6
Considerações Finais

Seleção de variáveis é um problema fundamental em muitas diferentes áreas do conhecimento. Todas as variáveis podem ser importantes dentro de um determinado contexto, mas para algum conceito específico, apenas um subconjunto pequeno de variáveis é normalmente relevante. Além disso, seleção de variáveis aumenta a inteligibilidade de um modelo, ao mesmo tempo reduzindo a dimensionalidade e a necessidade de armazenamento. Vários estudos experimentais mostraram que variáveis irrelevantes e redundantes podem reduzir drasticamente a exatidão preditiva de modelos construídos a partir dos dados. Nesta dissertação, abordou-se o Seletor de Variáveis Baseado em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U). Este algoritmo, como foi mostrado, envolve o cálculo da entropia e da informação mútua, com respeito a variáveis discretas e contínuas. No primeiro caso, o cálculo é direto, mas para variáveis contínuas, há inevitáveis integrais em todas as definições de entropia e de informação mútua, que traduzem-se na dificuldade principal depois da estimação da densidade. Portanto, a estimação da densidade e as medidas de entropia e informação mútua devem ser escolhidas adequadamente, de forma que as integrais correspondentes possam ser simplificadas. Mostrou-se como a entropia quadrática de Rényi e a informação mútua quadrática de Cauchy-Schwartz, ao invés da entropia e informação mútua de Shannon, podem ser combinadas com a função núcleo Gaussiana para estimar densidades, resultando em um método geral e efetivo para o cálculo da entropia e informação mútua, sem a necessidade de qualquer hipótese acerca da densidade desconhecida – em quase todos os problemas do mundo real, a única informação disponível está contida nos dados coletados. Deve-se sempre ter em mente que o processo de seleção de variáveis deve ser o mais exato possível, sem, contudo, perder a sua simplicidade. Na prática, simplicidade se torna uma consideração suprema. Se tal processo envolver técnicas muito complexas, ele acaba se tornando em si mesmo um problema, ao invés de ser um facilitador para uma fase posterior de

PUC-Rio - Certificação Digital Nº 0510468/CA

classificação, por meio, por exemplo, da aprendizagem de uma Rede Neural Artificial (RNA).

Experimentos com dados reais foram realizados, comparando-se o Método Cauchy-Schwartz / Parzen-Rosenblatt (CSPR), apresentado nesta dissertação, com o Método Shannon / Histograma (SH), largamente utilizado, baseado na definição de entropia de Shannon, e que faz uso da discretização das variáveis contínuas, como um passo de pré-processamento dos dados. Os resultados, enfocando-se o conjunto das cinco primeiras variáveis selecionadas, foram semelhantes. Como a comparação foi meramente especulativa, ressalta-se que uma análise mais cuidadosa deve ser realizada, através da aplicação de um classificador (ou mais de um), para que se possa comparar o efetivo desempenho dos conjuntos de variáveis selecionadas pelo MIFS-U, com base em ambos os métodos. Além disso, é altamente recomendada a participação de um profissional da área de conhecimento a que dizem respeito as bases de dados abordadas nesta dissertação, pois permitiria, certamente, uma avaliação ainda melhor dos métodos.O que fica como (auto-)sugestão para trabalhos futuros. O método CSPR trabalha diretamente com os dados, propiciando, teoricamente, maior exatidão. O SH que faz uso da discretização, o que, em princípio, poderia mascarar alguma “informação” relevante dos dados, é mais simples, o que explica a sua larga utilização. Uma análise, mais profunda, portanto, se seguirá a esta dissertação.

PUC-Rio - Certificação Digital Nº 0510468/CA