RESUMO DO TRABALHO

A THOROUGH EXPLOITATION OF DISTANCE-BASED META-FEATURES FOR AUTOMATED TEXT CLASSIFICATION

Classificação Automática de Texto (CAT) têm adquirido notória importância em uma variedade de tarefas, como a categorização de notícias, organização de bibliotecas digitais, criação de diretórios da web, análise de sentimentos em conteúdos gerados por usuários e detecção de spam. Dado um conjunto de documentos de treinamento classificados em uma ou mais categorias predefinidas, a tarefa do CAT é aprender automaticamente como classificar novos documentos (não classificados), usando uma combinação de atributos desses documentos que os associam a categorias. Devido ao fato de o problema do CAT ocorrer em vários contextos, diversos algoritmos de aprendizado de máquina foram propostos para lidar com CAT.

Embora o próprio algoritmo de classificação tenha um papel importante na CAT, os atributos que representam documentos podem ser igualmente importantes para determinar a eficácia da classificação. Especificamente, representar documentos em um espaço de atributos é um trabalho que precede a CAT, pois esses algoritmos de classificação são projetados para descobrir padrões discriminativos usando esses atributos. Nesse sentido, uma tarefa importante consiste em promover a manipulação espaço de atributos para abordar a CAT do ponto de vista da engenharia de dados. Nesse contexto, abordamos o problema de aprender a classificar textos de forma automática, explorando informações derivadas de meta-atributos, ou seja, atributos criados a partir da representação original dos documentos. Particularmente, os meta-atributos explorados contam com medidas de distância capazes de sumarizar relacionamentos potencialmente complexos entre documentos e apresentar informações relevantes para classificação.

Neste trabalho, não apenas propomos novos meta-atributos que fornecem evidências discriminativas para classificação, mas também novos mecanismos para analisar e selecionar meta-atributos. Nesse sentido, utilizamos estratégias multiobjetivo capazes de minimizar o número de meta-atributos e maximizar a eficácia da classificação, considerando a adequação dos meta-atributos selecionados a uma coleção de dados ou método de classificação específico. Além disso, fornecemos contribuições adicionais para aprimorar a eficiência e a eficácia da utilização de meta-atributos. Em particular, propomos o uso de GPUs (Graphical Processing Units) para reduzir o tempo computacional da geração de meta-atributos, o uso de aprendizado supervisionado para o enriquecimento dos relacionamentos de distância com dados rotulados, e a construção de novos meta-atributos específicos para o contexto da análise de sentimento.