 \tag{42}$$

$$\hat{V}_{c} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_1(i) - x_2(j)\right)^2 \tag{43}$$

de forma que

$$\hat{D}_{CS}\left(x_1, x_2\right) = \log\frac{\hat{V}_{x_1}\hat{V}_{x_2}}{\hat{V}_{c}} \tag{44}$$

Para o caso da informação mútua quadrática, a desigualdade de Cauchy-Schwarz envolve a função conjunta, as funções marginais e o produto delas. As três funções densidade de probabilidade estimadas são:

$$\hat{f}_{x_1}\left(x_1\right) = \frac{1}{N}\sum_{i=1}^{N} G_{\sigma}\left(x_1 - x_1(i)\right) \tag{45}$$

$$\hat{f}_{x_2}\left(x_2\right) = \frac{1}{N}\sum_{i=1}^{N} G_{\sigma}\left(x_2 - x_2(i)\right) \tag{46}$$

$$\hat{f}_{x_1x_2}\left(x_1, x_2\right) = \frac{1}{N}\sum_{i=1}^{N} G_{\sigma}\left(x - x(i)\right) \tag{47}$$

que geram os seguintes potenciais:

$$\hat{V}_j = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x(i) - x(j)\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_1(i) - x_1(j)\right) G_{\sigma\sqrt{2}}\left(x_2(i) - x_2(j)\right)$$

$$\hat{V}_m = \hat{V}_1\hat{V}_2 \quad \text{com } \hat{V}_k = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_k(i) - x_k(j)\right), \text{ para } k = 1,2$$

$$\hat{V}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_1(i) - x_1(j)\right)\right)\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}\left(x_2(i) - x_2(j)\right)\right) \tag{48}$$

Exemplificando para o caso de $\hat{V}_c$, usando as equações 45, 46 e 47:

$$\hat{V}_c = \int\int \hat{f}(x_1, x_2)\hat{f}(x_1)\hat{f}(x_2)dx_1dx_2$$

$$\hat{V}_c = \int\int \left[\frac{1}{N}\sum_{k=1}^{N} G_\sigma(x_1 - x_1(k))G_\sigma(x_2 - x_2(k))\right]\left[\frac{1}{N}\sum_{i=1}^{N} G_\sigma(x_1 - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_\sigma(x_2 - x_2(j))\right]dx_1dx_2$$

$$\hat{V}_c = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{N}\sum_{k=1}^{N}\int G_\sigma(x_1 - x_1(i))G_\sigma(x_1 - x_1(k))dx_1\int G_\sigma(x_2 - x_2(j))G_\sigma(x_2 - x_2(k))dx_2$$

$$\hat{V}_c = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sqrt{2}\sigma}(x_1(k) - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(k) - x_2(j))\right]$$

e aplicando a equação 44, finalmente tem-se:

$$\hat{I}_{CS}(X,Y) = \log\frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))\cdot G_{\sqrt{2}\sigma}(x_2(i) - x_2(j))\right)\left(\hat{V}_{x_1}\hat{V}_{x_2}\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\left(\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))\right)\left(\frac{1}{N}\left(\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(i) - x_2(j))\right)\right)\right)^2} \quad (49)$$

3. ÁRVORES DE DECISÃO

Árvores de decisão são modelos simbólicos de mineração de dados cuja estrutura apresenta-se no formato de uma árvore. Cada nó interno da árvore indica um teste sobre um atributo, cada ramo representa um resultado do teste, e os nós terminais (folhas) correspondem a classes ou distribuições de classes. A profundidade de uma árvore é definida pela maior distância entre uma folha e a raiz (primeiro nó). Com isso, tem-se uma técnica que constrói regras de classificação passíveis de avaliação, interpretação e posterior aplicação. Por esses motivos, as árvores de decisão tornam-se interessantes para os problemas de reconhecimento de padrões.

Algumas das vantagens apresentadas pelas árvores de decisão são sua flexibilidade, pois não assumem uma distribuição única dos dados, sendo métodos não-paramétricos; robustez, uma vez a seleção interna de características produz árvores que tendem a ser bastante robustas mesmo com a adição de variáveis irrelevantes; interpretabilidade, já que todas as decisões são baseadas nos valores (conhecidos) dos atributos usados para descrever o problema; e velocidade, pois a maioria dos algoritmos constrói rapidamente as árvores de decisão (Gama, 1999).

Em geral, o procedimento de uma árvore de decisão consiste em apresentar um conjunto de dados ao nó raiz da árvore e avaliá-lo segundo um teste lógico. Dependendo do resultado, a árvore ramifica-se para um dos nós descendentes e este procedimento é repetido até que uma folha conceda a classificação dos dados. A Figura 2 exemplifica a estrutura de uma árvore de decisão de resposta binária.

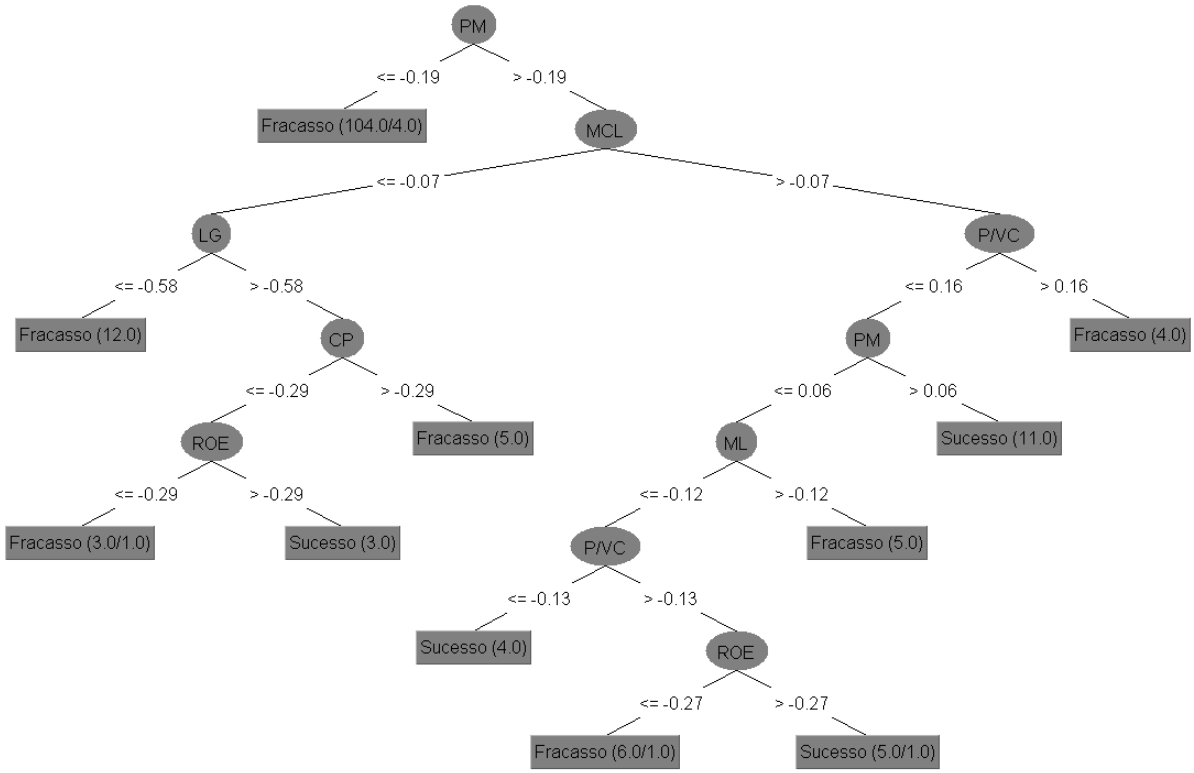

Figura 2 – Exemplo de Árvore De Decisão

O aprendizado de uma árvore de decisão é normalmente supervisionado, ou seja, o método aproxima funções-alvo de valor discreto, na qual a função aprendida é representada por uma árvore de decisão. As árvores treinadas podem ser representadas como um conjunto de regras *if-then* para melhor compreensão e interpretação.

As árvores de decisão são construídas usando um algoritmo de partição recursiva. Uma das possibilidades é este algoritmo construir uma árvore por divisões recursivas binárias que começa no nó raiz e desce até os nós folhas. Nesse caso, têm-se dois fatores principais no algoritmo de partição: a forma de selecionar uma divisão para cada nó intermediário (crescimento) e uma regra para determinar quando um nó é terminal (poda).

A maioria dos algoritmos de construção (também chamados de indução) de árvores de decisão corresponde a um procedimento guloso[4] que recursivamente constrói a árvore do nó raiz em direção aos nós terminais. Em cada iteração, a partir da base de dados de treinamento, os algoritmos procuram pelo atributo que melhor separa as classes para realizarem a ramificação da árvore, e recursivamente processam os subproblemas resultantes das ramificações.

Essa abordagem de divisão e conquista adotada pelos algoritmos de indução de árvores de decisão foi desenvolvida e refinada ao longo de vários anos por John Ross Quinlan. Sua contribuição inicial foi o algoritmo ID3 (Quinlan, 1986). Várias melhorias foram realizadas nesse algoritmo, culminando no surgimento do algoritmo C4.5 (Quinlan, 1993), muito utilizado em aplicações práticas e pesquisas acadêmicas.

Em ambos, a escolha do atributo que geram as ramificações é feita a partir de uma medida conhecida como Ganho de Informação, que nada mais é do que a informação mútua entre a variável escolhida e a variável resposta condicionada às escolhas feitas na iteração i, como será mostrado a seguir. O atributo que proporcionar o maior ganho de informação é selecionado como atributo teste do nó corrente, em outras palavras, é escolhida a variável que promove a maior minimização da entropia.

[4] Técnica para problemas de otimização que sempre faz uma escolha ótima local com a intenção de atingir uma solução ótima global.

Originalmente, o Ganho de Informação é baseado na entropia de Shannon (eq. 9). Segundo Merschmann (2007), procedendo-se dessa forma para a seleção de atributos constrói-se árvores mais simples e minimiza-se o número de testes necessários para a classificação.

Para isso, compara-se a entropia da variável do nó imediatamente acima (nó-pai) com a entropia condicional da variável resposta. O atributo que gerar uma maior diferença (maior redução da entropia/ganho de informação) é escolhido como condição de teste.

$$\text{Ganho de informação} = \text{entropia}\left(\text{nó-pai}\right) - \sum_{j=1}^{n} \frac{N\left(v_j\right)}{N} \text{entropia}\left(v_j\right) \qquad (50)$$

onde n é o número de nós-filhos, N é o número total de objetos do nó-pai e $N(v_j)$ é o número de observações associadas ao nó-filho. Essa medida corresponde à informação mútua entre o atributo analisado e a variável resposta (desfecho) daquela iteração, ou seja:

$$\text{Ganho de informação} = I\left(A, D \mid i\right) = H\left(A \mid i\right) - H\left(A \mid D, i\right)$$

onde i representa o conjunto de escolhas, testes lógicos e classificações feitas na iteração i, A é a variável a ser analisada no nó e D é a variável desfecho.

A afirmação anterior pode ser clarificada pelo seguinte exemplo. Suponha haver dois atributos candidatos a nó em uma determinada iteração da i árvore de decisão com a distribuição de classe Sucesso e Fracasso nas folhas desses atributos, conforme mostra a Figura 3. Nela também estão representadas as observações presentes em cada grupo Sucesso/Fracasso resultantes do teste lógico.

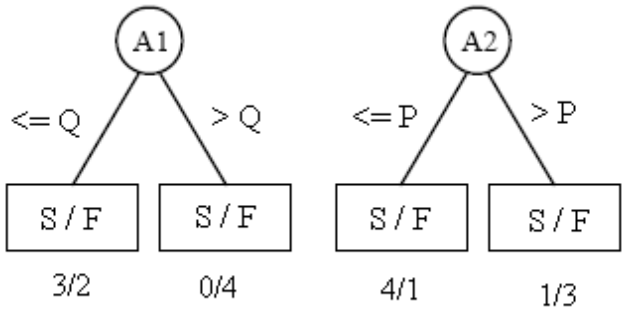

Figura 3 – Exemplo de escolha de atributos em duas possíveis partições a partir da informação mútua

Dada a separação proposta para as variáveis e a classificação das respostas para o nó i, tem-se os valores necessários já condicionados à i.. O ganho de informação, usando o logaritmo de base 10 e as equações 9, 14 e 18, para os atributos A_1 e A_2 é calculado, para a entropia de Shannon, da seguinte maneira:

$$H(A_1) = -\sum_{k=1}^{n} p_k \log\left(p_k\right)$$

$$H(A_1) = -\frac{3}{9}\log\left(\frac{3}{9}\right) - \frac{6}{9}\log\left(\frac{6}{9}\right) = 0,2764$$

$$H(A_1 \mid D) = -\sum_{k} p_D\left(k\right) \sum_{j} p_{A_1|D}\left(j \mid k\right) \log p_{A_1|D,i}\left(j \mid k\right)$$

$$H(A_1 \mid D) = \frac{5}{9}\left(-\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)\right) + \frac{4}{9}\left(-\frac{0}{4}\log\left(\frac{0}{4}\right) - \frac{4}{4}\log\left(\frac{4}{4}\right)\right) = 0,1624$$

$$I\left(A_1 \mid D\right) = H\left(A_1\right) - H\left(A_1 \mid D\right)$$

$$I\left(A_1 \mid D\right) = 0,2764 - 0,1624 = 0,1141$$

$$H(A_2) = -\frac{5}{9}\log\left(\frac{5}{9}\right) - \frac{4}{9}\log\left(\frac{4}{9}\right) = 0,2983$$

$$H(A_2 \mid D) = \frac{5}{9}\left(-\frac{4}{5}\log\left(\frac{4}{5}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right)\right) + \frac{4}{9}\left(-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right) = 0,2293$$

$$I(A_2 \mid D) = H(A_2) - H(A_2 \mid D)$$

$$I(A_2 \mid D) = 0,2983 - 0,2293 = 0,0691$$

Como ganho de informação do atributo A_1 é maior, ele seria o atributo considerado para entrar no modelo.

Vale lembrar que a probabilidade utilizada no cálculo de entropia é a fração de observações pertencentes a determinada classe no nó. Sendo assim, quanto menor a entropia, mais desbalanceada é a distribuição de classes. Em um determinado nó, a entropia é nula se todos os exemplos nele pertencerem à mesma classe. Analogamente, a entropia é máxima no nó se houver o mesmo número de casos para cada classe possível.

Desse modo, o critério de ganho de informação seleciona como atributo-teste aquele que minimiza a entropia, por ter a maior informação mútua com a variável resposta naquela iteração. O grande problema ao se utilizar o ganho de informação é que ele dá preferência a atributos com muitos valores possíveis. Um exemplo claro de problema ocorreria ao utilizar um atributo totalmente dispensável (por exemplo, um identificador único). Nesse caso, seria criado um nó para cada valor possível, e o número de nós seria igual ao número de identificadores. Cada um desses nós teria apenas um exemplo, o qual pertence a uma única classe, ou seja, os exemplos seriam totalmente discriminados. Assim, o valor da entropia seria mínima porque, em cada nó, todos os exemplos (no caso um só) pertencem à mesma classe. Essa divisão geraria um ganho máximo, embora a regra de classificação construída seja irrelevante.

Para contornar esse problema do ganho de informação, foi proposto em Quinlan (1993) o uso da Razão de Ganho (*Gain Ratio*), que consiste no ganho de informação relativo como critério de avaliação:

$$\text{Razão de Ganho} = \frac{\text{Ganho de informação}}{\text{entropia(nó)}} = \frac{I(A \mid D, i)}{H(A \mid i)} \qquad (51)$$

Pela equação 51, é possível perceber que a razão não é definida quando o denominador é igual a zero. Além disso, a razão de ganho favorece atributos cujo denominador – a entropia – possui valor pequeno. Em Quinlan (1988), é sugerido que a avaliação pela Razão de Ganho seja realizada em duas etapas. Na primeira, é calculado o ganho de informação para todos os atributos. Após isso, consideram-se apenas aqueles que obtiveram um ganho de informação acima da média, e então se escolhe aquele que apresenta a melhor razão de ganho. Com esse procedimento, Quinlan mostrou que a Razão de Ganho supera o Ganho de Informação tanto em termos de acurácia, quanto em termos de complexidade das árvores de decisão geradas, sendo esse o método implementado no algoritmo C4.5 (Quinlan, 1993).

O exemplo anterior refere-se à entropia de Shannon. Ao se optar, por exemplo, pela entropia quadrática de Rényi, devem-se, naturalmente, modificadas as equações de entropia e de informação mútua. Da mesma forma, caso sejam usados dados amostrais, são os estimadores da entropia e da informação mútua a serem utilizados.

A indução de Árvores de Decisão não se resume ao cálculo do ganho da informação e da razão de ganho, existindo outras questões envolvidas, como o valor atribuído para o teste do nó ou técnicas de podagem que buscam evitar o superajustamento aos dados. Contudo, são questões que fogem ao escopo desse trabalho e, portanto, não serão abordadas.

O uso dos conceitos da Teoria da Informação no âmbito de reconhecimento de padrões foram mostrados para Árvores de Decisão, mas existem diversas outras possibilidades. Duas, em especial,

são seu uso na composição de redes neurais artificiais ou em algoritmos de seleção de variáveis MIFS (*Mutual Information Feature Selector*). Uma implementação dessa última sugestão foi feita, para o critério de distribuições uniformes (MIFS-U), por Gonçalves e Macrini (2011). Outra sugestão é se desenvolver diferentes estimadores dos aqui empregados para o cálculo da entropia e da informação mútua.

4. CONCLUSÃO

Neste artigo foram discutidos conceitos da Teoria da Informação para sua aplicação em pesquisas no âmbito de técnicas de reconhecimento de padrões. Mais especificamente, refere-se ao uso da entropia e da informação mútua como medidas capazes de fornecer informações sobre variáveis que melhor discriminam os dados.

Para isso, foi debatida a origem do conceito de entropia e mostrada a validade limitada, porém existente, da comparação com a Física. Também foi apresentada a generalização de Rényi para a entropia de Shannon e como estimá-la por meio de uma Janela de Parzen.

Por fim, exemplificou-se a aplicação dos conceitos de entropia e informação mútua na técnica de Árvores de Decisão, evidenciando que o conceito ganho de informação corresponde à informação mútua.

Assim, espera-se contribuir, ainda que minimamente, para preencher a lacuna existente na literatura nacional sobre o tema e, desse modo, evitar que equívocos no uso dos conceitos da Teoria da Informação se avolumem. Por outro lado, ressalta-se que a maneira apresentada é apenas uma forma de se aplicar a Teoria da Informação em problemas de reconhecimento de padrões, não querendo se limitar as possibilidades de métodos a somente ela.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AVERY, J. *Information Theory and Evolution*. Cingapura: World Scientific, 2003.

COVER, T. M.; THOMAS, J. A. *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

GAMA, J. M. P. *Combining classification algorithms*. Porto, 1999. 195p. Tese (Doutorado em Ciência de Computadores) – Departamento de Ciência de Computadores, Universidade do Porto.

GONCALVES, L. B.; MACRINI, J. L. R. Rényi entropy and cauchy-schwartz mutual information applied to mifs-u variable selection algorithm: a comparative study. *Pesquisa Operacional*, Rio de Janeiro, v. 31, n. 3, p. 499-519, dez. 2011.

GRAY, R. M. *Entropy and Information Theory*. New York: Springer Verlag, 2009.

HARTLEY, R. V. L. The Transmission of Information. *Bell System Technical Journal*, v. 7, n. 3, p. 535-563, jul. 1928.

JAYNES, E. T. Information Theory and Statistical Mechanics. *Physical Review*, v. 106, n.4: p. 620–630, 1957.

JENSSEN, R.; PRINCIPE, J. C.; ERDOGMUS, D.; ELTOFT, T. The Cauchy-Schwarz Divergence and Parzen Windowing: Connections to Graph Theory and Mercer Kernels. *Journal of the Franklin Institute*, v. 343, n 6, p. 614–629, Set. 2006.

MERSCHMANN, L. H. C. *Classificação probabilística baseada em análise de padrões*. Niterói, 2007. 117 f. Tese (Doutorado em Otimização Combinatória) – Programa de Pós-Graduação em Computação, Universidade Federal Fluminense.

PARZEN, E., On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, v. 33, n. 3, p. 1065-1076, 1962.

PRINCIPE, J. C. *Information theoretic learning*: Rényi's entropy and kernel perspectives. New York: Springer Verlag, 2010.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81-106, 1986.

_____. Decision trees and multivalued attributes. *Machine Intelligence*, n. 11, p. 305-318, 1988.

_____. *C4.5*: Programs for Machine Learning. San Diego (EUA): Morgan Kaufmann, 1993.

RÉNYI, A. On measures of entropy and information. In: FOURTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 1960. Berkeley: University of California Press, v.1, 1961. *Anais...* p. 547-561.

SCOTT, D. W. *Multivariate Density Estimation*. New York: John Wiley & Sons, 1992.

SHANNON, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal*, n. 27, p. 379-423 e p. 623-656, jul./out. 1948.

ANEXO A – DEDUÇÃO DA ENTROPIA DE SHANNON A PARTIR DA ENTROPIA de rÉnyI

A entropia de Rényi é dada por[5]

$$H_\alpha = \frac{1}{1-\alpha}\ln\left(\sum_{i=1}^{n} p_i^\alpha\right) \tag{52}$$

Para α = 1 seu resultado é a indefinição 0/0. Assim, o limite da entropia de Rényi quando $\alpha \to 1$ pode ser encontrado pela regra de l'Hôpital:

$$\lim_{\alpha\to k}\frac{f(\alpha)}{g(\alpha)} = \lim_{\alpha\to k}\frac{f'(\alpha)}{g'(\alpha)} \tag{53}$$

fazendo $f(\alpha) = \ln\sum_{i=1}^{n} p_i^\alpha$ e $g(\alpha) = 1-\alpha$:

$$g' = -1$$

$$f' = \frac{1}{\sum_{i=1}^{n} p_i^\alpha}\sum_{i=1}^{n}\frac{d}{d\alpha}\left(p_i^\alpha\right)$$

[5] A base logarítmica é indiferente, foi adotado o logaritmo natural apenas para simplicidade da demonstração.

a derivada $\frac{d}{d\alpha} p_i^{\alpha}$ é:

$$\frac{d}{d\alpha} p_i{}^{\alpha} = \frac{d}{d\alpha} e^{\alpha \ln p_i} = e^{\alpha \ln p_i} \frac{d}{d\alpha} \alpha \ln p_i = a^{\alpha} \ln p_i$$

$$f' = \frac{1}{\sum_{i=1}^{n} p_i^{\alpha}} \sum_{i=1}^{n} p_i^{\alpha} \ln p_i$$

$$\lim_{\alpha \to 1} H_{\alpha} = \lim_{\alpha \to 1} \frac{f'(\alpha)}{g'(\alpha)} = -\frac{1}{\sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i \ln p_i$$

como $\sum_{i=1}^{n} p_i = 1$:

$$\lim_{\alpha \to 1} H_{\alpha} = -\sum_{i=1}^{n} p_i \ln p_i \qquad (54)$$

que é a entropia de Shannon.