Predição de resultados de futebol com Machine Learning

Victor Domiciano Moraes[1]

Resumo

Este trabalho busca conceber um modelo matemático capaz de prever resultados de partidas de futebol a partir de um treinamento prévio com partidas já realizadas e contidas em uma base de dados. Além, este trabalho sugere a implementação de algoritmos de aprendizado de máquina para afinar o modelo matemático e alcançar melhor taxa de acerto nas previsões.

Palavras-chave: Aprendizado de Máquina. Artigo científico. Futebol. Predição de resultados.

1 – Introdução

O futebol é indubitávelmente uma paixão nacional. Acompanhada dessa paixão vem a alegria de ver o time do coração vencer jogos e, consequentemente campeonatos. Porém, essas partidas nem sempre são vencidas facilmente. Em tais partidas, há uma dosagem extra de emoção e ansiedade, até mesmo desespero, que acompanha o torcedor até o último minuto do segundo tempo.

Este tipo de emoção nem sempre é bem-vinda. O homem sempre quis prever o futuro das mais diferentes maneiras e nas mais diferentes áreas. Este trabalho tenta unir estas características e tem como objetivo fina prever o resultado de uma partida de futebol cujas equipes e data são previamente informados.

Esta é uma aplicação simples e que pode ser usada para entender, tanto na teoria quanto na prática os usos de algoritmos de aprendizado supervisionado de máquina.

2 – Desenvolvimento

Antes de se desenvolver o código da solução para o problema proposto, é preciso entender os desafios que acompanham este problema. São eles: **(1)** Encontrar um volume de dados suficiente para o aprendizado **(2)** Como transformar e modelar esses dados de forma a se transformar numa entrada plausível para o modelo matemático? **(3)** Qual o melhor algoritmo de aprendizado para a predição?

2.1 – Primeiro desafio

Para o primeiro desafio, é preciso frisar que todos os tipos de problemas que são resolvidos com aprendizado de máquina supervisionado devem ter um conjunto de dados de treinamento de tamanho satisfatório com o objetivo de cobrir todas as possibilidades e saídas, assim permitindo que o algoritmo 'conheça' todo o escopo do problema e possa gerar saídas com grande nível de precisão.

No tocante à predição de resultados de partidas de futebol isso não é diferente. É necessário uma grande base de dados que contenha o histórico de milhares de partidas de futebol de várias temporadas. Por isso, foram estudados diversos conjuntos de dados e foi selecionado o conjunto de dados chamado 'European Soccer Database', que contém uma amostra de 25 mil partidas da principal liga de onze países da Europa entre os anos de 2008 e 2016.

Mais detalhadamente falando, essa base de dados possui informações sobre jogos de futebol, jogadores, equipes, países, campeonatos e habilidades técnicas de jogadores. Trata-se

[5]Autor correspondente: vdmoraes94@gmail.com
[1]Centro Federal de Educação Tecnológica de Minas Gerais

de um conjunto de dados bastante completo e apto a ser usado para a modelagem do problema proposto.

2.2 – Segundo desafio

Com a base de dados em mãos, foi necessário um estudo sobre como preparar os dados de entrada das amostras para servirem de dados de entrada para o algoritmo de aprendizado. A primeira decisão tomada foi a de não usar 100% dessas amostras, uma vez que em cada um dos onze países contidos na base de dados as influênc

2.3 – Próximos passos

ias no resultado de um jogo e até mesmo a maneira de praticar o esporte é diferente. Ou seja, há variáveis que influenciam mais do que outras no resultado de uma partida em um determinado campeonato, e essas variáveis podem influenciar menos em detrimento de outras em outra liga de outro país.

Por isso, decidiu-se por usar as amostras apenas de uma liga em particular, a 'Premier League', a primeira divisão do campeonato inglês de futebol. Essa liga foi escolhida por além de ser uma das mais famosas do mundo, é a liga que tem o nível de detalhe mais rico no que diz respeito à qualidade dos dados extraídos. Ademais, a 'Premier League' é a liga que mais movimenta dinheiro no mundo, dando um total de 400 milhões de libras esterlinas, o equivalente a 1,7 bilhões de reais em premiação para as equipes, além de ser a liga que mais movimenta dinheiro em direitos de transmissão e transferência de jogadores.

Finalmente, foram escolhidas para treinamento um total de 3.040 amostras, ou seja, 3.040 jogos de futebol do campeonato inglês entre os anos de 2008 e 2016. Após, foi preciso uma nova análise desses dados para decidir como o modelo matemático de previsão do resultado das partidas seria construído.

Para a previsão em si, um sistema de influências foi criado. Para efeitos de simplificação, foram escolhidas apenas quatro fatores que podem influenciar o resultado de um jogo:

(1) Mando de campo

(2) Pontuação da equipe no campeonato desde o início até a rodada a qual a partida será realizada

(3) Pontuação da equipe no campeonato nos últimos sete jogos

(4) Média quantificada das habilidades do time titular de cada equipe

É importante mencionar que todas os fatores têm a mesma importância, ou seja, implicitamente falando todos os fatores tem peso 1.

Decidiu-se em cada fator comparar as equipes entre si e, após a comparação, a cada fator é atribuido um valor, avaliado em:

(a) 1, em caso de vitória do mandante no fator

(b) -1, em caso de vitória do visitante no fator

(c) 0, em caso de empate no fator

Assim, o resultado da partida é previsto após a soma dos valores de cada fator, semelhante às mesmas regras listadas acima, caso a soma dos valores seja

(a) maior ou igual a 1, vitória do time mandante

(b) menor ou igual a -1, vitória do time visitante

(c) 0, empate

2.4 – Construção do algoritmo

O objetivo final é que o algoritmo receba uma partida do campeonato inglês como parâmetro de entrada, este parâmetro é subdividido em três parâmetros: 'nome do time da casa', 'nome do time visitante', 'ano em que a partida é realizada'.

Com isso, o algoritmo busca a rodada da liga pertence esta partida, além de calcular dados pertinentes a cada uma das equipes, como pontuação, jogadores que iniciarão aquela partida como titulares, número de gols marcados, número de gols sofridos e a média quantificada da habilidade técnica desses jogadores titulares.

Após a extração desses dados, estes são colocados em análise com o objetivo de se calcular os fatores listados na subseção acima e, assim, gerar a previsão do resultado do jogo.

Tecnicamente falando, optou-se por usar a linguagem de programação Python com a biblioteca Pandas para o tratamento dos dados e a comunicação com a base de dados, que foi criada usando a linguagem de banco de dados SQL e permite visualização pelo script Python através da biblioteca SQLite.

O script criado, chamado de Analyst, recebe por parâmetro em linha de comando os dados descritos acima, sua execução é feita por:

```
python analyst.py <mandante> <visitante> <ano>
```

Quando executado, o script se conecta com a base de dados pela biblioteca SQLite, extrai os dados da base de dados tomando como referência as informações obtidas por linha de comando usando o Pandas, os converte em tipos de variável primários, faz os cálculos pertinentes explicados anteriormente e retorna para o usuário uma mensagem indicando qual será o resultado da partida (em inglês).

2.5 – Treinamento e precisão da predição

Após prever o resultado da partida, o algoritmo reexecuta todos os passos N vezes, sendo N o tamanho da amostra, ou seja, agora o modelo é executado para todas as partidas da amostra. Porém, dessa vez o resultado da previsão é comparado com o resultado real da partida (1 para vitória do mandante, -1 para vitória do visitante e 0 para o empate) e, em caso de igualdade, é contabilizado um acerto para o algoritmo, caso contrário, é contabilizado um erro para o algoritmo. A precisão é calculada como a soma dos acertos dividida pelo número de amostras.

2.6 – Resultados e terceiro desafio

Para o caso explanado acima, onde todos os fatores/variáveis têm a mesma influência sob o resultado final, sem quaisquer tipo de aprendizado supervisionado, foi obtida uma precisão na ordem de 42,23% sob as 3.040 amostras comparadas. Portanto, é fortemente recomendado o uso de um algoritmo de aprendizado de máquina supervisionado onde o desafio final torna-se ajustar as influências (ou pesos) de cada variável para que haja um aumento no número de acertos da solução proposta e, consequentemente, um aumento na precisão. Afinal de contas, trata-se de um espelho da realidade, pois de fato alguns fatores são mais preponderantes no resultado final de uma partida do que outros, por isso é fundamental o uso de aprendizado de máquina, ou *Machine Learning*.

Na visão de aprendizado de máquina, conclui-se que prever os resultados de uma partida de futebol trata-se de um problema de classificação, onde as partidas devem ser classificadas em vitória do mandante, vitória do visitante ou empate. Por isso, três algoritmos são propostos para treinar os pesos:

(1) Redes Neurais Recorrentes **(2)** Supporting Vector Machines (SVM) **(3)** Regressão Logística

A escolha do melhor algoritmo a ser usado depende diretamente de como os dados são modelados. Algumas aplicações interessantes podem ser utilizadas a partir desse modelo. Como por exemplo a otimização de esforços em determinadas áreas por parte de um time de futebol. Por exemplo, se em um país A o modelo após treinamento detectou uma influência maior do mando de campo nas vitórias

do time da casa, é possível concentrar recursos na melhoria do estádio desse time. Ou em um país B onde a eficácia defensiva é fundamental nos resultados positivos de uma equipe, é possível criar uma formação tática ou estratégia que tenha ênfase na defesa.

2.7 – Próximos passos

Além da efetiva implementação de um algoritmo de treinamento no modelo proposto, várias outras melhorias podem ser efetuadas visando um aumento na precisão da solução, como por exemplo:

(1) Uso de mais fatores, como entrosamento da equipe **(2)** Melhor modelagem dos atributos dos jogadores **(3)** Análise de sensibilidade usando sistemas Fuzzy

Tudo isso, combinado, traria importante crescimento no desenvolvimento do modelo de negócio e, após, em resultados positivos.

2.8 – Conclusão

Prever resultados de futebol é só uma simples aplicação de métodos matemáticos e aprendizado de máquina e que torna fácil o entendimento da teoria que cerca esses dois assuntos. É impossível alcançar a perfeição nas predições, uma vez que o fator humano envolvido nos resultados das partidas é ainda impossível de se modelar. Além disso, um algoritmo hipotético que acertasse 100% das previsões tiraria aquela que é a motivação em se acompanhar esportes, que é a emoção envolvida acerca da confiança em um resultado positivo da sua equipe que o torcedor.

Sendo o fator humano tão importante até mesmo para a rentabilidade do futebol no meio profissional, existem certas fronteiras as quais, por uma questão ética, a Inteligência Artificial deve ser impedida de alcançar a perfeição, e prever resultados do esporte mais amado do planeta com 100% de acertos é certamente uma dessas fronteiras.

3 – Conclusão

Edite esta seção para colocar a conclusão de seu trabalho de pesquisa.

Procure fazer uma análise crítica de seu trabalho, destacando os principais resultados e as contribuições deste trabalho para a área de pesquisa.

Também deve indicar, se possível e/ou conveniente, como este trabalho pode ser estendido ou aprimorado.

Abstract

This work explains and builds a mathematical model able to predict final results of football (soccer) matches from a previous training in past games stored in a dataset. Furthermore, this work suggests an implementation of a machine learning algorithm which adjusts the influences involved in each deciding factor in order to improve prediction accuracy.