I Congresso Brasileiro de
Jovens Pesquisadores em
Matemática Pura e Aplicada

Proposta de um Algoritmo para Indução de Árvores de Classificação para Dados Desbalanceados

Marcelo Lauretto* e Claudio Frizzarini

*EACH-USP, São Paulo, Brasil

Resumo

As técnicas de mineração de dados, e mais especificamente de aprendizado de máquina, têm se popularizado enormemente nos últimos anos, passando a incorporar os Sistemas de Informação para Apoio à Decisão, Previsão de Eventos e Análise de Dados. Por exemplo, sistemas de apoio à decisão na área médica e ambientes de Business Intelligence fazem uso intensivo dessas técnicas, envolvendo particularmente árvores de decisão. A mineração de informação e conhecimento a partir de grandes bases de dados tem sido reconhecida como tema chave de pesquisa em sistemas de banco de dados e aprendizado de máquina.

Concomitantemente a essa popularização, faz-se necessário o desenvolvimento de ferramentas de modelagem acessíveis, conceitualmente simples e com baixa necessidade de parametrização, que possam ser utilizadas (ao menos em análises mais simples) por profissionais que não sejam necessariamente especialistas nos métodos de modelagem subjacentes.

Algoritmos indutores de árvores de classificação, particularmente os algoritmos TDIDT (Top-Down Induction of Decision Trees), figuram entre as técnicas mais comuns de aprendizado supervisionado. A construção de uma árvore de decisão consiste em partições sucessivas do conjunto de treinamento original em subconjuntos menores. Uma das vantagens desses algoritmos em relação a outros é que, uma vez construída e validada, a árvore tende a ser interpretada com relativa facilidade, sem a necessidade de conhecimento prévio sobre o algoritmo de construção. Em um contexto de mineração de dados, mesmo que não sejam necessariamente utilizadas na classificação de novas instâncias, árvores de classificação podem ser construídas para fornecer descrições (na forma de regras de classificação) das características comuns aos membros de cada classe.

Todavia, são comuns problemas de classificação em que as frequências relativas das classes variam significativamente. Algoritmos baseados em minimização do erro global de classificação tendem a construir classificadores com baixos erros de classificação nas classes majoritárias e altos erros nas classes minoritárias. Esse fenômeno pode ser crítico quando as classes minoritárias representam eventos como a presença de uma doença grave (em um problema de diagnóstico médico) ou a inadimplência em um crédito concedido (em um problema de análise de crédito).

Diversos algoritmos TDIDT não possuem métodos adaptativos automáticos, demandando a calibração de parâmetros ad-hoc de custos ou, na ausência de tais parâmetros, a adoção de métodos de balanceamento dos dados. As duas abordagens não apenas introduzem uma maior complexidade no uso das ferramentas de mineração de dados para usuários menos experientes, como também nem sempre estão disponíveis.

Este trabalho apresenta uma descrição e alguns resultados empíricos de um algoritmo TDIDT em desenvolvimento, para construção de árvores na presença de dados desbalanceados. Esse algoritmo, denominado atualmente DDBT (Dynamic Discriminant Bounds Tree), utiliza um critério de partição de nós que, ao invés de se basear em frequências absolutas de classes, compara as proporções das classes nos nós com as proporções do conjunto de treinamento original, buscando formar subconjuntos com maior discriminação de classes em relação ao conjunto de dados original. Para a rotulação de nós terminais, o algoritmo atribui a classe com maior prevalência relativa no nó em relação à prevalência no conjunto original. Essas características fornecem ao algoritmo a flexibilidade para o tratamento de conjuntos de dados com desbalanceamento de classes, resultando em um maior equilíbrio entre as taxas de erro em classificação de objetos entre as classes.