

poster:02

## Algoritmos de Aprendizado de Máquina para Predição de Resultados das Lutas de MMA

Leandro A. A. Silva, Johnnatan Messias, Mirella M. Moro,  
Pedro Olmo Vaz de Melo, Fabricio Benevenuto

<sup>1</sup>Departamento de Ciência da Computação, UFMG, Belo Horizonte, Brasil

{leandroaraujo, johnnatan, mirella, olmo, fabricio}@dcc.ufmg.br

**Resumo.** Este artigo propõe a utilização de algoritmos de aprendizado de máquina que prediz o resultado de lutas de MMA baseados nas características de dois lutadores e dos seus adversários recentes. Os resultados da avaliação experimental mostram uma abordagem para criação de uma base de dados aplicável a esportes individuais e que um dos algoritmos avaliados é capaz de obter sucesso em 67% das predições.

**Abstract.** This paper proposes using machine learning algorithms to predict the outcome of an MMA fight based on the characteristics of the two fighters and their recent opponents. Our experimental evaluation shows an approach to create a dataset applicable to individual sports and one of the evaluated algorithms has 67% of successful predictions.

### 1. Introdução

Esportes de luta sempre tiveram público, e os combates em solo, como o *Mixed Martial Arts* (MMA), têm crescido nos últimos tempos. Tal crescimento é acompanhado pelo seu público que inclui torcedores presentes no evento durante uma luta, espectadores que acompanham pela televisão ou por outro meio de comunicação como a Internet [Kim et al 2008]. A popularidade também é aparente no número de apostadores em todo o mundo. O MMA envolve dois lutadores no *ring* usando vários estilos e técnicas de lutas das artes marciais. Com a alta demanda, em 2006 o *Ultimate Fighting Championship* (UFC), a maior empresa organizadora de MMA do mundo, superou a popularidade do esporte boxe nos canais de TV HBO e WWE. De fato, em 2008 o UFC obteve lucro estimado de US \$250 milhões [Frederick et al 2012].

Nesse contexto, uma abordagem para predição de resultados de lutas é interessante por vários motivos. Primeiro, por se tratar de um mercado crescente, estimar um possível resultado pode tornar as apostas mais inteligentes. Segundo, com acesso a um algoritmo de predição de resultados, os *matchmakers* (pessoas que arranjam lutas) podem arranjar lutas mais parelhas. Terceiro, os *matchmakers* podem tornar popular um lutador: arranjando lutas contra adversários apontados pelo algoritmo como prováveis perdedores. Entretanto, a predição de lutas de MMA é particularmente difícil, pois podem ser decididas em apenas um golpe, às vezes de sorte, o que leva frequentemente o lutador considerado favorito a ser derrotado.

Neste artigo, apresentamos algoritmos baseados em Aprendizado de Máquina (AM) para prever resultados de lutas com base no perfil dos dois lutadores e dos seus adversários recentes. O modelo de predição proposto considera uma base de dados confiável

da qual as características necessárias são coletadas. Especificamente, a base foi extraída do site *Sherdog*<sup>1</sup>, considerado o maior site sobre MMA do mundo [Schneiderman 2006]. Os resultados da avaliação experimental mostram que um dos algoritmos avaliados é capaz de obter sucesso em 67% das predições.

A seguir, a Seção 2 discute trabalhos relacionados. A Seção 3 apresenta a metodologia utilizada para a coleta e a extração dos dados. A Seção 4 apresenta as predições das lutas juntamente com os algoritmos utilizados, o ranking das *features* e a classificação dos dados. Finalmente, a Seção 5 conclui este artigo.

## 2. Trabalhos Relacionados

Existem diferentes propostas para mensurar estatisticamente o desempenho de times de um determinado esporte como, por exemplo, futebol, baseball e basquete [Ben-Naim and Hengartner 2007, Ben-Naim et al 2013, Eli et al. 2006, Vaz de Melo et al. 2008]. Entretanto, apesar dos diversos estudos realizados, ainda existem desafios a serem solucionados como, por exemplo, a predição de resultados de jogos de maneira eficaz e eficiente.

Especificamente [Vaz de Melo et al 2012] propõem dois modelos (*NetFor* e *NetForY*) baseados em rede para prever o comportamento de equipes de ligas esportivas. Esses modelos não possuem parâmetros, isto é, são completamente livres de qualquer configuração, e podem ser aplicados diretamente em qualquer campeonato esportivo de equipe. Neste caso, foram aplicados nas ligas esportivas de equipe *National Basketball Association* (NBA) e *Major League Baseball* (MLB). O modelo proposto obteve bons resultados quando comparado com outros modelos, atingindo uma melhoria de acurácia na predição de ranking de aproximadamente 14% em relação ao seu melhor concorrente.

[Lemmer et al. 2014] também utilizaram dois modelos para prever resultados de partidas de críquete. O sucesso da predição é difícil porque se dois times jogam duas ou mais partidas um contra o outro, e se cada um vence algumas vezes, os resultados serão inconsistentes. Então como desafio, eles buscaram uma maneira de compensar esses resultados inconsistentes. O primeiro modelo utilizado teve taxa de sucesso de 76.4%, que antes era de apenas 56.8%. O segundo modelo obteve uma melhoria de 52.7% para 70.9%. O mesmo método pode ser usado em qualquer série de esportes onde uma equipe joga contra outra mais de uma vez, como o futebol, por exemplo.

Mais relacionado ao nosso estudo, [Procópio Jr. et al. 2012] apresentam um método para cálculo de ranking para redes sociais esportivas que considera fatores temporais. Especificamente, eles propuseram uma modelagem para essas redes de modo que os pesos das arestas decaem com o tempo. Isso é necessário para capturar a noção de que os resultados dos confrontos antigos não são tão importantes quanto os mais recentes para prever um resultado. O método proposto foi aplicado em dois esportes: tênis e MMA. Como resultado, obtiveram rankings com maior acurácia do que os da literatura, que ignoram os fatores temporais.

Diferentemente dos mencionados que atuam na predição de rankings ou comportamento de times e análises estatísticas sobre o desempenho dos times esportivos, nosso trabalho propõe a predição do resultado de lutas de MMA utilizando algoritmos de AM.

<sup>1</sup>Sherdog UFC and MMA: <http://sherdog.com>

### 3. Metodologia

Esta seção resume a metodologia empregada para coleta e extração de dados.

#### 3.1. Coleta dos Dados

Seja  $G = (E, N)$  um grafo onde cada nó  $n \in N$  representa um lutador e cada aresta  $e_{n_1 n_2} \in E$  implica que  $n_1$  e  $n_2$  já se enfrentaram pelo menos uma vez. Para construir o grafo de MMA, realizamos uma busca em largura (*BFS* [Moore et al 1959]) aplicada a  $G$  a partir de um nó considerado a semente da busca. Então, os dados de todos os lutadores só são explorados se há um único componente conectado a partir do nó semente.

Essa metodologia foi aplicada nos dados coletados do website *Sherdog*[Schneiderman 2006]. O nó semente foi o campeão mundial José Aldo Junior, e o período de lutas considerado foi de 25 de abril 1980 a 1 de julho de 2015. No total, foram coletados dados sobre 75.773 lutadores e 165.841 lutas. É importante notar que a partir do nó semente escolhido, todos os campeões do *UFC* foram encontrados pela busca.

#### 3.2. Extração dos Dados

Após a coleta, foi necessário tratar os dados e colocá-los no formato da nossa aplicação. A coleta retornou dados no formato *HTML*. Porém, neste formato, consultas como “Retorne o nome de todos lutadores que José Aldo enfrentou até 16 de Setembro de 2014” seriam extremamente trabalhosas. Então, foi definido um *parser* para transformar o *HTML* em *JSON* com os seguintes dados: (i) data de nascimento, país de residência, lista de lutas, altura e peso de cada lutador, e (ii) data, nome do evento, adversário, árbitro, resultado, tempo de duração e a forma de término (nocaute, finalização, decisão) para cada luta.

#### 3.3. Características utilizadas

Os algoritmos de AM utilizados recebem como entrada características (*features*) extraídas dos dados. Em resumo, é um problema de classificação de dados: existem características referentes a dois lutadores  $f_a$  e  $f_b$  em cada instância; caso  $f_a$  vença, a classe é +1; caso contrário, a classe é -1 (lutas que terminaram em empate podem ser desconsideradas). Dois conjuntos de características foram consideradas, como detalhado a seguir.

- **Do lutador:** idade, percentual de vitórias, número de vitórias nas últimas cinco lutas, altura, peso, número de vitórias/derrotas (subdivididos em: por decisão, por decisão dividida, por decisão unânime, por nocaute, e por finalização), total de vitórias/derrotas, total de vitórias/derrotas nas últimas cinco lutas, percentual de vitórias contra o adversário, percentual de vitórias contra adversários comuns, totalizando 21 características.
- **De seus cinco adversários mais recentes:** idade, percentual de vitórias, número de vitórias nas últimas cinco lutas, altura, peso, resultado da luta, totalizando  $6 \times 5 = 30$  características.

Como ambos lutadores são importantes para realizar a predição, o total final de características consideradas é:  $(21 + 6 \times 5) \times 2 = 102$ .

### 4. Predição das Lutas

Dentre os vários algoritmos de AM existentes, foram utilizados um algoritmo de Regressão Logística [Peng et al. 2002] e um algoritmo de Floresta Aleatória

**Tabela 1. Acurácia dos algoritmos de classificação**

Algoritmo	Acurácia média	DP
Aleatório	0.5000	–
%Vitórias	0.6456	–
%Vitórias nas últimas 5 lutas	0.6343	–
Regressão Logística	0.6592	0.0255
Floresta Aleatória	0.6390	0.0222

[Breiman 2001]. Tais algoritmos foram escolhidos pois são apropriados para classificação (diferentemente da Regressão Linear, por exemplo) e além de retornar uma classe, eles retornam um “grau” de confiança para cada classificação. Os algoritmos foram implementados em *Python*, com auxílio da biblioteca *scikit-learn*.

#### 4.1. Conjunto de Dados e Algoritmos Base

O conjunto de dados considerado pelos algoritmos é uma amostra do nosso conjunto de lutas. A amostra foi definida de acordo com as seguintes restrições: (i) a luta teve um vencedor; (ii) ambos lutadores realizaram um mínimo de cinco lutas que tiveram um vencedor; (iii) a data máxima de realização da luta é 31/12/2013 para o conjunto de treino; e (iv) a data mínima de realização da luta é 1/1/2014 para o conjunto de testes (ou seja, apenas lutas do passado são utilizadas para fazer as predições). Desse modo, os conjuntos de treino e teste ficaram com 32.878 e 6.290 lutas respectivamente.

Para avaliar os resultados obtidos, foram considerados três algoritmos base: (i) um que retorna aleatoriamente o vencedor da luta; (ii) um que considera o lutador com maior percentual de vitórias como vencedor; e (iii) um que considera o lutador com maior percentual de vitórias nas últimas cinco lutas como vencedor.

#### 4.2. Acurácia dos Algoritmos

Para estimar a acurácia dos algoritmos de Floresta Aleatória, foram utilizados 20 estimadores, profundidade máxima 3, e índice *gini* e entropia como critérios para medir o grau de impureza. Valores entre 1 e 10 foram distribuídos uniformemente para número de características, número mínimo de amostras exigido para dividir um nó interno, e número mínimo de amostras em novos nós-folha. Os conjuntos de treino e teste foram divididos aleatoriamente em cinco partes balanceadas e de mesmo tamanho. Cada uma das partes de treino com cada uma das partes de teste foram devidamente testadas, totalizando  $5 \times 5 = 25$  execuções para cada algoritmo de aprendizado de máquina.

A Tabela 1 apresenta a acurácia de cada algoritmo utilizado (os três base e os dois de aprendizado de máquina). É importante notar que o algoritmo de Floresta Aleatória teve desempenho inferior ao que considera o número de vitórias. Tal resultado pode indicar a presença de viés para esta característica no algoritmo de AM. Além disso, o algoritmo de Regressão Logística foi superior aos demais.

#### 4.3. Ranking das Características

Para entender melhor a acurácia, esta seção apresenta o ranking das características para o algoritmo de Regressão Logística, o qual retorna o coeficiente de ganho de informação

**Tabela 2. Ranking do ganho de informação das características**

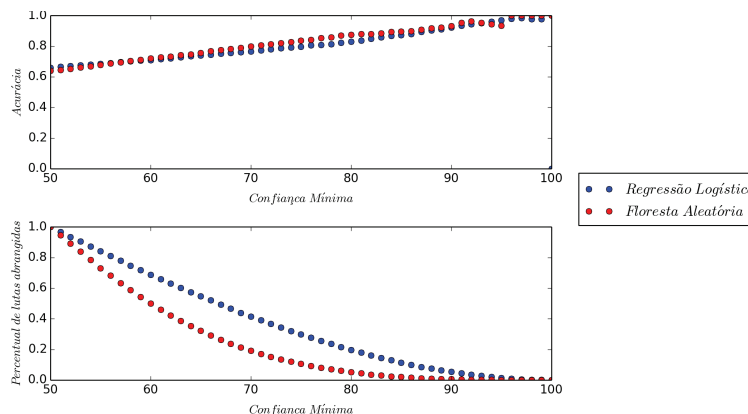
Característica	Coefficiente $f_a$	Coefficiente $f_b$
Derrotas por finalização	-0.21	0.26
% vitórias contra o adversário	0.24	-0.21
% vitórias do 1.º dos últimos 5 adversários	0.18	-0.19
Derrotas por decisão unânime	0.2	-0.17
% vitórias	0.17	-0.18

de cada uma. A Tabela 2 apresenta o ranking das características mais relevantes para a classificação dos dados. São dois coeficientes por característica, um por lutador  $f_a$  e  $f_b$ .

Esses resultados mostram que um alto número de derrotas por finalização de  $f_a$  colabora para a classe  $-1$  (sinal negativo), e um número alto para  $f_b$  colabora para a classe  $+1$ . As análises para as demais características são análogas. Diferentemente do óbvio, idade e altura possuem coeficientes baixos. A partir dessa observação, também foi conferido quantas vezes o lutador mais velho e o mais alto venceram. O resultado foi respectivamente 55.2% e 57% das vezes. Em contrapartida, o lutador que tem maior percentual de vitórias vence em 64.6%.

#### 4.4. Análise Relativa ao Grau de Confiança

Apesar de o algoritmo de Regressão Logística ser o que alcançou melhores valores de predição, o comportamento dos algoritmos de aprendizagem de máquina se comportam de uma forma diferente de acordo com o grau de confiança exigido. Dessa forma, para complementar os resultados, a Figura 1 ilustra a variação de acurácia e o respectivo percentual de lutas abrangidas para valores crescentes de confiança mínima. Dentre os resultados, verifica-se que se considerarmos o grau de confiança mínimo de 70%, as acurácias são de 76.66% e 79.86% para a regressão linear e a floresta aleatória. A correlação de Pearson entre o grau de confiança e a acurácia é de 0.9909, o que indica uma forte correlação.

**Figura 1. Resultados considerando o grau de confiança**

## 5. Conclusões e Trabalhos Futuros

Predizer resultados de qualquer esporte é um desafio com dificuldades associadas às características do esporte em questão. Neste trabalho, apresentamos algoritmos de aprendizado de máquina (Regressão Logística e de Floresta Aleatória) para prever o resultado

da luta com base nas informações de dois lutadores de MMA e seus adversários recentes. Os experimentos mostraram que um deles é capaz de obter sucesso em 67% das predições.

Como trabalhos futuros, planeja-se testar outros algoritmos de Aprendizagem de Máquina (tais como Árvores de Decisão, Redes Neurais, *Support Vector Machines* e k-Vizinhos mais Próximos). Planeja-se também aplicar o algoritmo a organizações específicas (tais como UFC, Bellator, WSOF e Jungle Fight) e permitir que mais classes sejam retornadas, por exemplo, o vencedor e o método como a vitória foi alcançada.

**Agradecimentos.** Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

## Referências

- Ben-Naim, E. and Hengartner, N. W. (2007). Efficiency of competitions. *Phys. Rev. E*, 76(2):026106.
- Ben-Naim et al, E. (2013). Randomness in competitions. *Journal of Statistical Physics*, 151(3-4):458–474.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Eli, B.-N., Federico, V., and Sidney, R. (2006). Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4):1–14.
- Frederick et al, E. L. (2012). Characteristics of users of a mixed- martial-arts blog: A case study of demographics and usage trends. *International Journal of Sport Communication*, 5:109–125.
- Kim et al, S. (2008). An analysis of spectator motives in an individual combat sport: A study of mixed martial arts fans. *Sport Marketing Quarterly*, 17(2):109.
- Lemmer, H., Bhattacharjee, D., and Saikia, H. (2014). A consistency adjusted measure for the success of prediction methods in cricket. *International Journal of Sports Science and Coaching*, 9(3):497–512.
- Moore et al, E. F. (1959). The shortest path through a maze. *Proceedings of the International Symposium on the Theory of Switching*, pages 285–292.
- Peng, C. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, 96:3–14.
- Procópio Jr., P. S., Gonçalves, M. A., Laender, A. H. F., Salles, T., and Figueiredo, D. (2012). Time-aware ranking in sport social networks. *JIDM*, 3(3):195–210.
- Schneiderman, R. (2006). Ultimate fighting’s next battle. [http://www.forbes.com/2006/10/05/sports-martial-arts\\_tech\\_media\\_cx\\_rs\\_1005fight.html](http://www.forbes.com/2006/10/05/sports-martial-arts_tech_media_cx_rs_1005fight.html).
- Vaz de Melo, P. O. S., Almeida, V. A. F., and Loureiro, A. A. F. (2008). Can complex network metrics predict the behavior of nba teams? In *ACM SIGKDD*, pages 695–703.
- Vaz de Melo et al, P. O. S. (2012). Forecasting in the NBA and other team sports: Network effects in action. *ACM Transactions on Knowledge Discovery from Data*, 6(3):13 – 27.