Entendendo a evolução das habilidades de jogadores de futebol através das pontuações do jogo eletrônico FIFA

Ivan R. Soares Jr., Renato M. Assunção, Pedro O. S. Vaz de Melo

Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - MG - Brazil
{ivansoares,assuncao,olmo}@dcc.ufmg.br

Resumo. A popularidade do futebol gera interesse em caracterizar o desenvolvimento de jogadores de elite, seja por razões comerciais ou de entretenimento. A EA Sports, produtora da franquia de jogos eletrônicos FIFA, investe em avaliar os atletas para representá-los de forma realista. Neste artigo são estudadas as pontuações atribuídas em múltiplas atualizações como medições longitudinais e é avaliada a possibilidade de descrever as curvas de desenvolvimento através de um número relativamente pequeno de padrões. É proposta uma transformação das séries de medições que visa enfatizar formatos e são utilizadas técnicas de análise de agrupamentos nas observações, a saber k-means e Spectral Clustering. São avaliados os resultados para múltiplas habilidades de jogadores em diferentes grupos de posições e apresentados 11 Padrões de Evolução identificados nos agrupamentos. É utilizado o índice Average Silhouette Width.

Categories and Subject Descriptors: I.5.3 [**Pattern Recognition**]: Clustering

Keywords: clustering, data mining, sports analytics

1. INTRODUÇÃO

O futebol é o esporte mais popular do mundo. O número de estimado espectadores supera 3 bilhões[1] e há mais de 100 mil atletas profissionais registrados[2]. O mercado das negociações de contratos de jogadores é multi-milionário e, além das partidas reais, até as *video game* mobilizam milhões de pessoas. O FIFA é o carro chefe de sua produtora, a *EA Sports*, com mais de 15 milhões de unidades vendidas em 2016[3]. Esses são indicativos de que há interesse por parte dos clubes e dos fãs em caracterizar a evolução do desempenho dos jogadores.

O FIFA é construído a partir de avaliações reais, disponibilizadas por 25 produtores, 400 curadores e uma rede com mais de 8.000 técnicos, olheiros e sócios-torcedores. O resultado é um conjunto rico e de grande escala. Em vez de considerar apenas estatísticas simples, como posse de bola ou chutes a gol, são atribuídas a cada jogador notas de 0 a 100 em habilidades como *Finalização*, *Posicionamento*, *Marcação* e *Força* (há pelo menos 25). Alguns atletas até dizem usar as avaliações para entender seus pontos fracos ou jogar algumas partidas virtuais para começar a conhecer seus futuros oponentes[4].

Essas avaliações têm, portanto, grande potencial para aplicações de análise de dados esportivos. Parte desse potencial já foi explorado por outros trabalhos, como Cotta et al. [2016], que apresenta o conjunto de dados, Vroonen et al. [2017] e Soto-Valero [2017]. Dada a precisão com a qual os jogadores são descritos no FIFA, neste trabalho propomos uma abordagem para caracterizar essa evolução dos atletas a partir de sua base de dados.

A abordagem proposta é inspirada em um trabalho similar realizado dentro do contexto de obras literárias: ainda em meados do século XX, o escritor Kurt Vonnegut, percebendo semelhança entre

[1] https://www.fifa.com/worldcup/news/2014-fifa-world-cuptm-reached-3-2-billion-viewers-one-billion-watched--2745519

[2] https://www.fifa.com/mm/document/fifafacts/bcoffsurv/emaga_9384_10704.pdf

[3] https://www.forbes.com/sites/greatspeculations/2017/10/10/fifa-remains-eas-bread-and-butter/

[4] https://www.datamakespossible.com/meet-data-master-ea-sports-fifa/

algumas estórias, definiu o conceito de arcos emocionais para caracterizar livros de ficção e hipotetizou que há seis formas básicas dominantes. Vonnegut ainda sugeriu que computadores poderiam um dia ser usados para analisar sua proposta. Reagan et al. [2016] abordaram o problema com uma metodologia que incluía algoritmos para *clustering* utilizados em dados com estrutura sequencial e técnicas estatísticas de análise de sentimentos.

Neste trabalho, através da elaboração de uma representação para as séries de pontuações e utilizando técnicas de *Cluster Analysis*, investigamos a existência de formas análogas para as curvas de evolução das habilidades de jogadores de futebol, dividindo-os de acordo com os grupos de posições em campo. Encontramos evidências que sugerem a existência dessas formas e exploramos os resultados com 11 Padrões de Evolução, que ocorrem para múltiplas habilidades e se distinguem por suas características e velocidades de melhoria, deterioração ou estagnação. Trabalhos futuros podem abordar aplicações como ferramentas de auxilio à identificação de talentos e à tomada de decisões estratégicas em campo.

2. TRABALHOS RELACIONADOS

Cotta et al. [2016] propõem a utilização das notas por habilidade da franquia de *video games* FIFA em aplicações de análise de dados esportivos, exemplificada com uma avaliação dos resultados de partidas reais a partir das pontuações dos jogadores em campo. Vroonen et al. [2017], utilizando a mesma fonte de dados, apresentam o sistema APROPOS (*Algorithm for PRediction Of the Potential Of Soccer players*), baseado em *k-nearest neighbors regression*. Os autores avaliam critérios de similaridade absoluta e critérios baseados na evolução das pontuações dos jogadores. Soto-Valero [2017] utiliza *Gaussian mixture models* para agrupar os jogadores a partir dos dados do FIFA, mas apenas com uma visão transversal, sem incorporar aspectos temporais, e sem separar os jogadores por grupos de posições em campo, o que levou o autor a identificar apenas diferenças relacionadas a essas posições. Akhanli and Hennig [2017] apresentam uma metodologia para agrupar e visualizar estatísticas transversais sobre o desempenho de jogadores de futebol, principalmente variáveis de contagem.

Sardá-Espinosa [2017] apresenta uma visão geral do problema de agrupar séries temporais com base em sua forma. Diferentes estratégias são identificadas pela função de distância e pelo procedimento de obtenção de protótipos adotado. Há ênfase em estratégias baseadas em *Dynamic Time Warping* e no algoritmo *k-Shape* [Paparrizos and Gravano 2017], mas os autores apontam para a necessidade considerar a aplicação na escolha de uma estratégia apropriada. Reagan et al. [2016] e McFee and Ellis [2014] apresentam exemplos envolvendo identificação de padrões em dados com estrutura sequencial.

A possibilidade de fazer predições a partir da busca de jogadores com séries de pontuações similares apresentada por Vroonen et al. [2017] sugere a existência de padrões nessas séries. E este trabalho explora exatamente uma forma de identificá-los e revelar sua estrutura. Realizar uma análise em grande escala dos padrões de evolução das habilidades em cada um dos grupos de posições, para o melhor de nosso conhecimento, sem precedentes na literatura, representa uma extensão importante.

3. METODOLOGIA

3.1 Base de Dados

Proveniência. A fonte original do conjunto de dados é o site SOFIFA[5], mantido pela comunidade de jogadores do título. No site, há um registro de todas as pontuações desde a versão FIFA 07, incluindo atualizações. Cotta et al. [2016] realizaram uma coleta e apresentaram os detalhes de todos os dados disponíveis até o período (Outubro de 2015) e a mesma versão é utilizada neste trabalho.

[5] https://sofifa.com/

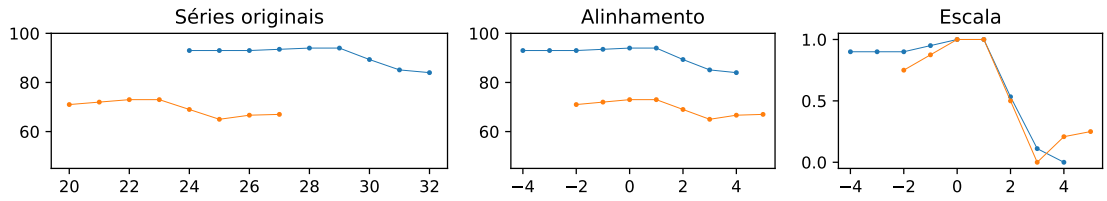

Fig. 1. Exemplo do passo-a-passo da transformação. Esquerda: Séries originais. Meio: Séries re-indexadas considerando a primeira ocorrência do valor máximo. Direita: Valores ajustados para intervalo entre 0 e 1.

Médias por idade. Principalmente para as versões mais recentes, o conjunto de dados inclui múltiplas avaliações de um mesmo jogador em um mesmo ano. Como no trabalho as pontuações são consideradas de acordo com as idades dos jogadores, a pontuação de um jogador em uma certa idade é obtida com a média das pontuações das avaliações que ele recebeu enquanto tinha aquela idade. As idades são calculadas a partir das datas de nascimento e das datas das atualizações.

Grupos de posições. Os registros dos jogadores informam sua posição. Exemplos: CB: *Center-back* (zagueiro central) e ST: *Striker* (finalizador). Essas posições são categorizadas em quatro grandes grupos, de acordo com o critério utilizado pelo próprio jogo, que atribui uma cor para cada grupo. As etapas de seleção de habilidades principais (Seção 3.2) e construção do grafo de similaridades (Seção 3.6) são feitas separadamente para cada um deles, a saber: Ataque, Meio-campo, Defesa e Goleiros.

Jogadores com história curta. É comum que um atleta que tenha recentemente ingressado pela primeira vez em uma das equipes disponíveis no jogo já apareça na próxima atualização. Dessa forma, a visão parcial de sua carreira é bem restrita, tornando difícil a análise de seu desenvolvimento. Há um problema análogo com os anos finais. Para mitigar essas limitações, as análises foram restritas aos jogadores que têm pontuações disponíveis em pelo menos 5 (cinco) idades, totalizando 11,061.

3.2 Seleção das habilidades principais

Foram analisadas como são ponderadas as habilidades para o cálculo da pontuação geral (*Overall*), de forma a determinar quais são as mais relevantes para cada grupo de posições. A estimativa foi feita com um modelo de regressão auxiliar, relacionando as demais pontuações com a pontuação geral para cada atualização. Foi usado o cálculo de importâncias de características[6] disponível na implementação do algoritmo *Random Forests* na biblioteca *scikit-learn*. As 5 habilidades mais relevantes para cada grupo de posições foram selecionadas (elas podem ser vistas na Figura 5).

3.3 Padrão de evolução constante

Para parte dos atletas, as pontuações variam relativamente pouco. O intervalo de variação foi definido como a diferença entre a maior e a menor pontuação. Considerando as notas de 0 a 100, para cada grupo de posições e para cada habilidade, jogadores cujo intervalo de variação fosse menor do que 5 (cinco) pontos (12.95% do total) foram considerados como membros de um agrupamento com *padrão de evolução constante*.

3.4 Alinhamento e escala

As séries temporais utilizadas consistem das pontuações médias por idade para cada jogador em cada uma das habilidades principais. Dado o número pequeno de medições – tipicamente 8 – estratégias

[6] http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

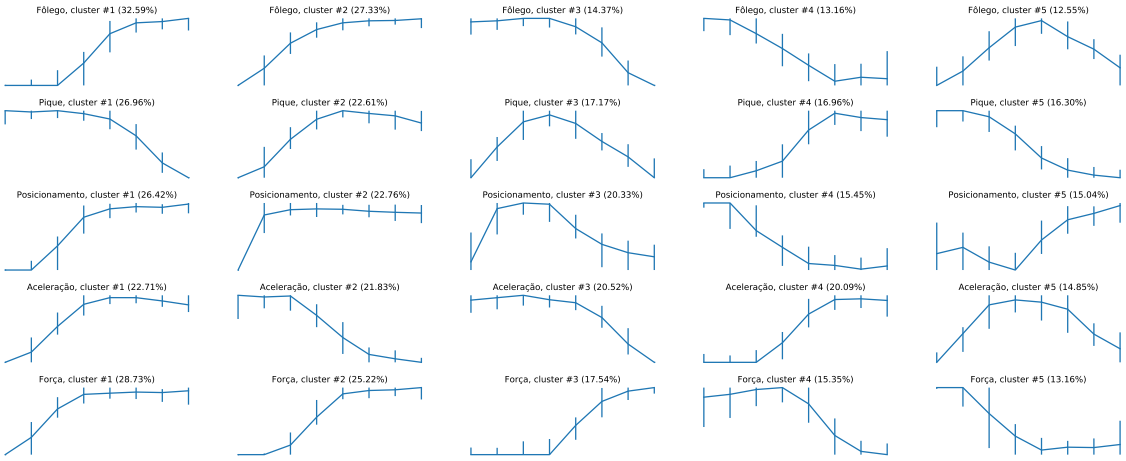

Fig. 2. Perfil dos agrupamentos encontrados utilizando *k-means* para jogadores de meio de campo com 8 pontuações em idades diferentes. Cada linha exibe os agrupamentos para uma habilidade e eles são ordenados da esquerda para a direita em ordem decrescente de número de jogadores. As curvas exibem a mediana das pontuações para os jogadores no agrupamento e as barras verticais indicam a dispersão, com extremidades nos quantis 25% e 75%.

baseadas em extração de características (*feature-based*) foram consideradas e julgadas como sendo de aplicabilidade limitada no problema. Consequentemente, uma abordagem baseada nas próprias medições (*observation-based clustering*) foi priorizada (Caiado et al. [2016] discutem essa distinção).

Um dos passos iniciais foi escolher uma transformação adequada para representá-las (Figura 1). A escolha dessa transformação foi feita para refletir a ênfase em capturar similaridades no formato das trajetórias, desassociando-as, por exemplo, do intervalo de variação. Além disso, como é possível ter jogadores com padrões de evolução semelhantes, mas que ingressaram na carreira em idades diferentes, é necessário considerar o alinhamento das séries.

Em relação à escala, as séries foram ajustadas para o intervalo entre 0 e 1, referidas como pontuações normalizadas, sendo 0 a pontuação mínima do jogador e 1 a máxima. Para mitigar problemas com o alinhamento, em vez de usar as idades como índice das pontuações, as séries foram reindexadas utilizando um intervalo de número inteiros de forma que a pontuação com índice 0 fosse a primeira ocorrência do valor máximo. Especificamente, sendo $x_i(t)$ a função que fornece a pontuação do i-ésimo jogador na idade t (já tendo sido selecionada uma habilidade), $\mathbf{D}_{x_i}$ seu conjunto domínio contendo as idades com pontuações disponíveis para o jogador e t_i^* a menor idade na qual o jogador atinge sua pontuação máxima, temos que as séries são representadas por vetores de tamanho $(z^+ - \ z^- + 1)$, indexados por $z \in [z^-, z^+]$, onde z^- e z^+ são o menor e o maior índice entre todos os jogadores, respectivamente. Ou seja, sendo $Z_i = \{t - t_i^* : t \in \mathbf{D}_{x_i}\}$ o conjunto de índices do i-ésimo jogador, $z^- = \min_i \ \min(Z_i)$ e $z^+ = \max_i \ \max(Z_i)$. A representação final $\mathbf{r}_i$ da evolução de cada habilidade dos jogadores é obtida com a Equação 1, em que $x_i^- = \min_t \ x_i(t)$ e $x_i^+ = \max_t \ x_i(t)$ representam as pontuações mínima e máxima do jogador, respectivamente.

$$\mathbf{r}_{iz} = \begin{cases} \frac{x_i(z+t_i^*) - x_i^-}{x_i^+ - x_i^-} & \text{se } z + t_i^* \in \mathbf{D}_{x_i}, \\ \text{N.A.} & \text{c.c.} \end{cases} \quad (z \in [z^-, z^+]) \tag{1}$$

3.5 *k-means Clustering*

Foi feita uma análise preliminar utilizando *k-means* para avaliar se mesmo uma metodologia simples poderia identificar regularidades nas séries, e uma inspeção dos resultados indica que sim. O número de agrupamentos foi fixado em 5 e foram selecionados apenas jogadores com pontuações em exatamente

8 idades diferentes. Não foi feito o alinhamento e as séries foram comparadas por suas distâncias euclidianas. Tais configurações foram escolhidas considerando a facilidade de inspeção e o compromisso entre número de jogadores disponíveis e a abrangência em relação ao tempo de carreira deles. Na Figura 2, são exibidos os resultados para os jogadores em posições de meio de campo (são 330). Os agrupamentos obtidos não são utilizados nas demais etapas e são apresentados apenas para referência.

3.6 *Spectral Clustering*

Em [Meila 2016], a família de métodos conhecida como *Spectral Clustering* é apresentada como uma alternativa para o problema de encontrar agrupamentos que envolve a construção de um grafo ponderado com similaridades par-a-par entre os elementos e a decomposição espectral de uma matriz calculada a partir da de adjacência. Os autovetores são utilizados para obter uma representação numérica dos elementos que reflete a estrutura do grafo. Vantagens desse tipo de método incluem a capacidade de encontrar agrupamentos com formatos arbitrários (desde que haja separação) e a flexibilidade na escolha de funções de similaridade, que são menos restritivas que distâncias. A implementação utilizada é a disponível na biblioteca *scikit-learn*. Foram configuradas opções para ativar o uso de afinidades pré-computadas e atribuir rótulos com o k-means nos *Spectral Embeddings*[7]. O parâmetro *n_ components* foi definido como 2. Os vetores obtidos para $\{\mathbf{r}_i\}$ são referidos como $\{\mathbf{u}_i\}$.

3.6.1 *Cálculo das similaridades.* Von Luxburg [2007] discute diferentes possibilidades para a construção do grafo de similaridades. A opção deste trabalho foi a de usar grafos totalmente conectados e função de similaridade gaussiana, com a qual a similaridade cai exponencialmente com o quadrado da distância ($K_{ij} = e^{-\frac{d_{ij}^2}{2\sigma^2}}$). Essa escolha foi feita para evitar grafos com múltiplos componentes ou a atribuição do mesmo valor de similaridade para pares de pontos com distâncias muito diferentes. O parâmetro σ está relacionado à dispersão. Foi utilizado $\sigma = 1$, valor na mesma escala em que os componentes, ajustados para estar entre 0 e 1.

Após a reindexação feita para alinhar as séries na primeira ocorrência do valor máximo, o que se obtém são representações de tamanho fixo com índices inteiros em torno de zero e alguns valores faltantes. Eirola et al. [2013] discutem questões associadas à aplicação de métodos estatísticos e de aprendizado de máquina em conjuntos com essa característica. Os autores indicam que algumas das estratégias comuns, como a imputação ou o uso de distâncias parciais, tendem a subestimar tanto as incertezas dos valores imputados quanto as distâncias. Dizem ainda que em situações práticas de reconhecimento de padrões, o foco costuma ser em pontos com alta similaridade e, portanto, falsos positivos são um problema maior do que falsos negativos. Por isso, apontam para estratégias que, em vez de subestimar, tendam a sobrestimar as distâncias. Os autores prescrevem o cálculo da distância esperada utilizando um modelo baseado na distribuição normal multivariada.

Neste trabalho, também são usadas distâncias sobrestimadas que privilegiam pontos com alta similaridade, conforme sugerem Eirola et al. [2013]. Os desvios amostrais $\hat{s}_z$ de cada componente (calculados ignorando os valores faltantes) são usados para preencher os vetores de diferenças entre representações $|\mathbf{r}_i - \mathbf{r}_j|$ $(\forall i, j)$ nos índices não disponíveis em $\mathbf{r}_i$ ou em $\mathbf{r}_j$. Sendo $\mathbf{\Delta}(i, j)$ os vetores de diferença resultantes, usa-se $\delta_{ij} = \|\mathbf{\Delta}(i, j)\|_2 - \|\mathbf{\Delta}(i, i)\|_2$ (em particular, $\delta_{ii} = 0$) e $K_{ij} = e^{-\frac{\delta_{ij}^2}{2}}$ (função gaussiana com $\sigma = 1$) para obter uma matriz simétrica $\mathrm{W} = \{(\frac{K_{ij}+K_{ji}}{2})_{ij}\}$ com as similaridades.

3.6.2 *Escolha do número de clusters.* Halkidi et al. [2016] discutem o problema de determinar a quantidade de *clusters* diante do grande número de métodos disponíveis e de ajustes possíveis. Como não há rótulos de referência, a escolha de índices de qualidade é mais difícil do que em um problema de classificação supervisionada ou de regressão. Os autores afirmam ainda que o número de índices

[7] http://scikit-learn.org/stable/modules/generated/sklearn.manifold.SpectralEmbedding.html

disponíveis na literatura é relativamente grande e que é importante reconhecer que *clustering* não é um problema unicamente definido. Geralmente, não há apenas um particionamento ótimo para um certo conjunto de dados e as escolhas tanto do método quanto dos índices de validade dependem dos objetivos da pesquisa e de critérios do domínio do problema.

Neste trabalho, o índice adotado foi o *Average Silhouette Width* (ASW) (também discutido em Halkidi et al. [2016]). Esse critério opera com a agregação das silhuetas, calculadas com uma razão s_i em que são comparadas as distâncias médias de uma amostra com as demais do seu *cluster* (a_i) e com as do *cluster* mais próximo (b_i) dada uma atribuição de rótulos $\{c_i\}$ que considera um número de *clusters* k. O índice varia entre -1 (atribuição incorreta, valores próximos a -1 indicam $b_i \ll a_i$) e 1 (atribuição ótima, valores próximos a 1 indicam $a_i \ll b_i$). Foi usada a definição $d_{ij} = \|\mathbf{u}_i - \mathbf{u}_j\|_2$ ($\{\mathbf{u}_i\}$ são os *Spectral Embeddings*, veja a Seção 3.6). Para cada uma das execuções, foi escolhido o número de grupos (k) que maximiza o índice (Equação 2). Adicionalmente, foi considerado o balanceamento entre os tamanhos dos *clusters*. Os resultados podem ser vistos na Figura 5.

$$a_i = \frac{1}{n_{c_i} - 1} \sum_{j\ :\ c_j = c_i} d_{ij} \qquad b_i = \min_{l \neq c_i} \frac{1}{n_{c_l}} \sum_{j\ :\ c_j = l} d_{ij} \qquad s_i = \begin{cases} 1 - \frac{a_i}{b_i} & \text{se } a_i < b_i \\ 0 & \text{se } a_i = b_i \\ \frac{b_i}{a_i} - 1 & \text{c.c.} \end{cases} \qquad ASW = \frac{1}{n} \sum_i s_i \quad (2)$$

4. RESULTADOS

Após obter separadamente os *clusters* para cada habilidade principal dos jogadores em cada grupo de posições, foram produzidas as visualizações e constatou-se que haviam protótipos muito semelhantes ocorrendo em resultados diferentes (Figura 3). Tal observação motivou um novo particionamento: foi feita uma aplicação de *clustering* para categorizar os protótipos de cada um dos grupos originais em Padrões de Evolução. Para essa tarefa, foi adotado um algoritmo que só depende das distâncias par-a-par, a saber, *Agglomerative Clustering with Complete Linkage*[8]. O mesmo índice (ASW, Equação 2) foi usado para selecionar o k e o valor obtido foi *10*. Ou seja, foram identificados mais 10 padrões de evolução além do constante, sendo 11 o total. Os resultados podem ser visualizados na Figura 4. Nela, há também os percentuais de jogadores identificados com os respectivos Padrões de Evolução.

Os protótipos (Figura 4) foram obtidos a partir dos vetores de representação (Equação 1). As cores são utilizadas para indicar as ocorrências dos padrões de evolução (Figura 5). As curvas representam as medianas das pontuações normalizadas dos jogadores no *cluster* para cada índice. As espessuras das linhas indicam o percentual dos jogadores com pontuação disponível naquele índice.

A Figura 5 exibe a frequência de cada padrão para 20 habilidades descritas no jogo. Observe que há habilidades com poucos padrões, como *Ataque Posicionamento*, com 3 padrões apenas, e outras com mais, como *Ataque Pique*, com 8 padrões. No entanto, note que a frequência dos padrões está razoavelmente balanceada para cada habilidade e o padrão dominante varia de acordo com a habilidade. Particularmente, observe que em habilidades como *Finalização* e *Posicionamento*, o padrão mais frequente apresenta tendência de melhoria, indicando um efeito da maturidade. Para *Pique*, o mais frequente é de deterioração. No Ataque e no Meio-campo, os padrões para *Aceleração* e *Pique* são similares. E o mesmo ocorre entre habilidades dos goleiros. Também há diferenças

Fig. 3. Exemplos de protótipos semelhantes em posições e habilidades diferentes.

[8] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

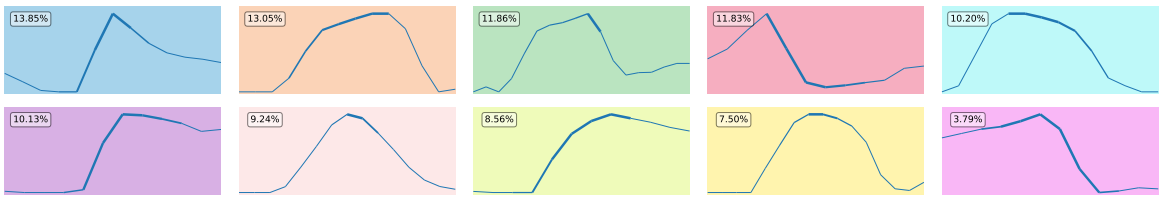

Fig. 4. Visualização dos 10 Padrões de Evolução obtidos, coloridos para identificar suas ocorrências (Figura 5).

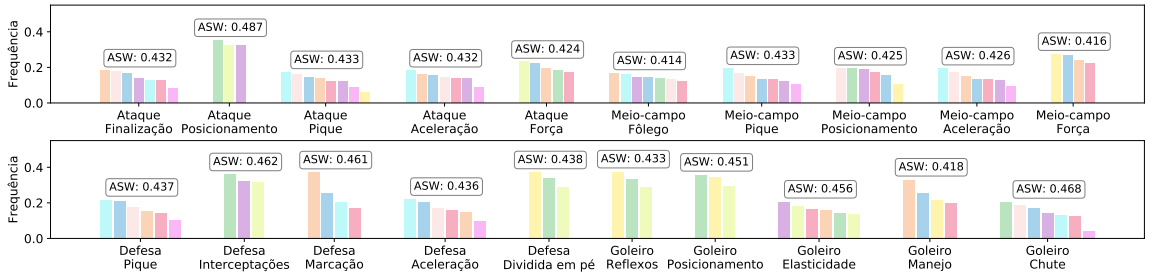

Fig. 5. Histogramas dos *clusters* para as habilidades de cada posição indicando seus Padrões de Evolução (Figura 4).

de diversidade de acordo com a habilidade considerada, vistas pelos diferentes números de padrões identificados.

Disponibilizamos exemplos de como os Padrões de Evolução aparacem nas séries de pontuações dos atletas. Foram selecionados jogadores com padrões com característica dominante de melhoria (Figura 6), de deterioração (Figura 7) e com melhoria seguida de deterioração (Figura 8). Em tons de azul, são apresentados os jogadores de topo, ordenados pela mediana do atributo *Overall*. Em tons de vermelho, são apresentados outros jogadores dos grupos, independentemente de sua pontuação geral.

5. CONCLUSÃO E TRABALHOS FUTUROS

Neste artigo, propomos uma metodologia para identificar padrões de evolução nas habilidades de atletas de elite do futebol a partir dos dados do FIFA. Em suma, a técnica normaliza as pontuações, as alinha considerando os valores máximos, define uma similaridade e utiliza algoritmos de *clustering*. Surpreendentemente, foram encontradas evidências de que as curvas de desenvolvimento das habilidades podem ser caracterizadas por apenas 11 padrões, sendo um deles o constante. As escolhas dos números de *clusters* foram feitas com o índice ASW, que apresentou resultados entre 0.414 e 0.487, indicando que as atribuições de rótulos representam agrupamentos com uma separação adequada. Os percentuais de jogadores em cada *cluster* são razoavelmente balanceados. Com uma análise qualitativa dos resultados, vimos que os padrões de evolução ocorrem em múltiplos grupos de posições e em habilidades diferentes: são visões distintas do conjunto de dados sendo processadas separadamente e

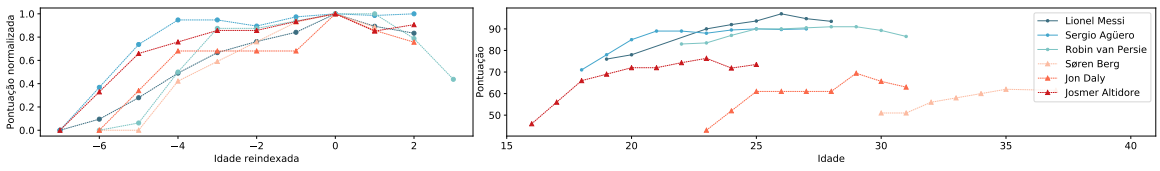

Fig. 6. Amostras de um agrupamento com tendência de melhoria. As pontuações são para a habilidade de *Finalização* dos jogadores de Ataque. O agrupamento é o mais frequente. Esquerda: Representação. Direita: Série original.

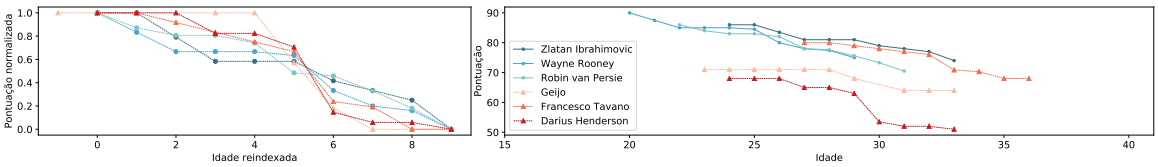

Fig. 7. Amostras de um agrupamento com tendência de deterioração. As pontuações são para a habilidade de *Pique* dos jogadores de Ataque. O agrupamento é o mais frequente. Esquerda: Representação. Direita: Série original.

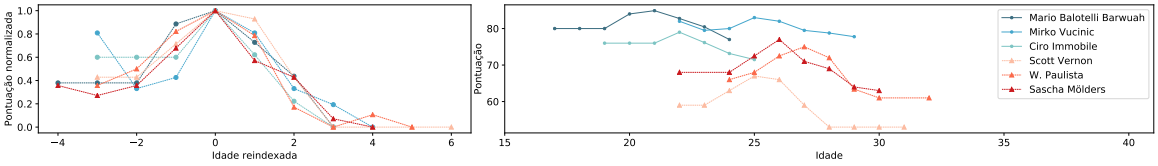

Fig. 8. Amostras de um agrupamento com tendência de melhoria seguida de deterioração. As pontuações são para a habilidade de *Aceleração* dos jogadores de Ataque. O agrupamento é o quinto em ordem de frequência.

resultando em protótipos similares. Também foram discutidas algumas intuições obtidas a partir dos agrupamentos e apresentados exemplos de como os padrões aparecem nas séries originais.

Há várias possíveis direções para estender este trabalho, incluindo: estudar a adição de outras informações sobre os atletas nas representações, tais como características físicas e mudanças de clube; investigar abordagens para incluir os jogadores com história curta nas análises, por exemplo, obtendo dados de outras fontes sobre o tempo de carreira deles; explorar outras possíveis decisões metodológicas, como modelar as estruturas de correlação das séries, inclusive entre séries de habilidades diferentes; e propor novas formas de avaliação e usos dos padrões de evolução em tomadas de decisão.

REFERENCES

Akhanli, S. E. and Hennig, C. Some issues in distance construction for football players performance data. *Archives of Data Science* 2 (1), 2017.

Caiado, J., Maharaj, E., and D'urso, P. Time series clustering. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.), *Handbook of Cluster Analysis*. CRC Press, USA, pp. 241–263, 2016.

Cotta, L., de Melo, P. O. V., Benevenuto, F., and Loureiro, A. A. Using fifa soccer video game data for soccer analytics. In *Proceedings of the KDD Workshop on Large-Scale Sports Analytics*. San Francisco, USA, 2016.

Eirola, E., Doquire, G., Verleysen, M., and Lendasse, A. Distance estimation in numerical data sets with missing values. *Information Sciences* vol. 240, pp. 115–128, 2013.

Halkidi, M., Vazirgiannis, M., and Hennig, C. Method-independent indices for cluster validation and estimating the number of clusters. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.), *Handbook of Cluster Analysis*. CRC Press, USA, pp. 595–618, 2016.

McFee, B. and Ellis, D. Analyzing song structure with spectral clustering. In *ISMIR*. pp. 405–410, 2014.

Meila, M. Spectral Clustering. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.), *Handbook of Cluster Analysis*. CRC Press, USA, pp. 147–164, 2016.

Paparrizos, J. and Gravano, L. Fast and accurate time-series clustering. *ACM Transactions on Database Systems* 42 (2): 8, 2017.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5 (1): 31, 2016.

Sardá-Espinosa, A. Comparing time-series clustering algorithms in r using the dtwclust package. https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf, 2017.

Soto-Valero, C. A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. *RICYDE. Revista Internacional de Ciencias del Deporte* 13 (49): 244–259, 2017.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing* 17 (4): 395–416, 2007.

Vroonen, R., Decroos, T., Van Haaren, J., and Davis, J. Predicting the potential of professional soccer players. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*. Skopje, Macedonia, 2017.