1
Introdução

1.1
Considerações Iniciais e Objetivo

Seleção de variáveis tem fundamental importância em sistemas de classificação, tais como Redes Neurais. Nesta dissertação, enfoca-se o Seletor de Variáveis Baseada em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U). Este algoritmo objetiva selecionar variáveis que sejam relevantes em relação à variável desfecho (variável explicada, de resultado ou de saída) e, paralelamente, reduzir a redundância. Para tanto, baseia-se, como o próprio nome revela, em conceitos da Teoria da Informação, quais sejam, entropia e informação mútua. Quando as variáveis envolvidas são discretas, o cálculo para obtenção de tais "medidas", baseadas na definição de Shannon, é simples e direto, visto que tanto a distribuição conjunta quanto as marginais podem ser estimadas meramente através da contagem das amostras. No entanto, quando pelo menos uma das variáveis em questão é contínua, o cálculo, envolvendo integração, torna-se difícil face ao número limitado de amostras. Uma solução, comumente utilizada, é a incorporação da discretização dos dados como um passo de pré-processamento, sendo a densidade desconhecida estimada através do histograma. Os cálculos ficam, assim, reduzidos a simples somatórios. Nem sempre, porém, a discretização é clara e feita de forma adequada. Exatamente, neste ponto, situa-se o objetivo desta dissertação: apresentar, como alternativa, um método, baseado na informação mútua quadrática de Cauchy-Schwartz e na entropia quadrática de Rényi, esta combinada ao famoso método de estimação de densidade Janela de Parzen, tornando, dessa forma, os cálculos diretos, sem necessidade de um passo de pré-processamento. E, por fim, aplicar ambos os métodos em conjuntos de dados reais e comparar os resultados. O escopo desta dissertação está restrito a tal comparação da ordem de seleção pelos dois métodos. Não está no escopo, portanto, a comparação, pela aplicação de um classificador, em termos de qual conjunto de variáveis selecionadas apresenta maior poder

explicativo em relação ao desfecho, o que fica como sugestão para trabalhos futuros.

1.2 Estrutura da Dissertação

A dissertação está organizada em mais cinco capítulos, além deste de introdução, e mais o apêndice e o anexo. O capítulo 2 reúne fundamentos teóricos da teoria da informação, sendo expressos de maneira descritiva, entre eles, os conceitos de entropia e informação mútua. O Capítulo 3 apresenta o Seletor de Variáveis Baseado em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U). O capítulo 4 apresenta os dois métodos de estimação da entropia e da informação mútua, através dos quais o algoritmo MIFS-U realizará a seleção de variáveis. Este capítulo apresenta, sucintamente, o método que utiliza a definição de entropia de Shannon e o histograma como estimador de densidade. Já o método que aborda a informação mútua de Cauchy-Schwartz e a entropia de Rényi, combinada, intrinsecamente, com o estimador de densidade Janela de Parzen, é apresentado em detalhes, pois constitui o cerne desta dissertação. No capítulo 5, são descritas as bases de dados reais utilizadas nos experimentos e apresentados os respectivos resultados da comparação entre os métodos. No capítulo 6, são feitas, por fim, algumas considerações. O apêndice apresenta as tabelas de informação mútua entre as variáveis, com respeito aos dois métodos abordados. E, finalmente, o anexo apresenta a prova do teorema da integração do produto de núcleos Gaussianos