UDESC UNIVERSIDADE DO ESTADO DE SANTA CATARINA
Seminário de Iniciação Científica
Universidade do Estado de Santa Catarina
29º SIC UDESC

Comparação de Algoritmos de Classificação Online Semi-Supervisionados

Gabriel Ziemann Ferreira[1], Fabiano Baldo[2]

[1]Acadêmico(a) do Curso de Ciência da Computação CCT - bolsista PROBIC/UDESC
[2] Orientador, Departamento de Ciência da ComputaçãoCCT– fabiano.baldo@udesc.br.

Palavras-chave:Aprendizado de Máquina, Aprendizado *Online*,*Massive Online Analysis*, Arvores de Decisão

O projeto envolveu a análise e comparação de diversos algoritmos de classificação com foco para o aprendizado *online*, onde há constante geração de novos dados. Devido a este fluxo, o modelo precisa estar em contínua atualização, com incorporação periódica dos novos dados aomodelo de forma que ele mantenha sua eficácia com o passar do tempo.

Para fazer melhor uso dos dados recebidos pelo fluxo, os quais raramente possuem classe definida,o projeto abrange também o aprendizado semi-supervisionado, onde instâncias sem classe definida podem ser incorporadas ao conjunto de treinamento. Calculando a proximidade entre uma instância não classificada e um conjunto de instâncias classificadas, o modelo é capaz de inferir a possível classe dainstânciacom um certo nível de confiança da resposta, que é utilizado para decidir se ela é ou não atribuída a instância sem classe.

Um dos principais desafios do aprendizado *online* é a chamadamudança de conceito, que ocorre quando o padrão dos dados sofre uma mudança externaà qual o modelo precisa se adaptar. A detecção damudança de conceito depende de detecção por parte da plataforma ou do algoritmo, criando um histórico dos valores de dados considerados "normais" ou detectando perdas bruscas nas fases de treinamento. Algoritmos que realizam substituição periódica de modelos, como variantes *online* do*bagging*, possuem adaptabilidade natural às mudanças.

Para os experimentos, foi utilizada a ferramenta *Massive Online Analisys* (MOA), da universidade de Waikato. O MOA é o sucessor da ferramenta de mineração de dados WEKA, também da universidade de Waikato, focado no treinamento e execução de modelos *online*, com versões adaptadas dos algoritmos de aprendizado de máquina tradicionais. Os testes foram realizados no modo *test-then-train*, que alterna unitariamente entre treinamento e classificação.

Os algoritmos testados para classificaçãoforam: Very Fast Decision Tree (VFDT), Random Tree, Random Forest, J48, M5' com regressão, Naive Bayes e Bagging. Para o algoritmo de Bagging foi utilizada a variante OzaBag que é online eusa janela adaptável.Oquantidade de instâncias

utilizadas foi definida como 10, pois testes com números maiores demonstraram pouco ou nenhum ganho de acurácia a um acréscimoconsiderável de tempo de processamento. Como classificadores agregados do OzaBag foram usadosos algoritmos Random Tree, VFDT e J48.

As bases utilizadas para teste foram obtidas do repositório público da Universidade da Carolina Irvine (UCI), EUA e convertidas para o formato ARFF, utilizado no MOA. As bases sãocompostas de dados reais obtidos de agências governamentais e instituições de diversos países, e compreendem dados como informações do censo, fenômenosnaturais e estudos biológicos entre outros.

No quesito da variedade dos dados, as bases contêm alta diversidade entre si, com algumas contendo somente atributos categóricos, outras somente numéricos, e outras uma combinação deles. Ainda, algumas das bases contém atributos de valor desconhecido, cujas lacunas precisam ser tratadas pelos algoritmos. As bases "Bank" e "EEG" foram particularmente úteis por conter ocorrências de mudança de conceito.

A performance foi avaliada utilizando a métrica da Área Sob a Curva ROC(AUC ROC), que é uma métrica calculada utilizando a área sobre a curvade razão entre verdadeiros positivos e falsos positivos previstos pelo modelo. Essa métrica é calculada à medida que um limitante de sensitividade é incrementado de zero a um. Ela pode ser vista como a união das curvas de probabilidade de verdadeiro positivo e negativo.Por considerar as áreas e interseção das previsões corretas ao invés de sua média, a curva ROC se mostra uma alternativa superior à medida de acurácia normalmente utilizada no aprendizado de máquina.

Em geral, as melhores performances foram obtidaspelos modelos Bagging com uma AUC média de 88% para VFDT e 82% para Random Tree, em sequência estão o algoritmo de regressão aplicada à arvore M5' com AUC média de 87% e o Naive Bayes com AUC média 87%,sendo que esse último apresenta quedas mais bruscas na presença de mudança de conceito.

É importante ressaltar que os dois últimos algoritmos mencionados são modelos únicos, ao contrário doBagging, que utiliza vários modelos com uso de memória e processamento por cada um. Por outro lado, os modelos que utilizamBagging demonstraram maior adaptabilidade às mudanças de conceito, devido àmultiplicidade dos classificadores e a substituição periódica dos piores classificadores individuais.