

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO



Leonardo Barroso Gonçalves

**Entropia de Rényi e
Informação Mútua de Cauchy-Schwartz
Aplicadas ao Algoritmo de
Seleção de Variáveis MIFS-U:
Um Estudo Comparativo**

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE ENGENHARIA ELÉTRICA
Programa de Pós-Graduação em
Engenharia Elétrica

Rio de Janeiro, abril de 2008



Leonardo Barroso Gonçalves

**Entropia de Rényi e
Informação Mútua de Cauchy-Schwartz
Aplicadas ao Algoritmo de
Seleção de Variáveis MIFS-U:
Um Estudo Comparativo**

DISSERTAÇÃO DE MESTRADO

Dissertação apresentada ao Programa de Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Carlos S. Kubrusly

Co-orientador: José Leonardo R. Macrini

Rio de Janeiro

Abril de 2008



Leonardo Barroso Gonçalves

**Entropia de Rényi e
Informação Mútua de Cauchy-Schwartz
Aplicadas ao Algoritmo de
Seleção de Variáveis MIFS-U:
Um Estudo Comparativo**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Dr. Carlos S. Kubrusly

Orientador

Departamento de Engenharia Elétrica – PUC-Rio

Dr. José Leonardo R. Macrini

Co-orientador

Programa de Pós-Graduação em Metrologia – PUC-Rio

Dr. Getúlio Borges da Silveira Filho

Departamento de Economia - UFRJ

Dra. Elisabeth Costa Monteiro

Programa de Pós-Graduação em Metrologia – PUC-Rio

Prof. José Eugenio Leal

Coordenador Setorial do Centro

Técnico Científico – Puc-Rio

Rio de Janeiro, 03 de abril de 2008

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Ficha Catalográfica

Gonçalves, Leonardo Barroso

Entropia de Rényi e informação mútua de Cauchy-Schwartz aplicadas ao algoritmo de seleção de variáveis MIFS-U: um estudo comparativo / Leonardo Barroso Gonçalves ; orientador: Carlos S. Kubrusly ; co-orientador: José Leonardo R. Macrini. – 2008.

106 f. : il. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui referências bibliográficas.

1. Engenharia elétrica – Teses. 2. Seleção de variáveis. 3. Entropia. 4. Informação mútua. 5. Janela de Parzen. I. Kubrusly, Carlos S. II. Macrini, José Leonardo R. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

A meus pais (*in memoriam*), Leonardo e Isabel,
meus eternos heróis, pelas lições de vida,
pelo incentivo aos estudos e, principalmente,
por terem-me transmitido os reais valores da vida:
caráter, honestidade, humildade e respeito ao próximo.

Agradecimentos

A Deus, pois devo absolutamente tudo a Ele.

Ao meu orientador Professor Carlos S. Kubrusly, por sua amizade, pelos seus conselhos e lições de vida.

Ao meu co-orientador Professor José Leonardo R. Macrini e à sua esposa Márcia, meus amigos, pelo estímulo e parceria na realização deste trabalho.

Aos demais membros da comissão examinadora, Professores Getúlio Borges da Silveira Filho e Elisabeth Costa Monteiro, pelas valiosas contribuições ao texto da dissertação.

Aos professores do Departamento de Engenharia Elétrica da PUC-Rio, pelos seus ensinamentos.

Aos funcionários do Departamento de Engenharia Elétrica da PUC-Rio, pelo seu trabalho e muita paciência. Em especial, Alcina e Márcia, por suas orientações.

À PUC-Rio, pelo acolhimento e auxílio concedido.

À minha amiga Neyde Martins Zambelli, por me mostrar o caminho a seguir.

À minha amiga Giselle Diniz Gonçalves, pelo constante incentivo e por sua paciência, compartilhando os bons e maus momentos.

Ao meu amigo Thiago Baptista Rodrigues, pelas valiosas dicas.

Às minhas irmãs, Albina e Fátima, pela compreensão que sempre tiveram. E aos meus queridos sobrinhos, Bruna, Gabriel, Giuliana e Guilherme, pela constante alegria.

Resumo

Gonçalves, Leonardo Barroso; Kubrusly, Carlos S. (Orientador); Macrini, José Leonardo R. (Co-orientador). **Entropia de Rényi e Informação Mútua de Cauchy-Schwartz Aplicadas ao Algoritmo de Seleção de Variáveis MIFS-U: Um Estudo Comparativo**. Rio de Janeiro, 2008, 106 p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A presente dissertação aborda o algoritmo de Seleção de Variáveis Baseada em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U) e expõe um método alternativo para estimação da entropia e da informação mútua, “medidas” que constituem a base deste algoritmo de seleção. Este método tem, por fundamento, a informação mútua quadrática de Cauchy-Schwartz e a entropia quadrática de Rényi, combinada, no caso de variáveis contínuas, ao método de estimação de densidade Janela de Parzen. Foram realizados experimentos com dados reais de domínio público, sendo tal método comparado com outro, largamente utilizado, que adota a definição de entropia de Shannon e faz uso, no caso de variáveis contínuas, do estimador de densidade histograma. Os resultados mostram pequenas variações entre os dois métodos, mas que sugerem uma investigação futura através de um classificador, tal como Redes Neurais, para avaliar qualitativamente tais resultados à luz do objetivo final que consiste na maior exatidão de classificação.

Palavras-chave

Seleção de Variáveis, Entropia, Informação Mútua, Janela de Parzen.

Abstract

Gonçalves, Leonardo Barroso; Kubrusly, Carlos S. (Advisor); Macrini, José Leonardo R. (Co-advisor). **Rényi Entropy and Cauchy-Schwartz Mutual Information Applied to the MIFS-U Variables Selection Algorithm: A Comparative Study**. Rio de Janeiro, 2008, 106 p. MSc. Dissertation – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation approaches the algorithm of Selection of Variables under Mutual Information with Uniform Distribution (MIFS-U) and presents an alternative method for estimate entropy and mutual information, “measures” that constitute the base of this selection algorithm. This method has, for foundation, the Cauchy-Schwartz quadratic mutual information and the quadratic Rényi entropy, combined, in the case of continuous variables, with Parzen Window density estimation. Experiments were accomplished with real public domain data, being such method compared with other, broadly used, that adopts the Shannon entropy definition and makes use, in the case of continuous variables, of the histogram density estimator. The results show small variations among the two methods, what suggests a future investigation through a classifier, such as Neural Networks, to evaluate this results, qualitatively, in the light of the final objective that consists of the biggest sort exactness.

Keywords

Variables Selection, Entropy, Mutual Information, Parzen Window.

Sumário

1 Introdução	14
1.1 Considerações Iniciais e Objetivo	14
1.2 Estrutura da Dissertação	15
2 Teoria da Informação	16
2.1 Entropia e Informação Mútua de Shannon	16
2.2 Principais Relações	21
2.3 Entropia Diferencial de Shannon	22
2.4 Entropia, Entropia Diferencial e Informação Mútua de Rényi	25
2.5 Informação Mútua de Cauchy-Schwartz	29
3 Seleção de Variáveis Baseada em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U)	32
3.1 Seleção de Variáveis de Entrada para Problemas de Classificação	32
3.2 Problema $FR_n - k$ e o Algoritmo de Seleção Ideal	33
3.3 Seleção de Variáveis Baseada em Informação Mútua (MIFS)	34
3.4 O Seletor de Variáveis MIFS-U	36
4 Métodos de Estimação da Entropia e da Informação Mútua	40
4.1 Método Shannon / Histograma	41
4.2 Método Cauchy-Schwartz / Parzen-Rosenblatt	42
4.2.1 Estimador de Densidade Janela de Parzen	43
4.2.2 Escolha da Largura da Janela	45
4.2.3 Cálculos Necessários ao MIFS-U	48
5 Experimentos com Dados Reais	56
5.1 Descrição das Bases de Dados Utilizadas	56
5.1.1 Base de Dados Ecocardiograma	57
5.1.2 Base de Dados Telescópio	58
5.1.3 Base de Dados Vinho	59

5.1.4 Base de Dados Dermatologia	60
5.1.5 Base de Dados Câncer de Mama	61
5.1.6 Base de Dados Doenças Cardíacas	62
5.2 Comparação dos Métodos	62
5.2.1 Resultado Comparativo da Seleção pelo MIFS-U Ecocardiograma	63
5.2.2 Resultado Comparativo da Seleção pelo MIFS-U Telescópio	64
5.2.3 Resultado Comparativo da Seleção pelo MIFS-U Vinho	65
5.2.4 Resultado Comparativo da Seleção pelo MIFS-U Dermatologia	66
5.2.5 Resultado Comparativo da Seleção pelo MIFS-U Câncer de Mama	68
5.2.6 Resultado Comparativo da Seleção pelo MIFS-U Doenças Cardíacas	69
6 Considerações Finais	70
Referências Bibliográficas	72
Apêndice (A) – Informação Mútua entre as Variáveis	78
A.1 Método Shannon / Histograma	78
A.2 Método Cauchy-Schwartz / Parzen-Rosenblatt	91
Anexo (B) – Integração do Produto de Núcleos Gaussianos	104

Lista de figuras

Figura 2.1 – Entropia ($H(p)$) versus p	18
Figura 2.2 – Relação entre entropia e informação mútua	22
Figura 3.1 – Relação entre variáveis de entrada e desfecho	35
Figura 3.2 – Esquema do algoritmo MIFS-U	39
Figura 4.1 – Estimador Núcleo: efeito da largura da janela	44

Lista de tabelas

Tabela 5.1 – Base de Dados ECOCARDIOGRAMA	57
Tabela 5.2 – Base de Dados TELESCÓPIO	58
Tabela 5.3 – Base de Dados VINHO	59
Tabela 5.4 – Base de Dados DERMATOLOGIA	60
Tabela 5.5 – Base de Dados CÂNCER DE MAMA	61
Tabela 5.6 – Base de Dados DOENÇAS CARDÍACAS	62
Tabela 5.7 – Resultado Comparativo da Seleção pelo MIFS-U ECOCARDIOGRAMA	63
Tabela 5.8 – Resultado Comparativo da Seleção pelo MIFS-U TELESCÓPIO	64
Tabela 5.9 – Resultado Comparativo da Seleção pelo MIFS-U VINHO	65
Tabela 5.10 – Resultado Comparativo da Seleção pelo MIFS-U DERMATOLOGIA	66
Tabela 5.11 – Resultado Comparativo da Seleção pelo MIFS-U CÂNCER DE MAMA	68
Tabela 5.12 – Resultado Comparativo da Seleção pelo MIFS-U DOENÇAS CARDÍACAS	69
Tabela A.1 – Informação Mútua entre variáveis ECOCARDIOGRAMA Método Shannon / Histograma	78
Tabela A.2 – Informação Mútua entre variáveis TELESCÓPIO Método Shannon / Histograma	79
Tabela A.3 – Informação Mútua entre variáveis VINHO Método Shannon / Histograma	80
Tabela A.4 – Informação Mútua entre variáveis DERMATOLOGIA Método Shannon / Histograma	81

Tabela A.5 – Informação Mútua entre variáveis CÂNCER DE MAMA Método Shannon / Histograma	85
Tabela A.6 – Informação Mútua entre variáveis DOENÇAS CARDÍACAS Método Shannon / Histograma	90
Tabela A.7 – Informação Mútua entre variáveis ECOCARDIOGRAMA Método Cauchy-Schwartz / Parzen	91
Tabela A.8 – Informação Mútua entre variáveis TELESCÓPIO Método Cauchy-Schwartz	92
Tabela A.9 – Informação Mútua entre variáveis VINHO Método Cauchy-Schwartz / Parzen	93
Tabela A.10 – Informação Mútua entre variáveis DERMATOLOGIA Método Cauchy-Schwartz / Parzen	94
Tabela A.11 – Informação Mútua entre variáveis CÂNCER DE MAMA Método Cauchy-Schwartz / Parzen	98
Tabela A.12 – Informação Mútua entre variáveis DOENÇAS CARDÍACAS Método Cauchy-Schwartz / Parzen	103

“Quanto maior a esfera do conhecimento, maior a superfície de contato com o desconhecido.”

(Autor desconhecido)