Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Projeto Final de Aprendizado Descritivo

Mineração de Dados de Eventos em Futebol

Alunos: Luís Felipe Ramos Ferreira, Igor Lacerda Faria da Silva, Matheus Tiago Pimenta de Souza

Professor: Renato Vimieiro

Belo Horizonte - Minas Gerais
2024

fala zezé, bom dia cara

Sumário

Mineração de Dados de Evento em Futebol

Luís Felipe Ramos Ferreira
Igor Lacerda iFaria da Silva
Matheus Tiago Pimenta de Souza

29 de julho de 2024

1 Introdução

O uso de ciência de dados e estatística para analisar esportes é algo que vem crescendo cada vez mais nos últimos anos. Em particular, o futebol tem sido um desses esportes [11]. Diversas empresas que atuam na área surgem a cada dia, e os times de futebol, no Brasil e no resto do mundo, estão investindo em seus departamentos de dados e estatística.

Nesse sentido, este trabalho propõe estudar e compreender melhor como funciona o uso de análises estatísticas no futebol, dado o interesse do público geral pelo esporte, e, para isso, propomos a aplicação de algoritmos de mineração de dados em dados futebolísticos, sendo eles dados de súmula e dados de eventos das partidas, para compreender como as informações acerca do jogo estão contidas dentro dos dados coletados e como isso pode ser utilizado a favor das equipes.

Os dados de eventos, especialmente, costumam ser mais fáceis de lidar e mais fáceis de acessar do que dados de *tracking*, enquanto trazem muito mais informações do que dados de súmula. Existem, atualmente, algumas bases gratuitas de dados de evento de partidas, disponibilizadas por diferentes empresas como *Wyscout* e *StasBomb*.

1.1 Base de dados

A principal base disponibilizada pela empresa *Wyscout* foi usada como base de dados. Ela contém dados de evento das 5 grandes ligas europeias na temporada 17/18.

Cada empresa fornecedora de dados possui seu próprio formato de representação dos dados de evento. De modo a facilitar a mesclagem entre as bases de dados utilizadas, iremos converter os dados coletados para uma representação geral proposta por pesquisadores denominada SPADL. A SPADL é uma boa escolha por ser uma representação concisa e fácil de utilizar. Ela é uma representação tabular de cada evento da partida, onde cada linha possui 12 colunas. A tabela abaixo ilustra o esquema de representação de um evento segundo o formato SPADL.

Atributo	**Descrição**
game_id	O ID do jogo no qual a ação foi realizada
period_id	O ID do período do jogo no qual a ação foi realizada
seconds	O tempo de início da ação
player	O jogador que realizou a ação
team	O time do jogador
start_x	A localização x onde a ação começou
start_y	A localização y onde a ação começou
end_x	A localização x onde a ação terminou
end_y	A localização y onde a ação terminou
action_type	O tipo de ação (por exemplo, passe, chute, drible)
result	O resultado da ação (por exemplo, sucesso ou falha)
bodypart	A parte do corpo do jogador usada para a ação

Tabela 1: Descrição dos dados no formato SPADL

2 Implementação

A linguagem escolhida para o desenvolvimento do trabalho foi `Python` (versão 3.10.12), devida a seu vasto ecossistema para ciência de dados e mineração de dados.

A manipulação dos dados foi feita com o uso de bibliotecas de análise numérica como `NumPy` e manipulação de *dataframes* como `Polars` e `Pandas`, uma vez que se tratam de ferramentas extremamente completas que facilitaram o desenvolvimento do projeto como um todo.

Para aplicar os algoritmos de descobertas de subgrupos, foram utilizados os pacotes `pysubgroup` e `subgroups`, que fornecem uma aglomeração de algoritmos do estado da arte de descoberta de subgrupos em um formato simples e leve para serem utilizados.

falar dos pacotes de mineração de seq aqui

Para organizar o ambiente de desenvolvimento, que englobava vários pacotes diferentes, foi utilizado o gerenciador de pacotes `Anaconda`, o que facilitou o trabalho com os pacotes de ciência de dados citados. O projeto final foi salvo em um `repositório` no GitHub para fácil versionamento e organização de código. As instruções de como utilizar o que foi implementado estão descritas no *README* do repositório.

3 Referencial Teórico

No decorrer do trabalho, duas importantes métricas de análise ofensiva no futebol serão utilizadas. Esta seção aglomera os conhecimentos necessários sobre elas para a compreensão do projeto e o como elas foram utilizadas.

3.1 Gols Esperados (xG)

Intuitivamente, existe a noção de que, quanto mais próximo um jogador está do gol, mais chance ele tem de conseguir marcar um ponto para sua equipe. Uma noção similar existe para o ângulo entre o gol e jogador: é mais difícil um atacante acertar se ele está em uma das laterais. Na área de *analytics* de futebol, essa noção é formalizada através da métrica de *expected goals*, ou xG. A ideia central é construir um modelo do que um jogador médio faria em dado estado de jogo, que, além de incluir os fatores mencionados anteriormente, pode ser mais (ou menos) extensivo.

A métrica de Gols Esperados captura um estado de jogo e retorna uma probabilidade estimada de um jogador marcar. Ela pode ser vista como um *framework*, cujos detalhes de implementação são decididos pelo usuário. Outros fatores considerados importantes são: parte do corpo associada à ação (pé dominante ou não; de cabeça), origem da assistência (cruzamento, passe) e a posição dos defensores. Para avaliar este último quesito, é necessário ter dados de *tracking*, o que limita consideravelmente sua adoção. Além disso, mesmo questões como a liga em análise podem ser relevantes.

O uso do xG é mais amplo do que se pode imaginar a princípio. Em um primeiro momento, partidas podem ser analisadas: é possível avaliar se um time fez mais "pressão" pela soma do seu xG na partida. No entanto, como com qualquer outro indicador, o xG não define absolutamente o resultado de um jogo, podendo um time aproveitar muito bem situações menos favoráveis. Outra aplicação é avaliação de jogadores: se um jogador consistentemente faz gols em situações desfavoráveis, isso é um grande indício de que ele é jogador de destaque, ou age muito bem sob pressão. O xG também pode ser usado taticamente, para proporcionar melhorias no posicionamento dos jogadores da *defesa* e para embasar estratégias.

É comum usar modelos de regressão logística para aplicar o xG, em que se considera uma combinação dos fatores mencionados anteriormente. Essa combinação não é necessariamente linear: considerar o quadrado da distância pode ser valioso, por exemplo. Na prática, no entanto, qualquer modelo de classificação binária pode ser usado. Um exemplo muito prevalente é o xG do *Statsbomb*, que usa um *XGBoost* e considera a quantidade e a posição dos jogadores entre o gol e o jogador que faz o lance.

3.2 Valuing Actions by Estimating Probabilities (VAEP)

O xG deixa a desejar em alguns aspectos. Principalmente pelo fato de focar exclusivamente nas chances de se fazer gols, sem considerar outros aspectos importantes do jogo. Por exemplo, existem passes que são cruciais para uma equipe se aproximar do gol. É justamente para suprir a necessidade de avaliar

as ações dos jogadores de forma mais ampla que foi desenvolvido o VAEP [5]. O VAEP avalia cada ação do jogo em termos do quanto ela aumenta (ou diminui) a chance de um time fazer (ou sofrer) um gol.

Com base no estado atual do jogo (placar, tempo restante, ações anteriores, posição da bola, etc), o VAEP estima a chance de um time fazer ou sofrer um gol, antes e após cada ação. Desse modo, é possível estimar a utilidade de uma ação com base na diferença entre as mudanças de se fazer e de se sofrer um gol. Formalizando, temos que, para um dado time t, uma ação a_i e um estado de jogo S_i:

$$\Delta P_{score}(a_i, t) = P^k_{score}(S_i, t) - P^k_{score}(S_{i-1}, t) \tag{1}$$

$$\Delta P_{concede}(a_i, t) = P^k_{concede}(S_i, t) - P^k_{concede}(S_{i-1}, t) \tag{2}$$

$$V_{\text{VAEP}}(a_i) = \Delta P_{score}(a_i, t) - \Delta P_{concede}(a_i, t) \tag{3}$$

Nas equações, k é um parâmetro que indica quantas jogadas futuras estão sendo consideradas para se avaliar a jogada atual. Assim, quando uma jogada aumenta muito a chance de um gol acontecer, seu VAEP é um número positivo e, caso a jogada ofereça mais risco do que recompensa, o VAEP será um valor negativo.

Então, para se calcular o VAEP, é necessário estimar as probabilidades de marcar ou sofrer um gol. Essa tarefa pode ser resolvida com Aprendizado de Máquina, com a criação de dois modelos: um para estimar a probabilidade da equipe com a posse da bola fazer um gol até as k ações após o estado atual S_i e outro para estimar a probabilidade de a equipe sofrer o gol no mesmo período. Um valor comummente utilizado para k é 3.

Tal como o xG, o VAEP pode ser utilizado para se avaliar a performance dos jogadores, ao se agregar as ações. Inclusive, os autores do *paper* usaram disso para validar que o VAEP é uma métrica consistente. Considerando os jogadores que participaram de jogos das ligas europeias nas temporadas de 17/18 (*coincidentemente* a mesma base de dados que foi usada nesse trabalho) por pelo menos 900 minutos, foram construídos gráficos mostrando o número de ações por 90 minutos pelo valor médio das ações. Liderando o *ranking*, nomes conhecidos como Messi, Bale e Cristiano Ronaldo aparecem, o que indica que o VAEP é condizente com outras avaliações de desempenho.

4 Análise Exploratória

A base de dados de eventos possui muitas informações interessantes que podem ser exploradas antes mesmo da aplicação de algoritmo de mineração de dados. Nesta seção, discutimos alguns *insights* interessantes observados na base das 5 grandes ligas europeias da temporada 17/18.

4.1 Distribuição de ações

A base de dados possui uma distribuição não uniforme de ações. Como pode-se analisar no histograma abaixo, os eventos de passes são extremamente mais frequentes do que qualquer outro. Isso está dentro do esperado, dado que o passe é o principal fundamento do futebol, mas deve ser levado em consideração quando modelos de mineração forem aplicados à base de dados.

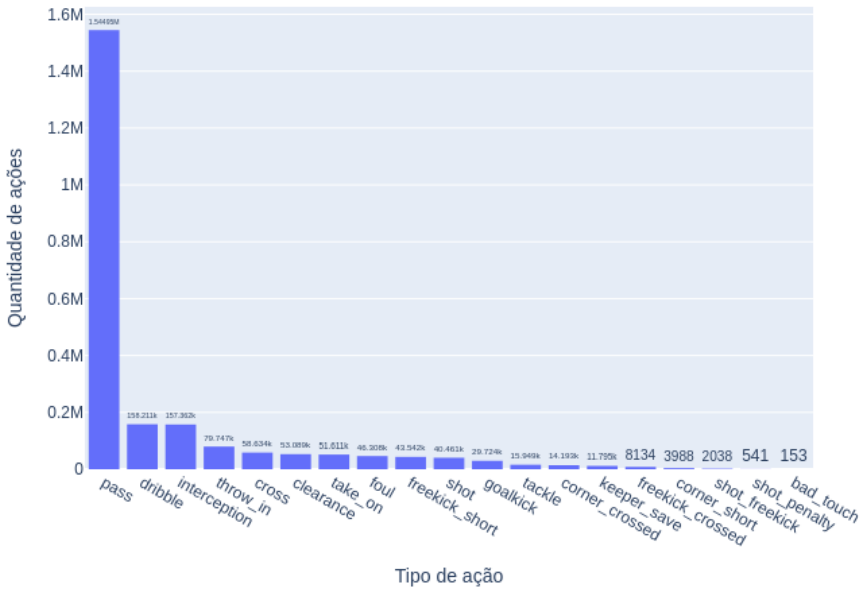

Figura 1: Distribuições dos tipos de ação

4.2 Distribuição da posição de chutes convertidos em gol

O mapa de calor abaixo permite que sejam analisadas a distribuição das posições dos chutes que se converteram em gol na base de dados.

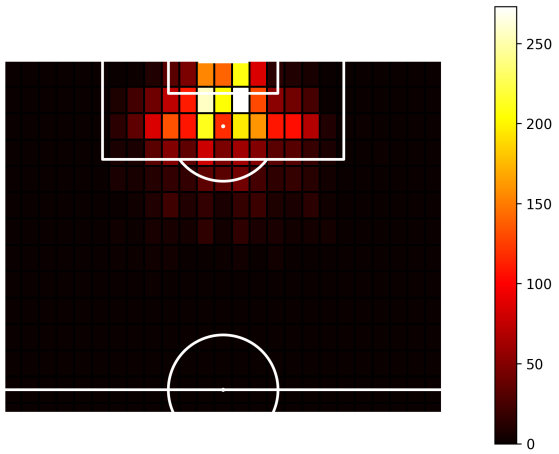

Figura 2: Distribuições dos tipos de ação

4.3 Distância média entre passes

O passe é um dos, se não o fundamento mais importante em um esporte coletivo como o futebol. Uma boa execução desse fundamento por parte dos jogadores de uma equipe indica um bom controle da posse de bola que está diretamente conectado com bons resultados [4]. Nesse sentido é interessante fazer um análise das distâncias entre os passes executados por cada equipe.

Em particular, foi coletada a distância entre todos os passes bem sucedidos de cada equipe e, posteriormente, coletadas estatísticas sobre essas distâncias. A distância média entre passes se destacou, uma vez que ela indica como funciona a dinâmica de troca de passes de uma equipe. Os resultados observados confirmaram suspeitas prévias acerca do assunto. Como pode-se notar nas tabelas abaixo,

para cada uma das grandes ligas, as principais equipes, isto é, as equipes com maior grandeza histórica e maior poderio financeiro figuram como aquelas que possuem a menor distância média entre passes bem sucedidos. Na Inglaterra, por exemplo, o Manchester City, equipe comandada pelo espanhol Pep Guardiola, foi a equipe com a menor média de distância entre passes. O estilo de jogo de Guardiola é muito focado no controle da posse de bola e na movimentação dos jogadores para receptar um passe [12], então o resultado está dentro do esperado.

É interessantes notar que os clubes com menor média de distância entre os passes da temporada 17/18 em cada liga mantiveram posições altas em seus respectivos campeonatos. O Manchester City e o Paris Saint Germain se sagraram campeões, enquanto Napoli e Atletico de Madrid ficaram com a segunda colocação. Na Alemanha, o RB Leipzig ficou com a sexta colocação. Pode-se notar então que existe uma correlação entre a o desempenho de um time no campeonato com a distância média entre os passes da equipe.

Equipe	**Distância média entre passes (m)**
Manchester City FC	17.208
Arsenal FC	17.729
Manchester United FC	17.823
AFC Bournemouth	18.417
Tottenham Hotspur FC	18.490
Crystal Palace FC	18.555
Chelsea FC	18.682
Southampton FC	18.801
Liverpool FC	18.808
West Ham United FC	18.832
Watford FC	18.967
Newcastle United FC	18.972
Swansea City AFC	19.077
Leicester City FC	19.189
Huddersfield Town FC	19.215
Stoke City FC	19.690
West Bromwich Albion FC	19.712
Everton FC	19.785
Brighton & Hove Albion FC	19.838
Burnley FC	20.634

Tabela 2: Premier League

5 Extração de *Features*

quais features usamos e qual target (paper original usou binario e xg, propomos usar vaep)

6 Descoberta de Subgrupos

sd nos dados. usar pysubgroupigual no artigo domiguel

7 Mineração de Sequências

Também foram aplicadas técnicas de mineração de sequências para encontrar padrões de tipos de ações realizadas antes de um chute ao gol. Para realizar a mineração de sequências foram usadas algumas técnicas.

Uma delas foi inspirada em [3]: o *Safe Pattern Pruning* (SPP). Usado originalmente no rugby, com resultados promissores, avaliou-se a viabilidade de aplicar o mesmo algoritmo no contexto do futebol. Para isso, os dados foram adaptados ao formato de entrada do programa publicado pelos autores

do artigo: cada linha possui um número para indicar a prosperidade da sequência de ações (1 para resultado promissor e -1 para caso contrário) e uma sequência de ações propriamente dita.

As sequências de ações são definidas com base na troca de posse de bola, até ocorrer algum chute ao gol (*shot*). As ações foram definidas como números, conforme a tabela 3.

ID	Ação
0	pass
1	interception
2	dribble
3	take on
4	tackle
5	foul
6	freekick short
7	cross
8	**shot**
9	clearance
10	throw in
11	goalkick
12	corner short
13	corner crossed
14	keeper save
15	freekick crossed
16	shot freekick
17	bad touch
18	shot penalty

Tabela 3: Mapeamento de Ações para IDs.

Para definir se o resultados das ações foi promissor, foram usadas 3 métricas distintas: o xG, o VAEP e uma classificação binária que indica o resultado da ação (no caso de um chute ao gol, se a ação marcou um ponto). O xG foi estimado com uso de um *RandomForest*, considerando a parte do corpo associada ao chute (categoricamente), a distância até o gol e o ângulo entre o chute e o gol. Já o VAEP foi treinado com um *XGBoost*, discretizando-se com um valor de P_{scores} igual a $\frac{1}{2}$. Ou seja, apenas chutes cuja probabilidade de marcar foi maior que $\frac{1}{2}$ foram considerados promissores.

Os parâmetros usados na execução foram: tamanho máximo dos padrões igual a 5, suporte mínimo igual a 50, multiprocessamento habilitado. Traduzindo-se para as *flags* utilizadas no programa:

```
-L 5 -m 50 -M 1
```

Apesar dos bons resultados no rugby em [3], as sequências obtidas deixaram a desejar quanto aplicadas ao futebol, independentemente da métrica usada. Em todos os casos, as sequências de jogadas obtidas eram muito redundantes, continham passes (0) em excesso ou não eram muito surpreendentes. Por exemplo, uma sequência de passes seguida por um cruzamento (7). Os arquivos de saída gerados podem ser obtidos aqui para o xG, aqui para o binário e aqui para o VAEP.

Uma possível explicação para os resultados encontrados é o fato de o futebol ser um esporte muito dinâmico e que, por essa natureza, exigiria outra abordagem para discretização dos dados, para além das sequências de ações. De fato, muitos gols são (e não são) marcados a partir de uma sequência de passes, por exemplo. Desse modo, uma abordagem que consiga mapear o campo de maneira mais sutil é necessária para que se possa realizar uma mineração de sequências usando o algoritmo proposto por [3].

8 Resultados

gfdfd

9 Conclusão

fefre

Referências Bibliográficas

[1] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435, 2002.

[2] Clive B. Beggs, Alexander J. Bond, Stacey Emmonds, and Ben Jones. Hidden dynamics of soccer leagues: The predictive 'power' of partial standings. *PLOS ONE*, 14(12):1–28, 12 2019.

[3] Rory Bunker, Keisuke Fujii, Hiroyuki Hanada, and Ichiro Takeuchi. Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PloS one*, 16(9):e0256329, 2021.

[4] M. Cox and T. Benjamin. *Entre Linhas: de Ajax a Zidane, a construção do futebol moderno nos gramados da Europa*. Editora Grande Área, 2022.

[5] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 1851–1861, New York, NY, USA, 2019. ACM.

[6] Pappalardo et al. A public data set of spatio-temporal match events in soccer competitions. *Nature Scientific Data*, 2019.

[7] J. Van Haaren, V. Dzyuba, S. Hannosset, and J. Davis. Automatically discovering offensive patterns in soccer match data. 2015.

[8] Miguel Paulo Martins Marques. Subgroup discovery in soccer data. Master's thesis, 2022.

[9] Jian Pei, Jiawei Han, B. Mortazavi-Asl, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. Prefixspan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering*, pages 215–224, 2001.

[10] Leszek Szczecinski and Aymen Djebbi. Understanding draws in elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 2020.

[11] Aristotelis Takvorian. *The Beautiful (Computer) Game: How Data Science Will Revolutionize the World's Most Popular Sport*. PhD thesis, 2021.

[12] A. Terzis. *Pep Guardiola - Coaching High Pressing Tactics & Sessions Against Different Formations*. SoccerTutor.com, 2023.

[13] Maaike Van Roy, Pieter Robberechts, Tom Decroos, and Jesse Davis. Valuing on-the-ball actions in soccer: A critical comparison of xt and vaep. In *Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports*, AITS. AI in Team Sports Organising Committee, dec 2020.