



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



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

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

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, a 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

1	Introdução	12
1.1	Tema	12
1.2	Objetivos	14
1.3	Metodologia	14
1.4	Estrutura da Tese	16
2	Boosting at Start	17
2.1	Boosting	17
2.2	AdaBoost	18
2.3	Boosting com uma distribuição arbitrária	24
2.4	Corretude do Algoritmo BAS	26
3	BAS Genético: Uma abordagem evolutiva	30
3.1	Algoritmos Genéticos	30
3.2	Modelagem Genética Utilizada	31
4	Comitê BAS	34
4.1	Inicialização dos pesos	34
4.2	Combinação dos Classificadores	39
4.3	Descrição do algoritmo	41
5	Algoritmos-Base de Referência	43
5.1	Toco de Decisão	43
5.2	Aprendizado Baseado em Transformações	44
5.3	TBL Genético	46
5.4	Aprendizado de Transformação guiado pela Entropia	47
6	Experimentos com Tarefas de Classificação de Dados	51
6.1	Medidas de Qualidade	51
6.2	Modelagem BAS para Classificação de Dados	52
6.3	Descrição dos Conjuntos de Dados	54
6.4	Software e Hardware	56
6.5	Resultados	56
6.6	Sumário	62
7	Experimentos com Tarefas de PLN	64
7.1	Medidas de Qualidade	64
7.2	Modelagem BAS para PLN	65
7.3	Corpora de Anotação Morfossintática	67
7.4	Corpora de Anotação de Sintagmas	67
7.5	Software e Hardware	69
7.6	Resultados	69
7.7	Sumário	75
8	Conclusões e Trabalhos Futuros	78

Lista de figuras

2.1	Dados do exemplo de utilização do AdaBoost.	20
2.2	Iterações do exemplo de utilização do AdaBoost.	21
2.3	Classificador final do exemplo de utilização do AdaBoost.	21
2.4	Novo conjunto de exemplos para execução do AdaBoost.	22
2.5	Iterações do AdaBoost para o novo conjunto.	22
2.6	Classificador AdaBoost final para o conjunto modificado.	23
2.7	Distribuição inicial de pesos para o novo conjunto utilizado.	23
2.8	Iterações do algoritmo para o novo conjunto utilizando uma distribuição arbitrária.	23
2.9	Classificador final gerado para o novo conjunto.	24
3.1	Exemplos de cromossomos.	31
3.2	Exemplo da codificação genética para o BAS.	32
3.3	Exemplo do operador de cruzamento.	33
3.4	Exemplo do operador de mutação.	33
4.1	Esquema de aprendizado do Comitê BAS.	41
5.1	Exemplos de tocos de decisão para atributos categóricos e numéricos.	43
5.2	Aprendizado de transformação guiado pela entropia.	48
5.3	Exemplo de árvore de decisão para a tarefa <i>Weather</i> .	48
6.1	Comparação entre as estratégias supervisionada e semi-supervisionada para a instância <i>bca</i> .	63

Lista de tabelas

1.1	Características dos corpora de texto utilizados.	15
1.2	Desempenho comparativo do algoritmo BAS nas tarefas de PLN.	15
4.1	Conjuntos de gabaritos e regras para o conjunto CMC.	40
5.1	Exemplo de uma lista de possíveis TAS para a codificação genética.	46
5.2	Exemplo de uma codificação genética para o TBL-GA.	46
5.3	Exemplo de gabaritos gerados para o TBL-GA.	47
5.4	Exemplo de lista de gabaritos para a tarefa <i>Weather</i> .	49
6.1	Exemplos das instâncias do conjunto de dados <i>Weather</i> .	52
6.2	Possíveis pesos iniciais para o algoritmo Comitê BAS.	53
6.3	Conjuntos de dados utilizados nos experimentos supervisionados.	54
6.4	Concatenação das classes para tarefas de classificação não-binária.	55
6.5	Conjuntos de dados utilizados nos experimentos semi-supervisionados.	55
6.6	Resultados para a abordagem BAS genético.	57
6.7	Resultados para a abordagem Comitê BAS com votação pela melhor maioria.	57
6.8	Resultados para a abordagem Comitê BAS com votação ETL.	58
6.9	Acurácias Médias para todas as abordagens supervisionadas.	59
6.10	Quantidade total de classificadores-base por abordagem.	60
6.11	Acurácias médias para todas as abordagens comparadas.	61
6.12	Comparação da abordagem semi-supervisionada com alguns algoritmos de estado-da-arte.	62
7.1	Exemplo de anotação morfossintática do Português.	67
7.2	Corpora de anotação morfossintática.	68
7.3	Exemplo de anotação de sintagmas do Português.	68
7.4	Corpora de anotação de sintagmas.	69
7.5	Desempenho dos algoritmos na tarefa de anotação morfossintática do Português.	70
7.6	Desempenho dos algoritmos na tarefa de anotação morfossintática do Português Histórico.	71
7.7	Desempenho dos algoritmos na tarefa de anotação morfossintática do Inglês.	72
7.8	Desempenho dos algoritmos na tarefa de anotação morfossintática do Alemão.	72
7.9	Desempenho dos algoritmos na tarefa de extração de sintagmas nominais do Português.	73
7.10	Desempenho dos algoritmos na tarefa de extração de sintagmas nominais do Inglês.	74
7.11	Desempenho dos algoritmos na tarefa de extração de sintagmas do Inglês.	75
7.12	Desempenho dos algoritmos na tarefa de extração de sintagmas do Hindi.	75

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

Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*.