Utilização de redes neurais para predição de ocorrência de escanteios no final de um dos tempos de uma partida de futebol

Adriel Filipe Lins Alves da Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
aflas@cin.ufpe.br

Daniel Cabral da Costa
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
dcc5@cin.ufpe.br

Juracy Emanuel Magalhães da Franca
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
jemf@cin.ufpe.br

Mario Diego Ferreira dos Santos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mdfs@cin.ufpe.br

Pedro Victor Eugencio de Souza
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
pves@cin.ufpe.br

***Resumo*—O mercado de apostas esportivas movimenta bilhões de dólares em todo o mundo. As bancas, cada vez mais, vêm utilizando cientistas de dados para coletar e predizer com maior precisão resultados de uma partida. Além disso, muitos trabalhos vem utilizando algoritmos de *Machine Learning*-(ML) e Redes Neurais Artificiais-(RNAs) para predizer o resultado de uma partida de futebol. Porém nenhum trabalho apresentou uma proposta para predição de escanteios em futebol, que faz parte de uma das modalidades de apostas. Este artigo apresenta o desempenho de das RNAs na tarefa de predizer a ocorrência de um escanteio nos a mais no final de cada tempo de uma partida. Foram testadas as redes *Multilayer perceptron* - MLP, Máquinas de Vetores de Suporte e a combinação destas redes com os extratores de características *Principal Component Analysis*-PCA e *Autoencoders*. Os resultados obtidos mostram que podem ser utilizadas arquiteturas diferentes a depender do perfil do *trader*: agressivo, moderado ou conservador. A combinação de diferentes arquiteturas em diferentes contextos possibilita um melhor retorno do investimento *Return on Investiment* - ROI. Os experimentos provaram que as redes podem prever com acurácia acima de 60%, sendo adequada para alcançar um ROI positivo podendo ser acessados na página[1].**

***Index Terms*—Redes Neurais, Rede Perceptron Multicamada, Máquina de Vetores de Suporte, Apostas Esportivas, Futebol.**

I. Introdução

O mercado de apostas esportivas, principalmente na modalidade de futebol, é um mercado que movimenta uma quantia substancial de dinheiro. Estima-se que esse mercado tenha movimentado no Brasil cerca de R$ 7 bilhões de reais apenas no ano de 2020. Embora denuncias de manipulação de resultados em ligas menores tenham sido realizadas, ainda não há evidências que confirmem essas denúncias em ligas com maior visibilidade como as ligas inglesa, espanhola, alemã, entre outras.

[1]link com o código fonte do algoritmos: https://github.com/yriuss/NN_pred

O mercado de apostas esportivas possui algumas similaridades ao mercado financeiro. Por exemplo, o mercado de apostas esportivas é um mercado de soma-zero, ou seja, há um *trader* em cada lado da operação. Portanto, enquanto um lado ganha, ou outro perde semelhante ao mercado de trade. Na aposta esportiva, as casas de aposta oferecem um preço com base nas probabilidades de ocorrência de um determinado evento, se esse evento ocorre, o *trader* que investiu naquele evento resgata o investimento com lucro e o trader que investiu no evento adverso perde toda a quantia [1].

As bancas de apostas esportivas utilizam times de cientistas de dados com o objetivo de modificar as *odds* de forma que as probabilidades favoreçam ao mercado. No entanto, quando esses essas probabilidades são calculadas de forma errada, surgem as chances para que o mercado seja vencido e os *traders* obtenham lucro nas operações [2].

O mercado de apostas esportivas é bastante diversificado, não limitando-se ao resultado final de uma partida, cada evento em uma partida pode ser utilizado como cartões amarelos, gols, laterais. Todos esses eventos pertencem a dois grandes mercados: os mercados de *over* e *under*, nos quais os *traders* apostam na ocorrência de uma quantidade X de eventos para cada categoria.

Enquanto muitos trabalhos concentram-se na predição do resultado de uma partida os mercados de *over* e *under* são menos explorados e possuem ótimas oportunidades, caso do mercado de escanteios. O mercado de escanteios possibilita a realização de uma entrada específica a partir do minuto 39 e do minuto 88 no mercado de *over* como ilustra a figura 1.

Como as operações no mercado de apostas esportivas são binárias, para o modelo apresentar lucratividade, ele precisa ter uma taxa de assertividade acima de um determinado valor, que é dependente da *odd*. Com uma taxa de assertividade de

Figura 1. Tela inicial das apostas.

75% e uma *odd* acima de 1.85 é possível obter um retorno cuja esperança seja de 38.75%.

A. Objetivo

O objetivo desse trabalho é construir um modelo de rede neural que consiga prever com uma precisão de 75% a possibilidade de ocorrência de, pelo menos, um escanteio a partir no minuto 39 ou a partir no minuto 88 de uma partida de futebol e verificar a lucratividade dessa estratégia para vencer as casas de apostas.

II. Trabalhos relacionados

Com o objetivo de encontrar furos nas predições das casas de aposta e sabendo que as *odds* disponibilizadas por elas possuem relação com a probabilidade de cada evento trabalhos que utilizam técnicas de Redes Neurais e Aprendizagem de Máquina podem ser utilizadas para realizar previsões sobre uma partida de futebol com base em estatísticas anteriores dos times.

Em [3], [4], foram utilizadas técnicas de Regressão Linear e Redes Neurais Artificiais para predizer as *odds* e o resultado de uma partida de futebol. Além de Redes Neurais e Regressão, séries temporais também foram utilizadas para predizer resultados de partidas de futebol em [5]. Por sua vez, um modelo de rede **bayesiana** para predição do resultado de uma partida é apresentado em [6].

Uma solução baseada em três modelos de rede - um modelo *gray fuzzy* baseado em redes neurais, uma máquina de aprendizagem *gray extreme* e um comitê montado com os dois modelos anteriores foi proposta por [7]. Eles conseguiram uma acurácia nos resultados de aproximadamente 80%.

Dois modelos híbridos os quais combinam um método de aprendizagem de máquina, *Naive Bayes*, e um método estatístico foram desenvolvidos por [8]. No processo de análise de características eles avaliaram: estatísticas individuais de times e ligas; o impacto de jogar em casa ou fora no resultado final do jogo; o impacto de considerar apenas resultados recentes em relação a considerar resultados globais e, por fim, o impacto de criar modelos separados para cada liga ou um único modelo para todos.

Em [9] foram utilizados dados de 5 ligas européias para predizer o resultado das partidas (Vitória, Empate, Derrota) utilizando um modelo baseado em RNA. Eles usaram o modelo de Dixon e Coles da econometria e uma rede neural implementada seguindo o procedimento do *deep learning*. Verificaram que a rede neural apresentou uma acurácia na predição melhor que o modelo Dixon e Cole. Porém avaliaram, que nos dois casos, há perda a longo prazo no *Return on Investiment*.

Para predição de taxa de vitória na copa de 2006, [10] utilizaram uma MLP com apenas uma camada escondida de 11 nodos, a qual chegou alcançar uma acurácia de aproximadamente 77% em alguns casos. Eles utilizaram 8 características por amostra similares aos utilizados neste trabalho (como gols da casa/fora e possessão de bola), as quais foram normalizadas com relação ao valor que cada time possuía (e.g. sendo 0.5 caso o valor de determinada característica fosse igual para os dois times).

III. Metodologia

A. Conjunto de dados

Os dados foram extraídos do site[2]. Dessa forma, para construção do conjunto de dados, foi implementado um *crawler* - responsável pela extração dos links de cada jogo - e um *scraper*, o qual extrai as características de cada amostra, aproveitando os padrões que existem no código HTML da página de cada jogo.

No total foram coletadas 5796 amostras, cada amostra com 21 características. Cada amostra corresponde a um tempo do jogo, primeiro ou segundo tempo. Para o primeiro tempo foi observado as características até o minuto 39 e para o segundo tempo foram observadas as características até o minuto 88. O conjunto de dados geral foi separado em dois menores contendo 2898 amostras cada, com a finalidade de analisar os dois tempos de forma independente.

B. Encoder

Foi desenvolvido um *autoencoder* para reduzir a dimensionalidade dos dados. Para isso, como não se tratavam de muitas características, foi construída uma arquitetura de 3 camadas e testadas combinações de 11, 12 e 14 nodos para a primeira camada e uma combinação de 8 e 6 nodos para a segunda camada, sendo a terceira camada uma repetição da primeira.

[2]https://www.scorebing.com/

Cada combinação foi treinada durante 100 épocas utilizando o otimizador Adam e uma taxa de aprendizagem de 10^{-3}, sendo a função de perda o método dos mínimos quadrados. Após os testes, a melhor topologia para o *encoder* foi de 12 nodos para a primeira camada e 8 nodos para a segunda camada.

C. Multilayer Perceptron - MLP

O modelo MLP foi desenvolvido através de uma busca randomizada em grade para um modelo de três camadas. As funções de ativação testadas em cada caso foram: identidade; logística; tangente hiperbólica e ReLu. Os otimizadores testados foram: L-BFGS (*Limited-memory Broyden–Fletcher–Goldfarb–Shanno*); SGD (*Stochastic Gradient Descent*) e Adam. Enquanto que a taxa de aprendizagem foi variada entre 1 e 10^{-5}. O objetivo dessa busca foi obter o melhor resultado com base na combinação dos hiperparâmetros considerando as métricas da acurácia, precisão, *recall* e F1-score.

D. PCA

A técnica do PCA *(Principal Component Analysis)* foi aplicada sobre o conjunto de dados com o objetivo de reduzir a dimensionalidade do problema de modo que o mínimo de informação possível seja perdida. Dessa forma, através de método iterativo, foi definido que os componentes principais resultantes da transformação deveriam explicar pelo menos 95% da variância do problema, pois com esse valor seria possível diminuir a quantidade de atributos para apenas onze colunas sem que houvesse grande prejuízo para o modelo.

Tabela I
COMPONENTES PRINCIPAIS

Componente	Variância Explicada (%)
PC1	33.7
PC2	12.25
PC3	10.31
PC4	9.11
PC5	8.17
PC6	5.73
PC7	5.40
PC8	4.18
PC9	3.23
PC10	2.53
PC11	2.15

E. SVM

O classificador SVM *(Support Vector Machine)* é um classificador binário que procura um hiperplano ótimo como uma função de decisão em um espaço de dimensões maiores. Com ele foi realizada uma busca randomizada dos hiperparâmetros para cada modelo de SVM treinado de modo a obter as combinações que trouxessem os melhores resultados possíveis para a situação analisada. Para isso, foram variadas:

- Função Kernel: função utilizada pelo algoritmo para alterar o espaço vetorial do problema. Teve como opções: RBF, Polinomial, Sigmoidal e Linear;
- C: parâmetro de regularização do SVM. É inversamente proporcional à força de regularização. Foi testada para valores entre 10 e 0.1;
- Gamma: é inversamente proporcional ao tamanho da esfera de influência de uma instância de treino. Variou entre 1 e 0.001.

F. Melhores Hiperparâmetros

As combinações ótimas para as arquiteturas que utilizaram o SVM foram variadas, sem que houvesse, por exemplo, repetição de função Kernel entre os modelos de uma mesma arquitetura. Em relação ao parâmetro C, percebe-se que as melhores combinações para o PCA-SVM apresentaram valores altos, indicando uma baixa força de regularização em comparação com as utilizadas pelas outras arquiteturas. Esses parâmetros podem ser vistos nas tabelas II, III e IV.

Tabela II
MELHORES HIPERPARÂMETROS DO SVM

Modelo	Função Kernel	Gamma	C
Jogo Completo	RBF	0.01	0.4
1° Tempo	Polinomial	1	0.1
2° Tempo	Sigmoidal	1	0.1

Tabela III
MELHORES HIPERPARÂMETROS DO PCA-SVM

Modelo	Função Kernel	Gamma	C
Jogo Completo	Sigmoidal	0.01	10
1° Tempo	RBF	0.01	10
2° Tempo	Polinomial	1	1

Quando se analisa os modelos treinados em cima da base completa, percebe-se que esses tiveram valores de Gamma baixos e que, quando combinados com técnicas de Encoder ou PCA, tiveram melhor desempenho ao utilizar uma função de kernel Sigmoidal.

Tabela IV
MELHORES HIPERPARÂMETROS DO ENCODER-SVM

Modelo	Função Kernel	Gamma	C
Jogo Completo	Sigmoidal	0.1	0.1
1° Tempo	Polinomial	1	0.6
2° Tempo	RBF	1	0.1

Já no MLP, observa-se que a estrutura ótima encontrada para o modelo da base de dados completa foi maior que as estruturas utilizadas para prever apenas o primeiro ou segundo tempo, como mostrado na Tabela V. Esse resultado é coerente, visto que uma maior quantidade de neurônios na rede aumenta a complexidade do algoritmo, algo necessário nesse caso, dado que o conjunto Jogo Completo apresenta mais desafios que os parciais.

Embora tenham camadas de mesmo tamanho, os modelos para primeiro e segundo tempo apresentaram grandes

diferenças na sua construção. A taxa de aprendizagem ótima para o modelo do primeiro tempo foi dez mil vezes menor que a obtida para o segundo tempo, indicando uma correção muito mais lenta dos seus pesos durante o processo de treinamento. Além disso, tanto os otimizadores e a funções de ativação foram discrepantes.

Tabela V
MELHORES HIPERPARÂMETROS DO MLP

Modelo	Estrutura	Otimizador	Função de Ativa.	Taxa de Aprendi.
Jogo Comp.	(18,14)	Adam	Identidade	0.0001
1° Tempo	(12,10)	Adam	Tanh	0.0001
2° Tempo	(12,10)	SGD	Identidade	1

IV. RESULTADOS

Nesta seção, são apresentadas as tabelas VI, VII, VIII, IX com os resultados dos algoritmos de redes neurais como **SVM, PCA-SVM, Encoder-SVM e MLP**. Neste sentido, também são reportados os resultados referentes às melhores soluções para o primeiro tempo, segundo tempo e jogo completo e a descrição de abreviação das métricas[3].

Tabela VI
SVM

Métrica	Primeiro Tempo	Segundo tempo	Jogo completo
Acurácia	**0.61**	0.60	0.59
PCP	0.64	0.74	**0.79**
PCN	0.57	**0.58**	0.54
RCP	**0.61**	0.14	0.13
RCN	0.59	0.96	**0.97**

Tabela VII
PCA-SVM

Métrica	Primeiro Tempo	Segundo tempo	Jogo completo
Acurácia	0.60	**0.66**	0.63
PCP	0.63	**0.73**	0.62
PCN	0.57	0.59	**0.63**
RCP	**0.65**	0.15	0.58
RCN	0.57	**0.96**	0.67

A. Estratégia

O lucro pode ser calculado pela equação 1

$$L = [(P * (M - 1) - (1 - P)] * 100 \quad (1)$$

,

[3]acurácia = quantidade de acertos. PCP = precisão classe positiva. PCN = precisão classe negativa. RCP = recall classe positiva. RCN = recall classe negativa.

Tabela VIII
ENCODER-SVM

Métrica	Primeiro Tempo	Segundo tempo	Jogo completo
Acurácia	**0.61**	0.57	0.58
PCP	0.65	**0.70**	0.67
PCN	**0.59**	0.57	0.57
RCP	**0.42**	0.10	0.15
RCN	0.79	**0.96**	0.94

Tabela IX
MLP

Métrica	Primeiro Tempo	Segundo tempo	Jogo completo
Acurácia	0.61	**0.63**	0.62
PCP	0.63	**0.65**	0.61
PCN	0.59	0.62	**0.63**
RCP	**0.63**	0.38	0.61
RCN	0.59	**0.83**	0.64

na qual L é o lucro, P é a precisão e M é o multiplicador. O multiplicador representa o retorno que a casa de apostas fornece ao se ganhar uma aposta, sendo subtraído por 1 para desconsiderar o dinheiro já investido para o cálculo do lucro. Sendo assim, o primeiro termo indica a porcentagem de lucro desconsiderando as perdas, enquanto que o segundo termo leva as perdas em consideração, dado que para cada erro, 100% do valor apostado é perdido.

Tendo isso em vista, pode-se traçar diferentes estratégias visando somente um maior ganho ou um ganho menor em troca de uma frequência alta de apostas, o que aqui é chamado de estratégia conservadora e estratégia agressiva respectivamente.

- **Estratégia agressiva para classe positiva:** Usar modelo MLP no primeiro tempo para classe positiva, dado que possui precisão similar ao *Encoder*-SVM aliado a um *recall* maior. Além disso, usar modelo SVM no segundo tempo para classe positiva, dadoq ue possui precisão muito melhor que os demais modelos.
- **Estratégia agressiva para classe negativa:** Usar modelo *Encoder*-SVM no primeiro tempo para classe negativa porque possui a maior precisão e *recall* para essa classe. Já para o segundo tempo, usar modelo MLP para classe negativa, visto que possui precisão similar e *recall* maior.
- **Estratégia conservadora:** Apostar apenas na ocorrência de escanteio (classe positiva), visto que o SVM possui uma precisão na classe positiva de 79%, embora deixe passar muitos falsos positivos (*recall* de 13%), o que faz com que a frequência de ganho nas apostas seja muito mais baixa, podendo não ser interessante a depender da necessidade de se ter uma frequência maior de ganhos (e.g. quando a frequência de apostas realizadas é baixa).

V. CONSIDERAÇÕES FINAIS

Neste trabalho, foram propostos estudos de técnicas de redes neurais para o problema de predição de ocorrência de escanteios no final de um dos tempos de uma partida

de futebol. Os modelos apresentados demonstram uma boa acurácia, visto que uma acurácia maior que 60%, considerando um Multiplicador de 1.80, resultaria em um ganho maior que 8% por. aposta.

Em termos de precisão, o melhor resultado foi o classificador SVM considerando os dados do primeiro e segundo tempo, para escanteios no final do segundo tempo, usando a estratégia conservadora.

VI. Trabalhos futuros

Durante a elaboração deste trabalho surgiram várias possibilidades de pesquisa que podem agregar uma maior qualidade ao trabalho proposto, dessa forma, buscamos elencar alguns pontos importantes a serem desenvolvidos nos trabalhos futuros.

- Testar outros modelos de RN (*deep learning*, outras combinações através de ensemble, treinamento semi-supervisionado).
- Fazer testes de hipóteses nos resultados (*Friedman Test*);
- Filtrar os dados por ligas mais relevantes;
- Aumentar a instância de dados;
- Refinar os dados: inserir mais atributos, considerar ordem cronológica dos eventos, gerar atributos artificialmente;

Referências

[1] S. D. Levitt, "How do markets function? an empirical analysis of gambling on the national football league," *NATIONAL BUREAU OF ECONOMIC RESEARCH*, 2003.

[2] L. Kaunitz, S. Zhong, and J. Kreiner, "Beating the bookies with their own numbers - and how the online sports betting market is rigged," 10 2017.

[3] S. Kumar, "Artificial neural network for betting rate in football," Master's thesis, Dublin, National College of Ireland, 2020.

[4] J. Goddard, "Regression models for forecasting goals and match results in association football," *International Journal of forecasting*, vol. 21, no. 2, pp. 331–340, 2005.

[5] J. Maria and T. Moniz, "Forecasting Football Outcomes to Invest in Betting Markets Forecasting Football Outcomes to Invest in Betting Markets," 2018.

[6] A. Constantinou, "Dolores: a model that predicts football match outcomes from all over the world," *Machine Learning*, vol. 108, 01 2019.

[7] S. Guan and X. Wang, "Optimization analysis of football match prediction model based on neural network," *Neural Computing and Applications*, vol. 34, no. 4, pp. 2525–2541, 2022.

[8] H. Elmiligi and S. Saad, "Predicting the outcome of soccer matches using machine learning and statistical analysis," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1–8, IEEE, 2022.

[9] J. Sibony, A. Tlemsani, Y. Hamchi, M. Gjergji, E. Shurmanov, and M. Frunza, "Neural networks predictive modeling for football betting," *Available at SSRN 3655700*, 2020.

[10] K.-Y. Huang and W.-L. Chang, "A neural network method for prediction of 2006 world cup football game," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2010.