
Relações entre Ranking, Análise ROC e Calibração em Aprendizado de Máquina

Edson Takashi Matsubara

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Relações entre Ranking, Análise ROC e Calibração em Aprendizado de Máquina¹

Edson Takashi Matsubara

Orientadora: *Prof^a Dr^a Maria Carolina Monard*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP como parte dos requisitos necessários à obtenção para do título em Doutor em Ciências de Computação e Matemática Computacional.

USP - São Carlos
julho/2008

¹Trabalho Realizado com Auxílio da FAPESP Proc. No: 2005/03792-9

*Aos meus pais,
Ryuiti e Ritie,*

*Aos meus irmãos,
Koiti e Sayuri,*

À Maria Carolina Monard.

Agradecimentos

Gostaria de agradecer a Deus por estar sempre ao meu lado, sempre me indicando caminhos que me levam a sua bondade, por nunca ter desistido de mim, por mostrar que nenhuma passagem pode ser sem esforço, por me dar coragem, por mostrar os valores que realmente valem, por colocar pessoas tão boas e especiais em meu caminho, por cuidar de cada simples detalhe em minha vida.

Como diria o poeta, mesmo que tivesse em minhas mãos todo o perfume das rosas, toda a beleza do céu, toda a pureza dos anjos, toda a inocência das crianças, toda a grandeza do mar, toda a força das ondas, mesmo que eu tivesse todas as coisas belas da vida e todos os belos lugares do mundo nada teria sentido se eu não tivesse o presente mais valioso, mais nobre e mais sagrado que o Senhor pode me dar... minha família. Agradeço meu pai Ryuiti, por me ensinar muito, mas muito mesmo, sobre todas as coisas da vida. Agradeço minha mãe Ritie por seu imenso carinho e amor. Agradeço aos meus irmãos Koiti, Sayuri, e agora a mais nova irmã Miwa, por serem os melhores irmãos que alguém pode ter.

Mestres nos conduzem, não somente ao conhecimento, mas também ao saber. Agradeço à professora Carolina, por tudo que tem me ensinado, pelos conselhos e pelas conversas, talvez a senhora nem saiba mas sem você, eu não teria feito pós-graduação. Tenho muito carinho por você. Gostaria de agradecer também ao professor Peter Flach e sua esposa Lisa por terem me ensinado muito no ano em que passei pela Universidade de Bristol. Por terem me acolhido em sua casa e pelas horas de utilizávamos os ladrilhos da cozinha me explicando sobre curvas ROC. Gostaria de agradecer aos professores Huei e Paulo, por serem exemplos de dedicação e ensino na formação de pessoas. Gostaria de agradecer também à professora Solange e ao professor André pelo apoio ao meu trabalho.

Além dessas pessoas que considero grandes mestres, existem as pessoas que dão um toque muito especial a minha vida. Considero essas pessoas

grandes amigos. O rio é a mistura de pequenos encontros e quando unidas possuem grande força. Assim considero as amizades que faço nos caminhos da vida. Muitas vezes essas amizades de distanciam de nós, mas elas estão sempre lá, sempre agregando força a esse rio, querendo o nosso bem gratuitamente. Gostaria muito de agradecer aos amigos que fiz aqui em São Carlos. Aos amigos de república Mauro, Ronaldo, Sidão, Danielzinho e Marcio. À dois grandes amigos, Gustavo Batista e Richardson pelos bons momentos projetando e fazendo aviõezinhos e pelas conversas das mais diversas coisas. Pelos amigos que fiz na graduação, Marcio, Alex, Kleber, Testa e Zóid. Agradeço a todos vocês por momentos onde a alegria, descontração e amizade se fizeram presentes.

Cada um que passa em nossa vida, passa sozinho, pois cada pessoa é única e nenhuma substitui outra. Também gostaria de agradecer aos amigos que fiz na Inglaterra: Sebastian, Virgínia, Tarek, Bill, Ksenia, Susanna e Rob. Ao pessoal do LABIC: André Maletzke, André Rossi, Andrés Ferrero, Brunos Ferres, Caneca, Capoli, Camila, Claudia Martins, Claudia Milaré, Christiane, Damiance, Débora, Edson Melanda, Eduardo Spinosa, Evandro, Fabiano, Flávia, Igor, Jaqueline, Jean, Katti, Leonardo, Lorena, Magaly, Marcio Basgalupp, Marcos Cintra, Marcos Quiles, Mariza, Merley, Murilo, Patrícia Rufino, Rafael Giusti, Renatinho, Roberta, Robson, Rodrigo Bianchi, Rodrigo Calvo e Valmir. Gostaria de agradecer à Pâmela, por ser tão única em minha vida e ter me ensinado coisas que somente você poderia ter me ensinado.

Agradeço ao pessoal da pós-graduação do ICMC, à Beth, à Laura, à Ana Paula, por serem tão atenciosas a cada um de nós pós-graduandos. É incrível como vocês decoram o nome de todos nós. Também à Marília por seus maravilhosos coffe breaks. Você tem um papel determinante na presença dos alunos e professores nas palestras.

Agradeço também as pessoas que mantêm a USP funcionando, como Paulinho, Dotta, Sonia, Dagoberto, Cabral, Seu Arly e tantos outros que vão do setor administrativo ao faxineiro e jardineiro.

Finalmente gostaria de agradecer à FAPESP pela minha bolsa de doutorado, à CAPES pela minha bolsa de doutorado sandwich e ao ICMC-USP, pelo suporte e estrutura disponibilizados para o desenvolvimento de minha formação.

Abstract

Supervised learning has been used mostly for classification. In this work we show the benefits of a welcome shift in attention from classification to ranking. A ranker is an algorithm that sorts a set of instances from highest to lowest expectation that the instance is positive, and a ranking is the outcome of this sorting. Usually a ranking is obtained by sorting scores given by classifiers. In this work, we are concerned about novel approaches to promote the use of ranking. Therefore, we present the differences and relations between ranking and classification followed by a proposal of a novel ranking algorithm called LEXRANK, whose rankings are derived not from scores, but from a simple ranking of attribute values obtained from the training data. One very important field which uses rankings as its main input is ROC analysis. The study of decision trees and ROC analysis suggested an interesting way to visualize the tree construction in ROC graphs, which has been implemented in a system called PROGROC. Focusing on ROC analysis, we observed that the slope of segments obtained from the ROC convex hull is equivalent to the likelihood ratio, which can be converted into probabilities. Interestingly, this ROC convex hull calibration method is equivalent to Pool Adjacent Violators (PAV). Furthermore, the ROC convex hull calibration method optimizes Brier Score, and the exploration of this measure leads us to find an interesting connection between the Brier Score and ROC Curves. Finally, we also investigate rankings build in the selection method which increments the labelled set of CO-TRAINING, a semi-supervised multi-view learning algorithm.

Resumo

Aprendizado supervisionado tem sido principalmente utilizado para classificação. Neste trabalho são mostrados os benefícios do uso de *rankings* ao invés de classificação de exemplos isolados. Um *rankeador* é um algoritmo que ordena um conjunto de exemplos de tal modo que eles são apresentados do exemplo de maior para o exemplo de menor expectativa de ser positivo. Um *ranking* é o resultado dessa ordenação. Normalmente, um *ranking* é obtido pela ordenação do valor de confiança de classificação dado por um classificador. Este trabalho tem como objetivo procurar por novas abordagens para promover o uso de *rankings*. Desse modo, inicialmente são apresentados as diferenças e semelhanças entre *ranking* e classificação, bem como um novo algoritmo de *ranking* que os obtém diretamente sem a necessidade de obter os valores de confiança de classificação, esse algoritmo é denominado de LEX-RANK. Uma área de pesquisa bastante importante em *rankings* é a análise ROC. O estudo de árvores de decisão e análise ROC é bastante sugestivo para o desenvolvimento de uma visualização da construção da árvore em gráficos ROC. Para mostrar passo a passo essa visualização foi desenvolvido um sistema denominado PROGROC. Ainda do estudo de análise ROC, foi observado que a inclinação (coeficiente angular) dos segmentos que compõem o feixe convexo de curvas ROC é equivalente a razão de verossimilhança que pode ser convertida para probabilidades. Essa conversão é denominada de calibração por feixe convexo de curvas ROC que coincidentemente é equivalente ao algoritmo PAV que implementa regressão isotônica. Esse método de calibração otimiza Brier Score. Ao explorar essa medida foi encontrada uma relação bastante interessante entre Brier Score e curvas ROC. Finalmente, também foram explorados os *rankings* construídos durante o método de seleção de exemplos do algoritmo de aprendizado semi-supervisionado multi-descrição CO-TRAINING.

Sumário

Sumário	xiii
Lista de Figuras	xv
Lista de Tabelas	xvii
Lista de Abreviaturas	xix
Lista de Algoritmos	xxi
1 Introdução	1
1.1 Motivação	1
1.2 Objetivos	1
1.3 Principais Contribuições	1
1.4 Organização	1
2 Aprendizado de Máquina	3
3 Ranking	5
4 Análise ROC	7
5 Calibração	9
6 Aplicando Análise ROC e Rankings em CO-TRAINING	11
7 Conclusões	13
7.1 Resumo dos Objetivos e Principais Resultados	13
7.2 Limitações	13
7.3 Trabalhos Futuros	13
Referências	15

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AD *Árvore de Decisão*

AM *Aprendizado de Máquina*

AUC *Area Under ROC Curve*

EM *Expectation Maximization*

FN *Falso Negativo*

FP *Falso Positivo*

IA *Inteligência Artificial*

KDD *Knowledge Discovery in Databases*

LABIC *Laboratório de Inteligência Computacional*

LR *Likelihood Ratio*

MAP *Maximum a Posteriori*

MCAR *Missing Completely at Random*

MT *Mineração de Texto*

NB *Naive Bayes*

OD *Odds Ratio*

PAV *Pool Adjacent Violators*

PET *Probability Estimation Trees*

ROC *Receiver Operating Characteristic*

SVM *Support Vector Machines*

TFP Taxa de Falso Positivo

TVP Taxa de Verdadeiro Positivo

VN Verdadeiro Negativo

VP Verdadeiro Positivo

Lista de Algoritmos

Introdução

Neste capítulo é apresentada uma descrição geral desta tese, com o objetivo de fornecer uma visão geral dos problemas tratados e dos objetivos principais do trabalho de pesquisa realizado. O capítulo está organizado da seguinte maneira: na Seção 1.1 é apresentada a motivação sobre o tema de pesquisa tratado nesta tese; na Seção 1.2 são apresentados os objetivos do trabalho em forma de questionamentos, que serão respondidos no decorrer dos capítulos; na Seção 1.3 são apresentados brevemente as principais contribuições deste trabalho; por fim, na Seção 1.4 é apresentada a organização da tese, com uma descrição resumida do conteúdo abordado em cada capítulo.

1.1 Motivação

1.2 Objetivos

1.3 Principais Contribuições

1.4 Organização

Aprendizado de Máquina

(Mitchell, 1997) é um livro de referência em aprendizado de máquina.

Em Mitchell (1997) descreve os principais algoritmos de aprendizado de máquina.

Ranking

Análise ROC

Calibração

Aplicando Análise ROC e Rankings em CO-TRAINING

Conclusões

Neste capítulo são apresentadas as conclusões deste trabalho. Na Seção 7.1 é realizado um paralelo entre os objetivos desta tese e os resultados obtidos. Na Seção 7.2 são discutidas algumas limitações das soluções propostas e na Seção 7.3 são apresentadas algumas direções de trabalhos futuros.

7.1 Resumo dos Objetivos e Principais Resultados

7.2 Limitações

7.3 Trabalhos Futuros

Referências Bibliográficas

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York. Citado na página 3.