Projeto de Pesquisa - Bolsa de Pós-Doutorado Júnior

Conselho Nacional de Desenvolvimento Científico e Tecnológico

Candidato

Davi Pereira dos Santos

Instituição

Universidade de São Paulo - Instituto de Ciências Matemáticas e de Computação
Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional
Avenida Trabalhador São-carlense, 400 - Centro CEP: 13566-590 - São Carlos - SP

Supervisor

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

Título

Recomendação automática de estratégias de aprendizado ativo

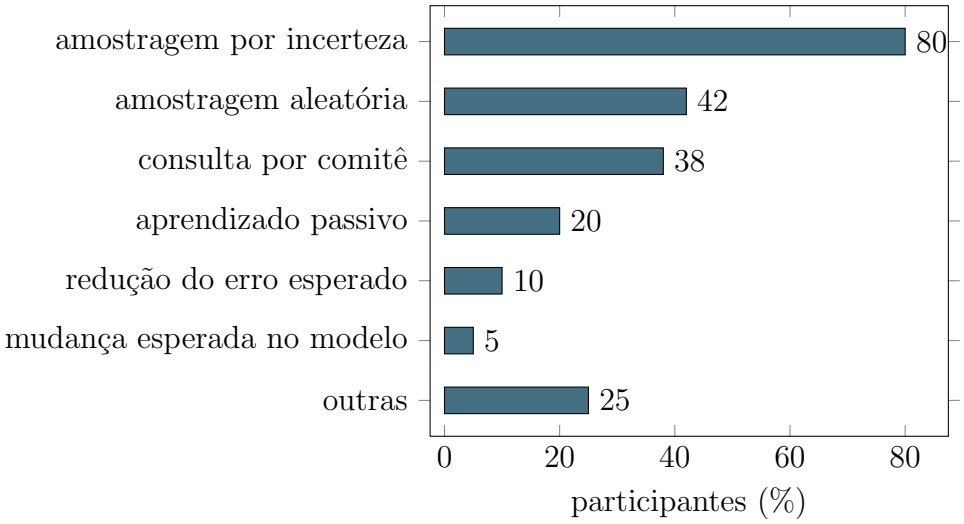
Introdução e Justificativa

Atualmente, o aprendizado de máquina permeia o cotidiano humano provendo auxílio em tarefas diversas. Seu bom desempenho, na tarefa de classificação, depende da existência de dados categorizados de qualidade para treinamento do sistema e de um algoritmo de aprendizado adequado. A categorização dos dados é um processo frequentemente custoso, pois é realizada por um supervisor humano que atribui categorias/classes para cada objeto/exemplo de interesse contido nos dados ou, dependendo da aplicação, é realizada por um processo químico ou mecânico, por exemplo. REF Dados os limites práticos de orçamento, esforço humano disponível, resistência à fadiga e crescente presença massiva de dados, faz-se necessária a *amostragem* de apenas um subconjunto de exemplos para a rotulação e subsequente construção do conjunto de treinamento. Quando se deseja eficiência no uso de recursos, essa amostragem de exemplos não é trivial, pois resultados teóricos e empíricos apontam para a existência de uma diferença de desempenho entre os métodos desenvolvidos na área de *aprendizado ativo* frente à amostragem aleatória. REF Apesar de ser uma área promissora, a diversidade de métodos de amostragem ativa aliada às diferentes características de cada base de dados e aos vieses dos algoritmos de aprendizado dificulta a fase de projeto de um sistema de aprendizado de máquina. Existem assim, dois problemas de escolha: estratégia de amostragem e algoritmo de aprendizado.

Tradicionalmente, a escolha do algoritmo de aprendizado é feita com base na experiência pessoal do especialista responsável pelo sistema ou em restrições próprias da aplicação. Outra possibilidade é o uso de um método automático de recomendação como o *meta-aprendizado*. REF Apesar dessas abordagens serem indicadas para o caso do aprendizado supervisionado convencional, em que todos os exemplos de treinamento já estão rotulados, elas não foram propostas para a situação de escassez ou ausência de rótulos, que é o caso de aplicações com orçamento limitado e ainda na fase de escolha da melhor forma de amostragem para rotulação.

Idealmente, a escolha do algoritmo de aprendizado é baseada em métodos de validação cruzada. Logo, a escolha se dá com o conjunto de treinamento completo, ou seja, amostrado e rotulado. Consequentemente, a escolha do processo de amostragem, precede

Figura 1: Frequência de uso de estratégias na competição de aprendizado ativo (alguns participantes adotaram mais de uma estratégia). *Adaptado de (?).*



a determinação do algoritmo de aprendizado. A escolha da estratégia de amostragem é mais crítica que a escolha do algoritmo de aprendizado, tendo-se em vista que o sucesso da estratégia escolhida só pode ser determinado após ter-se incorrido em custos financeiros com a atividade de supervisão. Dessa forma, é preciso adotar uma estratégia com base em expectativas de desempenho ou de acordo com características da base de dados. Na prática, essa escolha tem sido arbitrária, tendendo a se concentrar no uso da estratégia mais simples, chamada *amostragem por incerteza*. Essa preferência foi reportada numa competição de aprendizado ativo, onde também prevaleceu a ausência de estratégias possivelmente mais efetivas como as baseadas em densidade conforme Figura 1. Em consonância com esse panorama, está a ausência de estudos comparativos abrangentes que possam guiar o especialista na fase de amostragem.

Adicionalmente, diversas estratégias fazem uso interno de algoritmos de aprendizado para embasar a amostragem. Esse uso interno fecha um círculo impossível de dependências: a estratégia depende do algoritmo; o algoritmo depende da existência do conjunto de treinamento que requer rotulação; e, a rotulação é feita pela estratégia de aprendizado ativo.

A forma trivial de evitar o círculo de dependências, para além da amostragem aleatória, é a adoção de estratégias agnósticas. Essas estratégias não fazem suposições a respeito dos dados no que diz respeito ao viés de aprendizado, pois não requerem um algoritmo interno. Elas se baseiam, poe exemplo, em medidas de densidade ou estatísticas de agrupamento. REF A dispensa do algoritmo, entretanto, leva a amostragens que não consideram a fronteira de decisão que seria traçada durante o aprendizado. Ela contém informações que permitem consultas mais prospectivas que exploratórias, potencialmente acelerando a descoberta dos exemplos mais relevantes.

Assim, existem pelo menos dois problemas e respectivos subproblemas em aberto na área de aprendizado ativo:

- Que situações dispensam o uso de aprendiz (algoritmo interno)?

  Dentre as estratégias agnósticas, qual a mais indicada para um dado problema?

- Como quebrar o círculo de dependências?

Dentre as estratégias gnósticas, qual a mais indicada para uma dado problema?

Objetivo

A ideia é usar meta-aprendizado não supervisionado para indicar qual par estratégia-classificador é o mais adequado, pois idealmente é preferível não fixar o classificador previamente

Assim se resolvem os problemas: se é agnostica ou gnostica, qual estratégia dentro do grupo, e qual variante da estratégia (métrica de distância, etc.)

Plano de atividades

Cronograma