

Mineração de dados aplicada à dados de evento no futebol

Luis Felipe Ramos Igor Lacerda Faria da Silva Matheus
Tiago Pimenta de Souza

2024-07-28

Uso de ciência de dados aplicada ao esporte



Moneyball (2011)



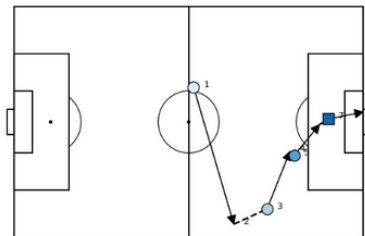
FAME

Dados de súmula no futebol



Estatísticas da final da Copa do Mundo de 2022

Dados de evento



	time_seconds	type	player	team
○ 1	1835.0	pass	Rodrigo Hernández Cascante	Manchester City
- - 2	1838.0	dribble	Kyle Walker	Manchester City
○ 3	1840.0	pass	Kyle Walker	Manchester City
- - 4	1841.0	dribble	Kevin De Bruyne	Manchester City
○ 5	1842.0	pass	Kevin De Bruyne	Manchester City
- - 6	1843.0	dribble	Gabriel Fernando de Jesus	Manchester City
■ 7	1845.0	shot	Gabriel Fernando de Jesus	Manchester City

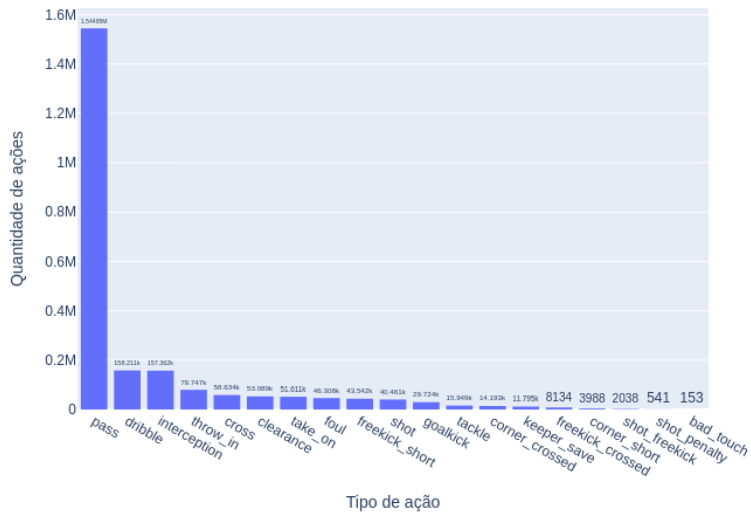
Jogada do Manchester City

SPADL

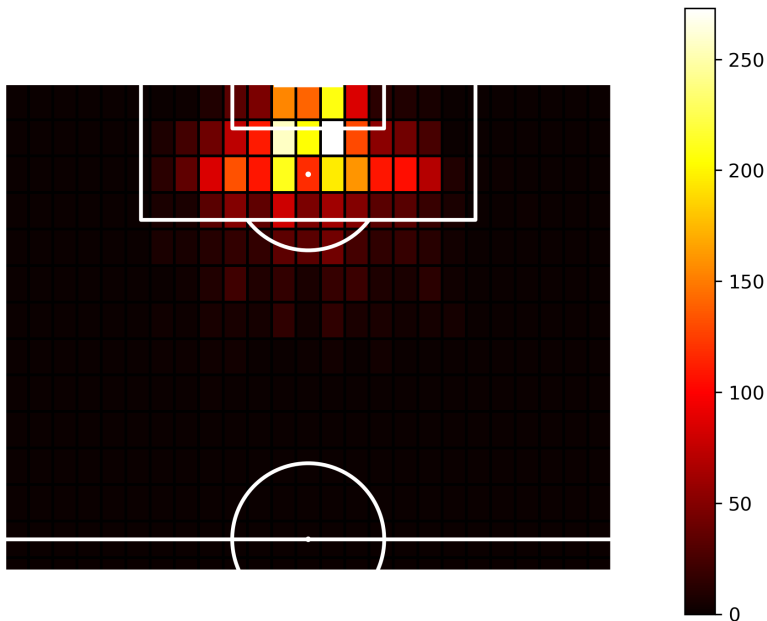
Atributo	Descrição
game_id	O ID do jogo no qual a ação foi realizada
period_id	O ID do período do jogo no qual a ação foi realizada
seconds	O tempo de início da ação
player	O jogador que realizou a ação
team	O time do jogador
start_x	A localização x onde a ação começou
start_y	A localização y onde a 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

Table: Descrição dos dados no formato SPADL

Distribuição dos tipos de ações



Mapa de calor dos gols



Objetivos

Descoberta de subgrupos

- ▶ Trabalho motivador: *Subgroup Discovery in Soccer Data*
- ▶ Reproduzir os experimentos do artigo e experimentar com novos alvos e um outro algoritmo de SD

Objetivos

Mineração de sequências

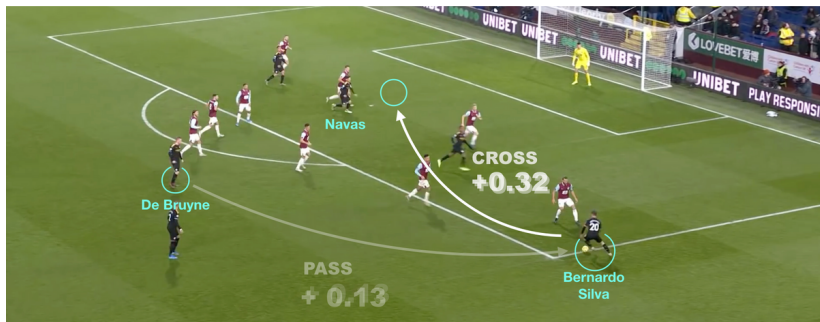
- ▶ Trabalho motivador: *Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: An application to rugby union*
- ▶ Aplicar as ideias em sequências de jogadas de futebol

Gols Esperados (xG)

- ▶ Métrica de cálculo de probabilidade de gol
- ▶ Diferentes *features* podem ser utilizadas:
 - ▶ Distância para o gol
 - ▶ Ângulo de chute



- ▶ Ações valem mais do que apenas gols
- ▶ O quanto cada jogada contribui para o aumento da chance de gol e diminuição da chance de tomar gol?



Descoberta de subgrupos - Extração de *Features*

Atributo	Descrição
start_x	A localização x onde a ação começou
start_y	A localização y onde a ação começou
num_events	Contagem total de eventos (ações) na jogada
num_passes	Contagem total de passes na jogada
num_dribles	Contagem total de dribles na jogada
play_duration	Duração total da jogada
player_rank	Ranking do jogador
play_distance	Distância (euclidiana) total percorrida na jogada
play_mean_distance_to_the_goal	Distância (euclidiana) média de cada jogada até o gol
play_std_distance_to_the_goal	Desvio padrão da distância (euclidiana) de cada jogada até o gol
play_distance_towards_goal	Distância percorrida nas jogadas em direção ao gol (considerando só o eixo x)
bodypart_name	A parte do corpo do jogador usada para a ação
ratio_distance	Razão entre play_distance e play_distance_towards_goal
total_time_per_play	Desvio padrão da distância (euclidiana) de cada jogada até o gol
play_duration	Duração total das jogadas
total_time_per_play	Média do tempo por jogada (razão entre play_duration e num_events)
play_speed	Razão entre play_distance e play_duration
play_speed_towards_goal	Razão entre play_distance_towards_goal e play_duration

Descoberta de subgrupos - Estratégias

- ▶ Avaliações em 3 tipos de granularidades
 - ▶ Avaliação de Todo o Dasatet VS Duas ligas diferentes
 - ▶ Avaliação de três times de uma mesma liga (1º, 10º e 20º)
 - ▶ Avaliação de duas abordagens de SD (Beam Search e SSD++)
- ▶ Todas as avaliações nos três alvos (Gol/Não Gol, xG e VAEP)
- ▶ As duas primeiras avaliações com Beam Search do pacote do *pysubgroup*, com profundidade 3, buscando 100 subgrupos com largura do Beam de 250
- ▶ *SSD++* com profundidade 3, largura do beam de 25 e máximo de regras sendo 20

Descoberta de subgrupos - Resultados Métricas

Algoritmo	Dado	Target	WRAcc do melhor	Cobertura Total
Beam Search	Todo dataset	Gol	0.0312	0.5099
Beam Search	Todo dataset	xG	0.0187	0.5046
Beam Search	Todo dataset	VAEP	0.0307	0.5099
Beam Search	Inglaterra	Gol	0.03192	0.4980
Beam Search	Inglaterra	xG	0.0203	0.4933
Beam Search	Inglaterra	VAEP	0.0317	0.4977
Beam Search	Espanha	Gol	0.0307	0.5110
Beam Search	Espanha	xG	0.0195	0.3983
Beam Search	Espanha	VAEP	0.0304	0.5120
Beam Search	Man City	Gol	0.0483	0.5721
Beam Search	Man City	xG	0.0303	0.4959
Beam Search	Man City	VAEP	0.0487	0.5656
Beam Search	Newcastle	Gol	0.0405	0.6561
Beam Search	Newcastle	xG	0.0247	0.3878
Beam Search	Newcastle	VAEP	0.0420	0.6561
Beam Search	West Bromwich	Gol	0.0293	0.5499
Beam Search	West Bromwich	xG	0.0307	0.3989
Beam Search	West Bromwich	VAEP	0.0287	0.5254
SSD++	Todo Dataset	Gol	0.0164	0.7818
SSD++	Todo Dataset	xG	0.0117	0.9956
SSD++	Todo Dataset	VAEP	0.0194	0.8114

Descoberta de subgrupos - Resultados entre Ligas

Dataset	Target	Subgrupo
Todo dataset	Gol	$\text{shot_angle_from_goal} \geq 0.60$ AND $\text{shot_distance_from_goal} < 11.26$
Todo dataset	xG	$\text{shot_angle_from_goal} \geq 0.60$ AND $\text{shot_distance_from_goal} < 11.26$ AND $\text{start_x} \geq 96.60$
Todo dataset	VAEP	$\text{num_dribbles} : [0 : 1[$ AND $\text{shot_angle_from_goal} \geq 0.60$ AND $\text{shot_distance_from_goal} < 11.26$
Inglaterra	Gol	$\text{shot_angle_from_goal} \geq 0.61$ AND $\text{shot_distance_from_goal} < 11.26$
Inglaterra	xG	$\text{num_dribbles} : [0 : 1[$ AND $\text{shot_distance_from_goal} < 11.26$ AND $\text{start_x} \geq 96.60$
Inglaterra	VAEP	$\text{play_duration} < 1.39$ AND $\text{shot_distance_from_goal} < 11.26$ AND $\text{total_time_per_play} < 0.65$
Espanha	Gol	$\text{shot_angle_from_goal} \geq 0.61$ AND $\text{shot_distance_from_goal} < 11.04$ AND $\text{start_x} \geq 96.60$
Espanha	xG	$\text{shot_angle_from_goal} \geq 0.61$
Espanha	VAEP	$\text{play_distance_towards_goal} : [0.0 : 7.35[$ AND $\text{ratio_distance} : [0.0 : 0.12[$ AND $\text{shot_distance_from_goal} < 11.04$

Descoberta de subgrupos - Resultados entre Times

Dataset	Target	Subgrupo
Man City	Gol	player_rank : [0.02 : 0.03[AND shot_distance_from_goal < 10.90 AND start_x > 96.60
Man City	xG	num_dribbles : [0 : 2[AND shot_angle_from_goal \geq 0.62 AND start_x \geq 96.60
Man City	VAEP	bodypart_name = foot_left AND shot_distance_from_goal < 11.03
Newcastle	Gol	shot_angle_from_goal \geq 0.61 AND shot_distance_from_goal < 11.04
Newcastle	xG	num_dribbles : [1 : 3[AND start_x \geq 96.60 AND start_y : [33.32 : 38.08[
Newcastle	VAEP	bodypart_name = foot_right AND shot_distance_from_goal < 11.04
West Bromwich	Gol	bodypart_name = head/other AND shot_distance_from_goal < 10.52 AND start_x \geq 96.60
West Bromwich	xG	num_dribbles : [0 : 2[AND start_x \geq 96.60
West Bromwich	VAEP	bodypart_name = head/other AND play_distance_towards_goal : [39.90 : 56.70[AND start_x \geq 96.60

Descoberta de subgrupos - Resultados entre Algoritmos

Dataset	Target	Subgrupo
BS	Gol	$\text{start_x} \geq 96.60$
BS	Gol	$\text{num_dribbles} : [0 : 1[\text{ AND } \text{shot_angle_from_goal} \geq 0.60 \text{ AND } \text{shot_distance_from_goal} < 11.26$
BS	xG	$\text{start_x} \geq 96.60 \text{ AND } \text{start_y} : [31.96 : 37.40[$
BS	xG	$\text{shot_angle_from_goal} \geq 0.60 \text{ AND } \text{shot_distance_from_goal} < 11.26$
BS	VAEP	$\text{start_x} \geq 96.60$
BS	VAEP	$\text{num_dribbles} : [0 : 1[\text{ AND } \text{shot_distance_from_goal} < 11.26$
SSD++	Gol	$\text{shot_angle_from_goal} \geq 0.5591 \text{ AND } \text{bodypart_name} = \text{foot_right} \text{ AND } \text{start_x} \geq 95.55$
SSD++	Gol	$\text{player_rank} \geq 0.016 \text{ AND } 2.0 \leq \text{num_passes} \leq 3.0 \text{ AND } 0.359 \leq \text{shot_angle_from_goal} \leq 0.5591$
SSD++	xG	$27.2 \leq \text{start_y} \leq 34.0 \text{ AND } 4.64 \leq \text{play_speed} \leq 6.87 \text{ AND } \text{num_dribbles} \geq 1.0$
SSD++	xG	$\text{play_speed} \geq 8.98 \text{ AND } \text{play_duration} \geq 2.07 \text{ AND } 0.004 \leq \text{player_rank} \leq 0.016$
SSD++	VAEP	$\text{shot_angle_from_goal} \geq 0.56 \text{ AND } 0.004 \leq \text{player_rank} \leq 0.016 \text{ AND } 0.93 \leq \text{total_time_per_play} \leq 1.76$
SSD++	VAEP	$\text{shot_angle_from_goal} \geq 0.56 \text{ AND } \text{play_speed_towards_goal} \geq 1.52 \text{ AND } 12.39 \leq \text{play_mean_distance_to_the_goal} \leq 28.09$

Descoberta de subgrupos - Conclusões

- ▶ Alvos
 - ▶ VAEP se mostrou coerente com xG e Gol/Não Gol
 - ▶ xG levou a métricas menores, mas grupos são parecidos
 - ▶ Nenhuma métrica se sobressaiu
- ▶ Algoritmos
 - ▶ Beam Search com maior WRAcc, mas menor cobertura
 - ▶ Acreditamos que os resultados são complementares
- ▶ Utilidade dos padrões
 - ▶ Beam Search mais simples, úteis para compara times
 - ▶ SSD++ mais diferentes, permitindo análises mais exóticas

Mineração de sequências

- ▶ Resultados não muito promissores usando o SPP

Fim

- ▶ Perguntas, comentários, etc