Treinamento Supervisionado para Previsão de Partidas de Futebol: Uma Abordagem usando Dados de Videogames

Bruno Guilherme Gomes[1], **Mário C. G. Moreira**[1], **Pedro H. F. Holanda**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

{brunoguilherme,mcesar,holanda}@dcc.ufmg.br

Resumo. *Encontrar fontes de dados para aplicações de aprendizado de máquina e mineração de dados pode ser uma tarefa muito complicada em alguns casos. Além disso, em muitas aplicações, os dados podem ser bastante escassos. Por essa razão, buscamos uma maneira alternativa de obter dados e informações através da grande e vasta indústria dos jogos eletrônicos, indústria essa que tem conseguido coletar e construir dados tão verossímeis a ponto de poderem ser usados para resolver problemas do mundo real. Assim, este trabalho apresenta uma abordagem para prever os resultados de partidas de futebol utilizando para isso os dados do jogo FIFA.*

Abstract. *Finding data sources for machine learning and data mining applications can be a very complicated task in some cases. Furthermore, in many applications, the data may be quite scarce. Therefore, we sought an alternative way to obtain data and information through the large and wide video game industry, an industry which has been able to collect and build so credible data that they can be used to solve real-world problems. Thus, this paper presents an approach to predict soccer match outcomes using for it the FIFA game data.*

1. Introdução

Um dos principais problemas enfrentados pelos cientistas em suas análises e pesquisas é a ausência de dados e informações. Além deste, outro problema igualmente preocupamente é a dificuldade de obtê-los e/ou acessá-los. Em muitos casos, tais restrições são tão severas que levam à frustração da conclusão do trabalho científico e inclusive ao abandono da própria pesquisa.

Diante dessas dificuldades, o avanço e o aprimoramento de técnicas de análises de dados, aprendizado de máquina, mineração de dados, entre outras, têm contribuído bastante para que se consiga solucionar a carência de dados em certas aplicações. Todavia, por mais complexa que seja uma técnica, é difícil imaginar que ela conseguirá ter êxito lidando com a ausência de dados nas mais diversas aplicações das áreas de medicina, economia, mineração de padrões, entre outras.

Dentro de economia, a área de mercados preditivos é um exemplo. Comumente conhecidos como mercados de apostas, os mercados preditivos são mercados especulativos criados com o propósito de fazer previsões para antecipar ou monitorar um provável evento futuro. E o mercado de apostas esportivas é o que vem ganhando mais adeptos a cada dia. Segundo o *website* Statista[1], o lucro bruto do mercado desse nicho específico

[1]http://www.statista.com/

de aposta foi de 11,5 bilhões de euros em 2012. Com isso, diversas técnicas para coletar informações úteis na previsão de eventos esportivos têm sido buscadas por apostadores e pela banca de aposta no intuito de diminuir o grau de incerteza ao predizer um resultado e, dessa maneira, aumentar o retorno do investimento.

Todavia, é grande a complexidade de se prever um evento esportivo e isso pode ser explicado pela grande dificuldade de se mensurar as diversas variáveis que influenciam o próprio jogo. Características especificadas de cada jogador, problemas psicológicos e físicos dos jogadores, informações climáticas durante a partida são exemplos de algumas dessas variáveis. Mesmo assim, diversos trabalhos na literatura tentam solucionar o problema da previsão de partidas de futebol. Geralmente estes trabalhos se atêm a dados estatísticos das partidas como, quantidade de gols, número de cartões distribuídos, porcentagem de posse de bola, quantidade de faltas ocorridas e etc. São exemplos disso [Constantinou et al. 2013] e [Dixon and Coles 1997].

Sendo assim, na tentativa de buscar novas fontes de dados para previsão de partidas de futebol, voltamos nossos olhares para a indústria dos jogos eletrônicos, uma indústria que vem investindo milhões de dólares, ano após ano, na tentativa de proporcionar a seus usuários um experiência cada vez mais próxima à realidade. Os jogos eletrônicos podem ser usados como fonte de dados em modelagens do mundo real, e, em especial, que os dados do jogo de videogame FIFA podem ser usados para solucionar problemas de previsão de resultados de partidas de futebol. Uma vez que, para conferirem um maior realismo ao jogo, as habilidades e performances dos jogadores são mensuradas em números, essa geração de dados cria também uma quantidade de informação capaz de aumentar o conhecimento sobre o universo dos esportes e, com isso, diminuir a incerteza ao se fazer uma previsão de um evento [Cover and Thomas 2006].

Esse trabalho propõe, para isso, uma metodologia supervisionada para se prever resultados de partidas de futebol a partir de dados de caracterização dos jogadores do jogo de videogame FIFA.

2. Trabalhos Relacionados

Diversos trabalhos têm utilizado métodos de inteligência computacional para fazer a predição de eventos esportivos. De fato, dado o volume de capital que gira dentro do mercado de apostas, o interesse em se modelar as ocorrências nos eventos esportivos e a busca de uma acurácia cada vez maior na predição de resultados têm levado muitos cientistas a se dedicarem a esses estudos.

Nesse contexto, grande parte dos trabalhos se diferenciam em alguns aspectos: a forma de predição pretendida do resultado, tais como números de gols ou vitória-empate-derrota, os modelos de parametrização adotados e as fontes de dados para previsões esportivas. Contudo, algumas características permeiam diversos trabalhos na área de previsão de eventos. Uma estratégia comum encontrada em grande parte deles é o uso de diversas técnicas de inteligência artificial.

Revemos na literatura o uso de diferentes algoritmos de aprendizado de máquina na previsão de resultados de jogos. Em [Joseph et al. 2006] são usadas redes bayesianas e outras técnicas de aprendizado de máquina, incluindo árvores de decisão e k-vizinhos mais próximos para a realização dessa tarefa. [Hucaljuk and Rakipović 2011] usam diversos algoritmos fazendo comparação de diferentes algoritmos de aprendizagem para prever

os resultados dos jogos da Liga Europeia, tais como redes neurais artificiais, naïve bayes, redes bayesianas, floresta aleatória e regressão logística. Já [Tsakonas et al. 2002] utilizam lógica difusa, redes neurais artificiais e programação genética para executar a tarefa de previsão e também fazer comparações. Técnicas de programação genética também são usadas em [Cui et al. 2013], obtendo resultados bastante satisfatórios.

[Constantinou et al. 2013] apresentam um estudo detalhado do mercado de apostas e explicita as razões do uso de redes bayesianas enquanto modelos não-paramétricos para alcançar bons resultados. Os autores fazem uma análise dos dados a serem usados em uma rede bayesiana para que esta apresente bons resultados e concluem que a estratégia adotada apresenta uma acurácia melhor que a apresentada por *bookmakers*, ou apostadores profissionais.

No que tange as fontes de dados para previsões, os trabalhos que não utilizam *bookmakers* fazem amplo uso de estatísticas de jogos para melhorarem seus resultados [Dixon and Coles 1997].Além disso, na literatura há uma carência de trabalhos que busque caminhos alternativos para prever eventos esportivos, sendo o presente trabalho um precursor no uso de dados de jogos eletrônicos para modelar eventos esportivos reais.

3. Coleta de Dados e Caracterização

3.1. Coleta dos Dados

Na coleta dos dados foi implementado um *crawler web* para minerar as informações referentes às características de todos os jogadores da English Premier League. Essas informações foram retiradas do *website* SoFIFA[2]. Nesse site, os jogadores são descritos por 33 características valoradas entre 0 e 100 pontos e agrupadas nos seguintes conjuntos: Ataque, Habilidade, Movimentação, Força, Mentalidade, Defesa e Goleiro (Tabela 1). Estas características são usadas no videogame FIFA para simular a ação de cada jogador durante as partidas.

Tabela 1. Características dos Jogadores

Ataque	Habilidade	Movimentação	Força	Mentalidade	Defesa	Goleiro
Cruzamento	Dribles	Aceleração	Força de Chute	Agressividade	Marcação	Elasticidade
Finalização	Curva	Pique	Impulsão	Interceptação	Dividida em pé	Manejo
Precisão do Cabeceio	Precisão nas Faltas	Agilidade	Fôlego	Posicionamento	Carrinho	Chute
Passe Curto	Lançamento	Reação	Força	Visão de Jogo		Posicionamento
Voleio	Controle de Bola	Equilíbrio	Chute de Longe	Pênaltis		Reflexos

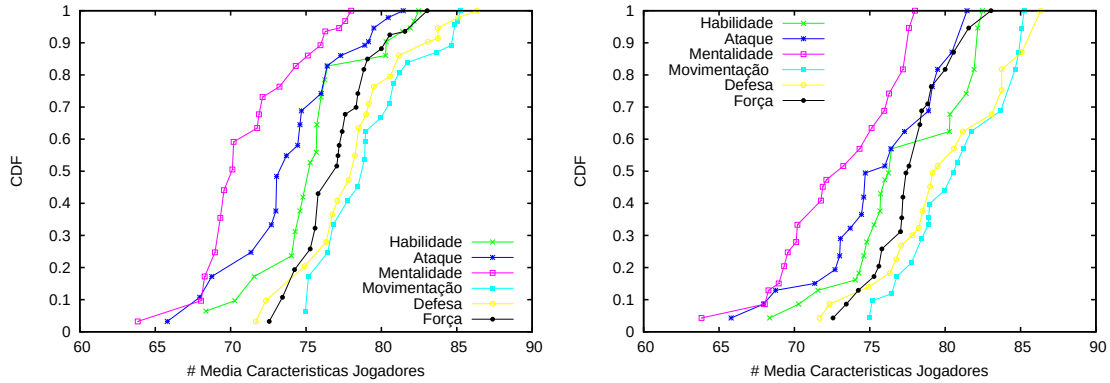
Os resultados das partidas da English Premier League temporada 2011-2012 foram obtidos através do site *http://www.football-data.co.uk* e representam um conjunto de 380 partidas jogadas por 20 times.

3.2. Caracterização dos dados

As Figuras 1(a) e 1(b) mostram a função de distribuição acumulada (CDF) em relação à média por time das características de um jogador e subdividas por subgrupos e a correlação com as derrotas obtidas em casa de um time mandante e as vitórias ocorridas fora de casa de um time visitante. Para esses gráficos consideramos apenas, para cada time, os 4 melhores jogadores do subgrupo Ataque, isto é, os 4 jogadores com maior soma

[2]http://sofifa.com/players

das características desse subgrupo, além dos 4 melhores do subgrupo Defesa, e os 5 melhores dos de cada um dos subgrupos Habilidade, Movimentação, Força e Mentalidade. Na Figura 1(a) podemos observar que no subgrupo Ataque os times que tiveram médias superiores a 76 perdem em casa em aproximadamente 10% das partidas. Em contrapartida, a característica Habilidade representa aproximadamente 15% das partidas para esse mesmo valor. Podemos notar ainda que as curvas seguem a dedução natural, ou seja, à medida que a média de um subgrupo de características aumenta a probabilidade de um time perder em casa tende a diminuir. Isso também pode ser constatado na Figura 1(b), porém neste caso as suavidades nas curvas tendem a amplificar as probabilidades, uma vez que aumentando a média de um subgrupo de características a probabilidade de ocorrer vitória fora de casa aumenta.



(a) Time Mandante - Derrotas em Casa

(b) Time Visitante - Vitórias Fora de Casa

Figura 1. CDF das Médias das Características agrupados por Times e a Correlação entre as Vitórias e Derrotas.

4. Metodologia Proposta e Resultados

A metodologia constrói um modelo de previsão por meio de análise de comparação das médias das características dos jogadores dos times e dos jogadores com as melhores médias das características dos times. Essas médias irão balizar os padrões dos times que apresentam melhor desempenho no decorrer da temporada 2011-2012 da English Premier League.

4.1. Agrupando *Features*

Para agrupar as *features* que foram usadas no processo de classificação, criou-se uma função de agregação que combina tanto características individuais dos jogadores quanto médias dos times.

Primeiramente, seleciona-se do subgrupo Ataque os 4 jogadores com maiores somas deste subgrupo e, de mesmo modo, 4 do subgrupo Defesa, 5 de cada um dos subgrupos Habilidade, Movimentação, Força e Mentalidade, e 1 do subgrupo Goleiro. A escolha dessas quantidades de jogadores se justifica por existirem sobreposições de características em um mesmo jogador, isto é, o jogador com melhor ataque pode ser também o jogador com melhor movimentação. Testamos outras quantidades de seleção de jogadores, porém pelas nossas análises o modelo que obteve melhores resultados utilizou essa seleção de jogadores. Depois disso, para cada uma das 33 características faremos uma média por time

e somaremos aos valores das características dos jogadores selecionados na fase anterior. Podemos sintetizar nossa função da seguinte forma:

$$\textit{Feature}\big(\text{time } x\big) = \begin{cases} \sum \big(\text{Top 4 Ataq}\big) + \mu\big(\text{Ataq (time } x)\big) \\ \sum \big(\text{Top 4 Defe}\big) + \mu\big(\text{Defe (time } x)\big) \\ \sum \big(\text{Top 5 Habil}\big) + \mu\big(\text{Habil (time } x)\big) \\ \sum \big(\text{Top 5 Movi}\big) + \mu\big(\text{Movi (time } x)\big) \\ \sum \big(\text{Top 5 Força}\big) + \mu\big(\text{Força (time } x)\big) \\ \sum \big(\text{Top 5 Ment}\big) + \mu\big(\text{Ment (time } x)\big) \\ \text{Melhor Goleiro} + \mu\big(\text{Goleiro (time } x)\big) \end{cases}$$

Assim, cada partida irá ser representada por um total de 66 *features*, já que cada jogo é disputada por 2 times.

4.2. Classificação das Partidas e Resultados

Para classificar os resultados das partidas (Empate, Vitória em Casa e Vitória fora de Casa) usaremos o classificador SVM [Cortes and Vapnik 1995] e a Regressão Logística [McCullagh and Nelder 1989], que são modelos de treinamento supervisionado. Esses classificadores irão aprender uma função de inferência através de um conjunto de treinamento. Assim, os classificadores serão treinados por meio do *5 fold cross validation One vs All*, que neste caso irá representar 80% das partidas para treino e 20% para teste. A precisão dos classificadores será dada pela média de cada *fold* (Tabela 2).

Tabela 2. Acurácia dos Classificadores

Modelo	Vitória em casa	Vitória fora de casa	Empate	Acurácia
Linear SVM	76.83%	67.89%	58.94%	67.88%
RBF SVM	75.52%	66.05%	58.67%	66.70%
Polly SVM	75.52%	65.78%	58.68%	66.66%
Regressão Logística	75.29%	67.10%	58.68%	67.01%

Podemos observar que o classificador Linear SVM obteve um bom desempenho para classificar os dados, alcançando uma média de 67.88% de precisão. Observa-se que foram alcançados resultados similares aos publicados em [Constantinou et al. 2013], em que a acurácia foi de 66,85%[3]. Os resultados podem ser comparáveis uma vez que se referem ao mesmo conjunto de partidas.

5. Conclusões

Esse trabalho investiga, através de técnicas supervisionadas, o problema da previsão de resultados de partidas de futebol que apresenta vários desafios correlacionados às incertezas inerentes aos próprios eventos. Técnicas de inteligência computacional vem sendo amplamente empregadas para a resolução desse problema. Porém, muitas das vezes essas abordagens sofrem do problema da falta de dados para auxiliar e aprimorar a acurácia das previsões já existentes.

[3]http://pi-football.com/2011-12_predictions.aspx

Com isso, propõe-se o uso de jogos eletrônicos como poderosa fonte de informação capaz de prover dados confiáveis, em virtude do volume de recursos investidos pela gigante indústria de videogames. Foi utilizado, por essa razão, o jogo FIFA com o objetivo de melhorar as previsões de resultados de partidas de futebol.

Essa abordagem atestou que os dados extraídos dessas fontes de informação resultaram em uma estratégia eficaz. Conclui-se assim que os jogos eletrônicos podem conter poderosas informações capazes de auxiliar na solução de problemas do mundo real.

6. Trabalhos Futuros

Em termos de trabalhos futuros, pretende-se concatenar os dados estatísticos das partidas com os dados de jogos eletrônicos e assim investigar essa junção no aprimoramento do estado da arte. Pretende-se ainda testar outros métodos supervisionados, como modelos gráficos probabilísticos, para fazer uma comparação entre os resultados.

Além disso, é possível melhorar a heurística da construção de variáveis usadas nos classificadores e, por fim, pretende-se ainda montar um histórico com as previsões de um campeonato corrente para ser disponibilizado virtualmente.

Agradecimentos. CNPq, FAPEMIG e CAPES. Aos laborátorios do DCC Wisemap e Locus

Referências

Constantinou, A. C., Fenton, N. E., and Neil, M. (2013). Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems*, 50:60–86.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory 2nd edition*. Wiley-interscience.

Cui, T., Li, J., Woodward, J. R., and Parkes, A. J. (2013). An ensemble based genetic programming system to predict english football premier league games. In *Evolving and Adaptive Intelligent Systems (EAIS), 2013 IEEE Conference on*, pages 138–143. IEEE.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *MIPRO, 2011 Proceedings of the 34th International Convention*, pages 1623–1627. IEEE.

Joseph, A., Fenton, N. E., and Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Tsakonas, A., Dounias, G., Shtovba, S., and Vivdyuk, V. (2002). Soft computing-based result prediction of football games. In *The First International Conference on Inductive Modelling (ICIM'2002). Lviv, Ukraine*. Citeseer.