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

wabba labba dub dub

Sumário

Mineração de dados de evento em futebol

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

1 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 têm sido um desses esportes [11]. A própria UFMG ofertou no ano passado e novamente neste semestre a disciplina 'Ciência de Dados aplicada ao futebol', o que mostra a relevância do tema. Diversas empresas que atuam na área surgem a cada dia, e os times de futebol, no Brasil e no resto do mundo, estão investimento em seus departamentos de dados e estatística.

Nesse sentido, nosso grupo optou por estudar e compreender melhor como funciona o uso de análises estatísticas no futebol, dado o interesse geral pelo esporte, e, para isso, nos propusemos a aplicar algoritmos de mineração de dados em dados futebolísticos, sendo eles dados de súmula, dados de eventos ou até mesmo dados de *tracking* dos jogadores, 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*. Como a ideia é ter um panorama geral de diversas partidas, campeonatos e jogadores, iremos utilizar as bases de dados disponibilizadas sobre as 5 grandes ligas de futebol europeu das temporadas 17/18 da empresa *Wyscout*.

1.1 Base de dados

As bases de dados utilizadas na ferramenta consistirá na base principal disponibilizada pela empresa *Wyscout*, consistindo em uma base de dados de evento das 5 grandes ligas europeias na temporada 17/18.

Cada empresa fornecedora de dados possuem 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, foi utilizado o pacote `pysubgroup`, que fornece uma aglomeração de algoritmos do estado da arte de descoberta de subgrupos em um formato simples e leve para serem utilizados.

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 arquivo *README.md* do repositório.

3 Referencial Teórico

No decorrer do trabalho, duas importantes métricas de análise ofensivano 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

A métrica

3.2 VAEP

falar da vaep [5]

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.

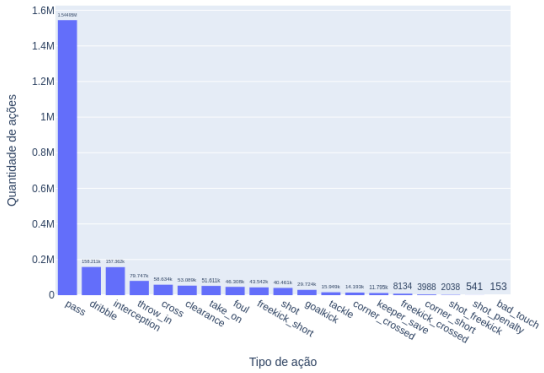

Figura 1: Distribuições dos tipos de ação

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.

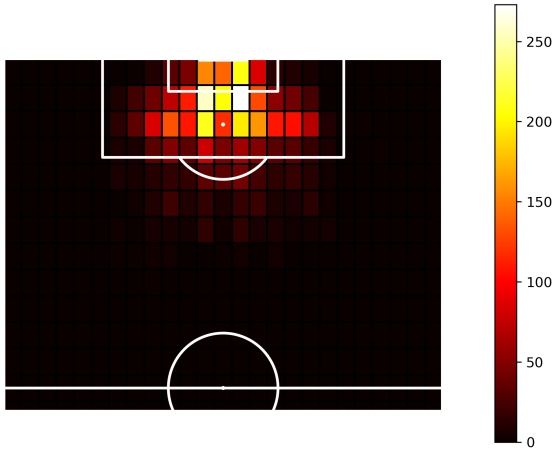

Figura 2: Distribuições dos tipos de ação

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

Equipe	Distância média entre passes (m)
Club Atlético de Madrid	17.581
FC Barcelona	17.618
UD Las Palmas	17.864
Real Madrid Club de Fútbol	17.905
Sevilla FC	18.070
Real Betis Balompié	18.659
Villarreal Club de Fútbol	18.703
Real Club Deportivo de La Coruña	18.990
Reial Club Deportiu Espanyol	19.018
Real Sociedad de Fútbol	19.051
Valencia Club de Fútbol	19.090
Deportivo Alavés	19.225
Real Club Celta de Vigo	19.324
CD Leganés	19.340
Levante UD	19.353
Málaga Club de Fútbol	19.527
Girona FC	20.133
Athletic Club Bilbao	20.137
Getafe Club de Fútbol	20.608
SD Eibar	20.982

Tabela 3: La Liga

Equipe	Distância média entre passes (m)
Paris Saint-Germain FC	17.205
O.G.C. Nice Côte d'Azur	17.923
Olympique de Marseille	17.977
Lille OSC Métropole	17.999
En Avant Guingamp	18.008
AS Saint-Étienne	18.320
Olympique Lyonnais	18.377
Angers SCO	18.414
FC Nantes	18.544
Amiens SC	18.570
Espérance Sportive Troyes Aube Champagne	18.604
FC Girondins de Bordeaux	18.621
AS Monaco FC	19.099
Stade Rennais FC	19.141
FC Metz	19.266
RC Strasbourg Alsace	19.424
Montpellier HSC	19.592
Toulouse FC	19.677
Stade Malherbe Caen	19.805
Dijon FCO	19.913

Tabela 4: Ligue 1

Equipe	Distância média entre passes (m)
SSC Napoli	16.864
FC Internazionale Milano	17.370
UC Sampdoria	17.980
AC Milan	18.514
Torino FC	18.564
Benevento Calcio	18.613
ACF Fiorentina	18.654
AC Chievo Verona	18.706
SS Lazio	18.712
Atalanta Bergamasca Calcio	18.757
AS Roma	18.759
Società Polisportiva Ars et Labor 2013	18.763
Genoa CFC	18.890
Cagliari Calcio	18.897
Juventus FC	18.983
Udinese Calcio	19.043
Bologna FC 1909	19.056
FC Crotone	19.103
Hellas Verona FC	19.202
US Sassuolo Calcio	19.646

Tabela 5: Serie A

Equipe	Distância média entre passes (m)
Rasen Ballsport Leipzig	18.123
TSV Bayer 04 Leverkusen	18.319
BV Borussia 09 Dortmund	18.386
Borussia VfL Mönchengladbach	18.648
FC Bayern München	18.835
TSG 1899 Hoffenheim	19.136
FC Schalke 04	19.255
Hannover 96	19.269
VfB Stuttgart 1893	19.616
1. FC Köln	19.626
Hertha BSC	19.762
Hamburger SV	19.834
VfL Wolfsburg	19.836
SV Werder Bremen	19.883
1. FSV Mainz 05	19.884
SC Freiburg	19.904
FC Augsburg	20.228
Eintracht Frankfurt	20.606

Tabela 6: Bundesliga

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

minerar sequencias antes de gols. nao so tipo deacao,mas pegar posição no campo, jogadores, etc

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.