Pontifícia Universidade Católica
do Rio de Janeiro

Julio Cesar Duarte

O Algoritmo Boosting at Start e suas Aplicações

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio como requisito parcial para obtenção do título de Doutor em Ciências - Informática

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Setembro de 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Julio Cesar Duarte

O Algoritmo Boosting at Start e suas Aplicações

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do título de Doutor em Ciências - Informática. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática — PUC–Rio

Prof. Marcus Vinícius Soledade Poggi de Aragão
Departamento de Informática — PUC–Rio

Prof. Bianca Zadrozny
Universidade Federal Fluminense — UFF

Prof. Raúl Pierre Rentería
Departamento de Informática — PUC–Rio

Prof. Geraldo Bonorino Xexéo
Universidade Federal do Rio de Janeiro — UFRJ

Prof. Flávio Luis de Mello
Universidade Federal do Rio de Janeiro — UFRJ

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 08 de Setembro de 2009

Julio Cesar Duarte

Graduou–se em 1998 no Instituto Militar de Engenharia (Rio de Janeiro) em Engenharia de Computação. Em 2003, obteve título de Mestre em Informática pela Pontifícia Universidade Católica do Rio de Janeiro. Trabalha desde 1998 no Centro Tecnológico do Exército desenvolvendo sistemas de Guerra Eletrônica.

Ficha Catalográfica

Duarte, Julio Cesar

O Algoritmo Boosting at Start e suas Aplicações / Julio Cesar Duarte; orientador: Ruy Luiz Milidiú. — Rio de Janeiro : PUC–Rio, Departamento de Informática, 2009.

v., 87 f: il. ; 29,7 cm

1. Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. 2. Aprendizado de Máquina. 3. Processamento de Linguagem Natural. 4. Algoritmos de Comitê. 5. Boosting. 6. AdaBoost. 7. Boosting At Start. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Ao meu professor e orientador, Ruy Luiz Milidiú, pela orientação, apoio, incentivo e, principalmente, pela amizade desenvolvida nesses oito anos de convívio.

À minha esposa, Ana Paula, pela força, carinho e compreensão em todos os momentos.

Ao meu pai, que sempre me apoiou e incentivou minha vida de estudos.

Ao Centro Tecnológico do Exército, pelo apoio e liberação de tarefas para a dedicação a esta tese.

À Banca examinadora, pelos conselhos e críticas positivas ao trabalho.

Aos meus companheiros da PUC–Rio: Cícero Santos, Roberto Cavalcante, Eraldo Rezende, Breno Faria e Frederico Pessoa pelo auxílio em diversas tarefas relacionadas a esta tese.

Ao Programa de Pós-Graduação em Informática da PUC–Rio, pelo excelente ambiente acadêmico.

Em especial, à minha mãe, que com certeza esteve ao meu lado durante toda a redação desta tese e que, por durante um longo período, abdicou de sua vida para que eu pudesse me tornar quem eu sou. Por isso, terei eternamente uma dívida de gratidão.

Resumo

Duarte, Julio Cesar; Milidiú, Ruy Luiz. **O Algoritmo Boosting at Start e suas Aplicações**. Rio de Janeiro, 2009. 87p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Boosting é uma técnica de aprendizado de máquina que combina diversos classificadores *fracos* com o objetivo de melhorar a acurácia geral. Em cada iteração, o algoritmo atualiza os pesos dos exemplos e constrói um classificador adicional. Um esquema simples de votação é utilizado para combinar os classificadores. O algoritmo mais famoso baseado em Boosting é o AdaBoost. Este algoritmo aumenta os pesos dos exemplos em que os classificadores anteriores cometeram erros. Assim, foca o classificador adicional nos exemplos mais difíceis. Inicialmente, uma distribuição uniforme de pesos é atribuída aos exemplos. Entretanto, não existe garantia que essa seja a melhor escolha para a distribuição inicial. Neste trabalho, apresentamos o Boosting at Start (BAS), uma nova abordagem de aprendizado de máquina baseada em Boosting. O BAS generaliza o AdaBoost permitindo a utilização de uma distribuição inicial arbitrária. Também apresentamos esquemas para determinação de tal distribuição. Além disso, mostramos como adaptar o BAS para esquemas de Aprendizado Semi-supervisionado. Adicionalmente, descrevemos a aplicação do BAS em diferentes problemas de classificação de dados e de texto, comparando o seu desempenho com o algoritmo AdaBoost original e alguns algoritmos do estado-da-arte para tais tarefas. Os resultados experimentais indicam que uma modelagem simples usando o algoritmo BAS gera classificadores eficazes.

Palavras–chave

Aprendizado de Máquina. Processamento de Linguagem Natural. Algoritmos de Comitê. Boosting. AdaBoost. Boosting At Start.

Abstract

Duarte, Julio Cesar; Milidiú, Ruy Luiz. **The Boosting at Start Algorithm and its Applications**. Rio de Janeiro, 2009. 87p. PhD Thesis — Department of Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Boosting is a Machine Learning technique that combines several *weak* classifiers with the goal of improving the overall accuracy. In each iteration, the algorithm updates the example weights and builds an additional classifier. A simple voting scheme is used to combine the classifiers. The most famous Boosting-based algorithm is AdaBoost. This algorithm increases the weights of the examples that were misclassified by the previous classifiers. Thus, it focuses the additional classifier on the hardest examples. Initially, an uniform weight distribution is assigned to the examples. However, there is no guarantee that this is the best choice for the initial distribution. In this work, we present Boosting at Start (BAS), a new Machine Learning approach based on Boosting. BAS generalizes AdaBoost by allowing the use of an arbitrary initial distribution. We present schemes for the determination of such distribution. We also show how to adapt BAS to Semi-supervised learning schemes. Additionally, we describe the application of BAS in different problems of data and text classification, comparing its performance with the original AdaBoost algorithm and some state-of-the-art algorithms for such tasks. The experimental results indicate that a simple modelling using the BAS algorithm generates effective classifiers.

Keywords

Machine Learning. Natural Language Processing. Ensemble Algorithms. Boosting. AdaBoost. Boosting At Start.

Sumário

Lista de figuras

Lista de tabelas

Os fatos fazem todos parte apenas do problema, não da solução.

Ludwig Wittgenstein, *Tractatus Logico-Philosophicus.*