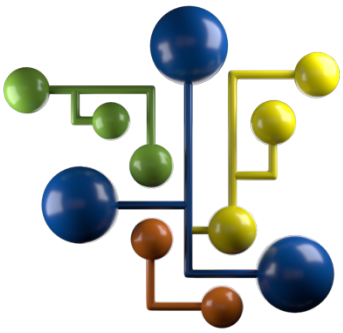

Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Classificação Por Compressão de Dados

Outra maneira de pensar sobre a classificação é como um problema na compressão de dados. Um algoritmo de compressão sem perdas considera uma sequência de símbolos, detecta padrões repetidos no mesmo e escreve uma descrição da sequência mais compacta do que a original. Por exemplo, o texto “0,142857142857142857” pode ser comprimido para “0, [142857]*3”. Os algoritmos de compressão trabalham com a construção de dicionários de subsequências do texto e depois se referem à entradas no dicionário. O exemplo aqui só tinha uma entrada no dicionário, “142857”.

Com efeito, os algoritmos de compressão estão criando um modelo de linguagem. O algoritmo LZW em particular modela diretamente uma distribuição de probabilidade de entropia máxima. Para fazer a classificação por compressão, primeiro reunimos todas as mensagens de spam de treinamento e as comprimimos como uma unidade. Fazemos o mesmo para o ham. Então, quando uma nova mensagem é fornecida para classificar, nós a anexamos às mensagens de spam e comprimimos o resultado. Também a anexamos ao ham e comprimimos. Qualquer classe que comprima melhor — adiciona o menor número de bytes adicionais para a nova mensagem — é a classe prevista. A ideia é que uma mensagem de spam tenderá a compartilhar entradas de dicionário com outras mensagens de spam e, portanto, vai comprimir melhor quando anexada a uma coleção que já contém o dicionário de spam.

Experimentos com classificação baseada em compressão em alguns dos corpora-padrão para classificação do texto — 20 conjuntos de dados de novos grupos corpora Reuters-10, corpora do setor da indústria — indicam que, enquanto a execução da compressão de algoritmos prontos para uso como gzip, RAR e LZW pode ser bastante lenta, sua precisão é comparável à de algoritmos de classificação tradicional. Isso é interessante em seu próprio direito e também serve para destacar que há promessa para os algoritmos que usam caractere n-gramas diretamente sem pré-processamento de texto ou seleção de característica: eles parecem estar capturando alguns padrões reais.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig