FEATURE ANALYSIS IN EVOLVING DATA STREAMS: ISSUES AND ALGORITHMS

Doutorado

Jean Paul Barddal, Fabricio Enembreck

Contexto: Este projeto de tese é dedicado à fluxos contínuos de dados com mudanças de relevância em seus atributos (feature drifts). Este tipo de mudança ocorre quando um subconjunto de atributos se torna, ou deixa de ser, relevante para a tarefa de aprendizagem. Por mais que este tipo de mudança tenha sido citado em trabalhos pioneiros da área, maior parte dos algoritmos existentes assume que os mesmos atributos são igualmente relevantes durante todo o processo. Este trabalho inclui uma revisão sistemática do tópico, abrangendo uma definição formal para feature drifts e técnicas capazes de lidar com este problema. Adicionalmente, novos geradores de dados são propostos e os existentes capazes de sintetizar este tipo de problema são apresentados. Estes geradores são utilizados para avaliar os algoritmos levantados e detectores de mudanças de conceito. As contribuições deste projeto incluem a proposta de dois novos operadores dinâmicos derivados da Teoria da Informação, i.e., Entropia Condicional e Incerteza Simétrica, que são capazes de dinamicamente verificar a relevância dos atributos durante o processamento dos fluxos de dados. Estes operadores são utilizados como parte de (i) um processo de ponderação dinâmica de atributos nos classificadores Naive Bayes, k-Vizinhos mais próximos e Hoeffding Adaptive Tree, e (ii) um algoritmo de seleção dinâmica de atributos, que por sua vez, é independente do classificador utilizado. Os resultados obtidos até o momento mostram que ambas as abordagens provêm melhorias significativas na acurácia de todos os classificadores, contudo, induzindo um aumento em seus custos computacionais.

Palavras-chave: Classificação de Fluxos Contínuos de Dados; Feature Drift; Seleção de Atributos; Ponderação de Atributos