Moneyball - Definição de times da NBA baseada em estatísticas visando maior performance por dólar gasto

Vítor Corrêa Silva

***Index Terms*— Moneyball, NBA, Multiobjective Optimization**

I. Resumo

Nesse projeto será modelado um algoritmo evolucionário para optimização da formulação de um time de basquete da NBA, utilizando o menor capital possível. Serão usados dados oficiais da NBA sobre os jogadores ao longo da temporada para modelar o possível comportamento dos times.

II. Introdução

No contexto dos esportes coletivos é bastante complexo definir a melhor formatação para um time de elite, tendo em vista que parâmetros fora do ambiente de jogo tem um peso imprevisível nos resultados observados em campo. Tentando tornar esses eventos caóticos, previsíveis estrategistas e estatísticos especializados em esportes desenvolveram métricas quantitativas e qualitativas que descrevem o esporte de uma forma que seja empírica e fiel. Estas

Mesmo com anos de estudo, o gap entre os parâmetros medidos e um modelo suficientemente assertivo de predição de vitórias esportivas ainda é abissal. Uma forma que vários estudiosos encontraram de tornar o processo de mais simples é o uso das métricas individuais dos jogadores de forma a ranquear os times pelas suas qualidades e defeitos.

Esse tipo de estudo estatístico do esporte foi primeiramente feito para o baseball em 1858 quando o jornalista esportivo Henry Chadwick o box score, que disponibilizava as estatísticas individuais de cada jogador, assim como do time após cada partida. O estudo se intensificou no século 20 com a fundação da Society for American Baseball Research que desenvolveu o que ficou conhecido como Sabermetrics, o que é definida como uma metodologia para a compreensão do baseball através de fatos objetivos e mensuráveis.

Tem se tornado muito comum o uso desse tipo de estatística é para formação ou preenchimento de vagas abertas em times esportivos. Isso por si só é suficiente para modelar um problema de otimização mono-objetivo, entretanto, nas grandes ligas profissionais não somente o resultado é importante, mas também o quanto esse resultado custou a organização na temporada. Isso torna a questão da formação de um time bem mais complexo, pois trata-se não só do melhor time, mais sim do melhor time que se pode comprar, gastando o mínimo possível.

O primeiro exemplo de sucesso da Sabermetrics aplicada a formulação de times, apelidada na época de Moneyball, foi feita pelos Oakland Athletics em 2002, aonde com um gasto de mais de um milhão de dólares menor por jogo ganho, alcançaram o mesmo número de vitórias do New York Yankees, 103. Esse fato chamou a atenção de diversas organizações dentro e fora do baseball. Sendo alguns exemplos o futebol, o vôlei e o basquete, que será o foco do modelo descrito neste trabalho.

III. Modelagem

O objetivo é utilizando a base de dados de atletas da temporada regular da NBA 2021-22 e o mapa salarial dos atletas para temporada 2022-23, formular o melhor time possível, gastando o mínimo possível. O time em questão será formado por doze jogadores, dirtribuídos na entre as posições da seguinte forma.

Posição	Sigla	Total de Jogadores
Armador	PG	3
Ala-armador	SG	2
Ala	SF	2
Ala de força	PF	3
Pivô	C	2

A. Função Objetivo

O problema possui duas funções objetivo uma que define a qualidade do time que foi montado e a outra que define o custo total do time. A função de avaliação pode ser descrita da seguinte forma:

$$f_1(x) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} P_j E_j(i) x_i \tag{1}$$

Aonde :

- i é o índice númerico que representa o jogador
- x_i é um vetor binário que indica se o jogador pertence ou não ao time formulado
- $E_1(i)$ é o total de pontos por jogo
- $E_2(i)$ é o número de assitências por jogo
- $E_3(i)$ é o número de roubos por jogo
- $E_4(i)$ é o número de bloqueios por jogo
- $E_5(i)$ é o número de rebotes por jogo
- $E_7(i)$ é o número de turnovers por jogo
- $E_8(i)$ é o número de faltas por jogo
- P_j é um vetor de contantes de peso para cada um dos parâmetros contabilizados, aonde :

  - $P_j > 0, j \leq 6$
  - $P_j < 0, 7 \leq j \leq 8$
- M é o total de parâmetros utilizados
- N é o total de jogadores

Todos os valores serão usados pós normalização para evitar que valores muito grandes como a pontução, quando comparado ao número de faltas, por exemplo, force uma tendência sobre o modelo.

O custo do time poderá ser descrito da seguinte forma :

$$f_2(x) = \sum_{i=1}^{N} S(i)x_i \tag{2}$$

Aonde :

- $S(i)$ é uma função que para um índice de indetificação do jogador, retorna o salário pago a ele
- x_i é um vetor binário que indica se o jogador pertence ou não ao time formulado

B. Restrições

As únicas restrições usadas nesse modelo serão as de quantidade de jogadores por posição já expressadas anteriormente.

C. Modelagem Formal

$$\begin{cases} \max f_1(x) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} P_j E_j(i) x_i \\ \min f_2(x) = \sum_{i=1}^{N} S(i) x_i \end{cases} \tag{3}$$

$$\text{Sujeito a : } \begin{cases} \sum_{N}^{i=1} P_g(i) x_i \leq 3 \\ \sum_{N}^{i=1} S_g(i) x_i \leq 2 \\ \sum_{N}^{i=1} S_f(i) x_i \leq 2 \\ \sum_{N}^{i=1} P_f(i) x_i \leq 3 \\ \sum_{N}^{i=1} C(i) x_i \leq 2 \end{cases}$$

Por se tratar de um sistema multivariável e não completamente mapeado é correto assumir que tratasse de um problema de otimização linear multimodal, pois é correto assumir que existam diversas combinações possíveis de times. O algoritmo que se pretende usar seria um algoritmo evolucionário modelado especificamente para o problema em razão do número de variáveis que o sitema apresenta.

REFERENCES

[1] Kubatko J, Oliver D, Pelton K, Rosenbaum D. A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports. 2007;3(3):1–24.

[2] Ahmed F, Deb K, Jindal A. Multi-objective optimization and decision making approaches to cricket team selection. Applied Soft Computing. 2013;13(1):402–14.

[3] Ahmed F, Deb K, Jindal A. Evolutionary multi-objective optimization and decision making approaches to cricket team selection. In: Panigrahi BK, Suganthan PN, Das S, Satapathy SC, editors. Swarm, Evolutionary, and Memetic Computing. SEMCCO 2011: Proceedings of the Second International Conference on Swarm, Evolutionary, and Memetic Computing; Berlin, Heidelberg: Springer-Verlag; 2011. p. 71–78.

[4] García J, Ibáñez SJ, Martinez De Santos R, Leite N, Sampaio J. Identifying Basketball Performance Indicators in Regular Season and Playoff Games. Journal of Human Kinetics. 2013;36:161–8. pmid:23717365