A momentum quantification method for tennis matches based on t-SNE and k means clustering

Wenhao Shen[1], Runchen Zhao[1], Tianqi Bu [2], Yuhan Wu[3], Yanmei Jiao[4,5]

[1]School of Civil Engineering, Nanjing Tech University, Nanjing 211800, China.
[2]School of Mechanical and Power Engineering, Nanjing Tech University, Nanjing 211800, China
[3]College of Chemical Engineering, Nanjing Tech University 211800, Nanjing, China.
[4]School of Physical and Mathematical Sciences, Nanjing Tech University, Nanjing 211800, China.

[5]jiaoym@njtech.edu.cn

Abstract. Momentum is a crucial factor influencing the outcomes of sports competitions. This study proposes a method for visualising and quantifying momentum using various modern analytical techniques. These techniques include feature engineering, dimensionality reduction with the t-SNE algorithm, and k-means clustering. By analysing data from the past five years of Wimbledon, the study extracted various performance features such as serve success rate and scoring rate. By calculating the similarities between features and conducting cluster analysis, the study reveals changes in momentum during matches. This research employs data visualisation and clustering results to not only display the conditions of matches under different momentum states but also quantifies momentum changes by calculating the Euclidean distance between data points and cluster centroids and standardising these distances using the Z-score method. This approach provides a visual perspective that reveals the dynamic changes that occur within matches and quantifies momentum. It demonstrates that these changes in momentum can indeed reflect the progression and outcomes of matches.

Keywords: Momentum, t-SNE algorithm, K-means clustering

1. Introduction

In sports competition, momentum is considered a key force that can bring about disruptive changes during a match, especially in tense and highly competitive circumstances. The creation of momentum is often based on consecutive scores, an outstanding performance, or a psychological advantage over the opponent. Understanding and mastering momentum can help coaches and athletes' better strategies and gain advantageous positions during matches. The 2023 Wimbledon tennis championship final between Carlos Alcaraz and Novak Djokovic is a classic example of how momentum can change the outcome of a match. Despite Djokovic's initial dominance [1], Alcaraz turned the tide by winning several key games in rapid succession. Not only did Alcaraz maintain this momentum, but he also showed incredible determination and skill throughout the match.

Momentum may be a more psychological concept than scientific. A positive momentum is generally thought of as a psychological or competitive advantage that is gradually built up by a side over the

DOI: 10.54254/2753-8818/39/20240611

course of a game as a result of successive successful moves [2]. The study of momentum and its impact on the outcomes of sporting competitions is a complex and challenging subject. The value of momentum is contingent upon the concept of predefined key moments, such as break points, game points, set points, and match points [3]. These situations are influenced by the relative strength of the players and the events that have already occurred in the match, which can result in changes in force and shifts in control.

Although the existence of momentum is widely accepted, research on its quantification and specific impact on matches is still in its early stages. Traditional statistical data, such as scores, dominance duration, and possession rates, can provide some clues about momentum, but often fail to fully reflect the dynamic changes and psychological impacts of momentum. Tennis matches are usually intense and long, with many factors on the court that are not observable yet affect the progress of the game. Nevertheless, it is of paramount importance to closely monitor the fluctuations in the score and the potential state of the players, as this can provide profound insights for players and coaches, enabling them to devise more effective game strategies. Quantifying momentum in matches is a challenging but significant endeavour, resulting from the interplay of various influencing factors in specific scenarios. These factors can also interact with each other, rather than existing independently.[4]

Therefore, we cannot simply view momentum as a linear combination of other features. Previous research on momentum has mainly remained at the superficial level, including individual analysis [5] and simple effects [6]. Modern analytical techniques are being used to study and quantify momentum more deeply to predict match outcomes more accurately and intuitively [7]. This study proposes a detailed framework through a mathematical model to assess the momentum in tennis matches, aiming to objectively measure the changes in players' momentum. Using feature engineering, it identifies key indicators such as scoring differences, game wins, set advantages, and service efficiency for analysis. The t-SNE algorithm reduces the dimensionality of the data, followed by k-means clustering. The distance between sample points and cluster centres is quantified using the z-score method, providing a precise measure of momentum. Analysing the initial match dataset revealed the dynamic interplay between scoring and momentum.

2. Methodology

We utilized information from the Wimbledon Tennis Championships over the last 5 years, including a variety of statistics from different tournaments. After processing, we extracted several features that describe different aspects of performance during the match. These characteristics include first serve, number of serves, number of serves, service success rate, average serve speed, points won, point success rate, break success rate, unforced errors, error count, net score rate, movement, average rebound count, cumulative points lost, and cumulative points won, etc. There are more than ten attributes, with a total of 1190 pieces of information. We selected these features based on their importance in reflecting player performance and match momentum, considering factors such as player performance on serves, points scored and errors. To calculate the value of each feature, we also used a dictionary data structure in Python to store the state of each player and imported the selected features into it. At each point in the game, we continuously update these states and use them for feature calculations. This approach allows us to accurately track player performance during different phases of the game. For example, the cumulative distance covered is calculated in a cumulative manner, which can provide some insight into the player's physical condition. We stored each feature using a dictionary data structure in Python. To better understand the player's momentum changes during the game, we used the t-SNE algorithm for feature dimensionality reduction. t-SNE [8] is a non-linear dimensionality reduction technique that maps high-dimensional data to a low-dimensional space. In our study, we chose to reduce the features to three dimensions to allow for subsequent visualization and cluster analysis. Firstly, we calculate the similarity between each pair of data points, where we utilize a Gaussian kernel function to measure similarity in equation 1:

$$p_{ij} = \frac{\exp\left(-\frac{|x_i - x_j|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{|x_i - x_k|^2}{2\sigma_i^2}\right)} \tag{1}$$

In the equation 1, p_{ij} represents the conditional probability between data points x_i, and x_j, while σ_i is the bandwidth parameter used to determine the similarity. Then, similarity construction in low-dimensional space is done. In the low dimensional space, we calculate the conditional probability q_{ij} between data points, typically represented by the t-distribution in equation 2:

$$q_{ij} = \frac{\left(1 + |y_i - y_j|^2\right)^{-1}}{\sum_{k \neq l}(1 + |y_k - y_l|^2)^{-1}} \tag{2}$$

With y_i and y_j denote the mappings of data points x_i and xj in the low dimensional space, while q_{ij} represents the conditional probability between them. Next, we aim to minimize the KL divergence between the conditional probability distributions in the high-dimensional and low-dimensional spaces, ensuring that data points maintain their similarity relationships in the low-dimensional space:

$$KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{3}$$

Next, we minimize this KL divergence through optimization algorithms such as gradient descent. we employ the gradient descent optimization algorithm to minimize the KL divergence and adjust the positions of data points in the low-dimensional space denoted as y_i to obtain the optimal low-dimensional representation.

$$\frac{\partial KL}{\partial y_i} = 2\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + |y_i - y_j|^2\right)^{-1} \tag{4}$$

In the reduced-dimensional feature space, we utilize the K-means clustering algorithm to identify the momentum shifts at different stages of the match. we opt for the K-means algorithm because it allows us to partition the data into distinct clusters, each representing a different momentum state. Through cluster analysis, we investigate the shifts in momentum during different phases of the match and identify stages with stronger momentum. Here, we haven data points 1, 2..., n, x1, x2 ..., then randomly select k data points as the initial cluster centres, denoted as 1, 2, ..., c1, c2, ... For each data point xi, calculate its distance to each cluster centre and assign it to the cluster associated with the nearest cluster centre. The distance measurement commonly used is the Euclidean Distance, where d is the dimension of the data point.

$$d(x_i, c_j) = \sqrt{\sum_{p=1}^{d} \left(x_{ip} - c_{jp}\right)^2} \tag{5}$$

For each cluster j, recalculate its cluster centre c_j as the mean of all data points within the cluster, where sj represents the set of all data points belonging to cluster j.

$$c_j = \frac{1}{|s_j|} \sum_{x_i \in s_j} x_i \tag{6}$$

Then, repeat the assignment step and update step until the stopping criteria are met, typically when the cluster centres no longer change or after reaching the maximum number of iterations. And last, output the Results of the final cluster assignments. The objective of the K-Means algorithm is to minimize the sum of squared distances be- tween data points within their cluster and their respective cluster centres. In other words, it aims to minimize the following objective function, where j represents the objective function, and s_j denotes all the data points within cluster j.

$$J(c_1, c_2, \dots, c_k) = \sum_{j=1}^{k} \sum_{x_i \in s_j} |x_i - c_j|^2 \tag{7}$$

3. Analysis

The visualization diagram after dimensionality reduction by the t-SNE algorithm is shown in Figure 1. The dimensionally reduced data is presented in two clusters of different colours, which may represent two different momentum states or categories. One colour represents the player with a greater advantage. or controlling phases of the game, while another colour might represent a relatively small advantage or a time when there is less momentum.

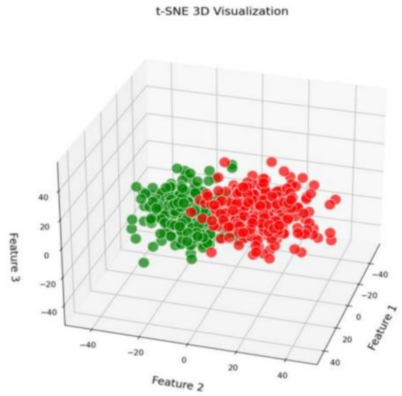
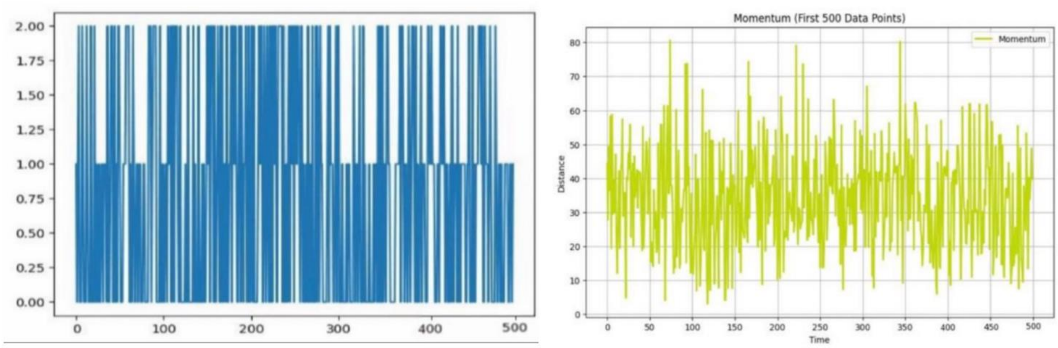


Figure 1. t-SNE 3D Visualization

After the analysis of polymorphism is completed, we calculate each polymorphism's characteristics. The clustering results may not necessarily represent momentum changes. Here, we use the distance between data points and k-means cluster centroids to signify momentum. To quantify this distance, we employ the Euclidean distance to calculate the distance between each record and the nearest centroid. To ensure that the distances are on the same scale and to eliminate the influence of differences in distance values, we apply Z-score standardization to the distances. This way, we can compare distances between different records without being affected by the magnitude of distance values. on the right of Figure 2 is a visualization of momentum states in a portion of the data. Momentum quantification has been completed, and to better observe the trends in momentum and score flow, we will now select specific matches for analysis.



(a) (b)

Figure 2. k-means scatter plot figure (a) and momentum change chart(b).

Here, we will select the first match and conduct a detailed comparison of the match con- tent. In the graph, negative momentum values indicate an advantage for player 1, while positive momentum values indicate an advantage for player 2. This will help us gain a clearer under- standing of the progress and outcome of the match.

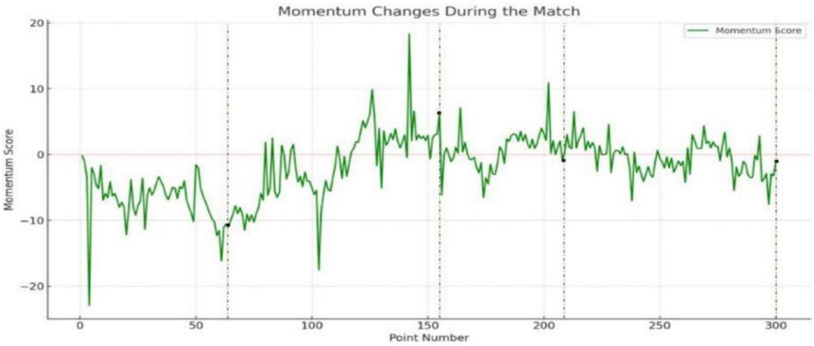
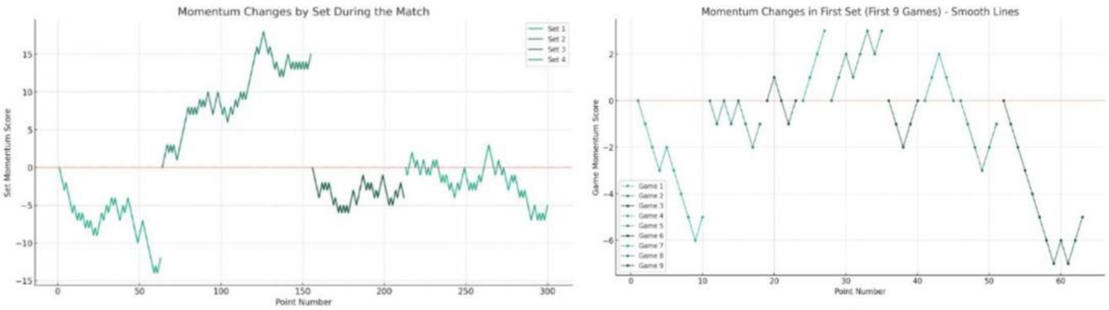


Figure 3. Momentum change chart of the first match.

The dividing line in the graph delineates the boundaries between different phases of the first match. Let’s review what transpired during the actual match: In the first set, player 1, carlos Alcaraz, initially took control of the match, securing the first set with a score of 6-3. In the second set, carlos Alcaraz maintained his lead, starting with a 3- 1 advantage. How- ever, player 2, Nicolas Jarry, mounted a strong comeback, leading to a tiebreaker, where Nico- las Jarry ultimately secured the valuable set, winning 7-6. In the third set, Nicolas Jarry, who had made a comeback in the previous set, couldn’t sustain his momentum. He started with a 4- 1 deficit and eventually lost the set 6-3. Moving on to the fourth set, carlos Alcaraz took control of the match right from the start, establishing a 4- 1 lead. Despite Nicolas Jarry’s efforts to stage a comeback and win three consecutive games, carlos Alcaraz ultimately secured victory with a score of 7-5. when comparing the actual match situation with the momentum flowchart above, it be- comes evident that the outcomes of each set remain entirely consistent. Additionally, for a clearer observation of the specific momentum changes within each set and the momentum and score flow variations within the first nine games of the first set, we have magnified and segmented the data as follows in Figure 4. It can be observed that in the first, third, and fourth sets of the match, carlos Alcaraz consistently performed as the stronger player, which ultimately led to his victory in the match. In the second set, Nicolas Jarry’s comeback added significantly to his momentum and made him the standout player in that set. Evidently, there is a high degree of alignment between the actual match outcomes and the changes in momentum. This validates the correctness of the momentum measurement method and the fact that momentum changes can reflect the progression of the match.



(a) (b)

Figure 4. Momentum chart for each set(a) and the momentum chart for games(b)

It should be noted that we use structured data, so there may be some information loss. Nowadays, the development of artificial intelligence, especially the development of visual technology [9], has given room for updating analysis technology. Therefore, we will consider adding visual factors in future modelling, such as using the evolving YOLO [10,11] technology to identify players' expressions after rounds and directly use them as features or directly performing information fusion [12]. In addition to the inclusion of additional information in future research. Some more machine learning algorithms will be added to the momentum modelling of this work to further improve the visualization and interpretability of the model [13].

4. Conclusion

This study proposes a mathematical model capable of quantifying changes in momentum during tennis matches. The model has been successfully validated in revealing the close connection between momentum and match outcomes. After feature processing of the data, we established a set of paradigms for quantifying momentum through dimensionality reduction using the t-SNE algorithm and k-means cluster analysis. The enhanced visual expression of momentum in the data, in conjunction with the ability to identify and track momentum changes during a race, provides a more accurate representation of the underlying dynamics. The momentum quantification method proposed in this study offers a novel perspective for coaches and athletes, providing insights into the dynamics of power transfer and control during the game. This, in turn, facilitates the development of more effective game strategies. Future research may further enhance this model framework by integrating additional data types and refining the machine learning model.

References

[1] Al Jazeera. (2023, July 16). Wimbledon 2023 final live: Novak Djokovic faces Carlos Alcaraz. Retrieved from https://www.aljazeera.com/sports/liveblog/2023/7/16/wimbledon-2023-final-live-novak-djokovic-faces-carlos-alcaraz.

[2] Gray, W. R., & Vogel, J. R. (2016). Quantitative Momentum: A practitioner's guide to building a momentum-based stock selection system. John Wiley & Sons.

[3] Moss, B., & O'Donoghue, P. (2015). Momentum in US Open men's singles tennis. International Journal of Performance Analysis in Sport, 15(3), 884-896.

[4] Seidl, R., & Lucey, P. Live Counter-Factual Analysis in Women's Tennis using Automatic Key-Moment Detection.

[5] Meier, P., Flepp, R., Ruedisser, M., & Franck, E. (2020). Separating psychological momentum from strategic momentum: Evidence from men's professional tennis. Journal of economic psychology.

[6] Gauriot, R., & Page, L. (2019). Does success breed success? A quasi-experiment on strategic momentum in dynamic contests. The Economic Journal, 129(624), 3107-3136.

[7] Shen, X., Luo, S., & Zhang, M. (2023). House quality index construction and rent prediction in New York City with interactive visualization and product design. Computational Statistics, 38(4), 1629-1641.

[8] Shi, S. (2021). Visualizing data using GTSNE. arXiv preprint arXiv:2108.01301.

[9] Dai, W., Mou, C., Wu, J., & Ye, X. (2023, May). Diabetic Retinopathy Detection with Enhanced Vision Transformers: The Twins-PCPVT Solution. In 2023 IEEE 3rd International Conference on Electronic Technology, Communication, and Information (ICETCI) (pp. 403-407). IEEE.

[10] Qi, Z., Ma, D., Xu, J., Xiang, A., & Qu, H. (2024). Improved YOLOv5 Based on Attention Mechanism and FasterNet for Foreign Object Detection on Railway and Airway tracks. arXiv preprint arXiv:2403.08499

[11] Ma, D., Li, S., Dang, B., Zang, H., & Dong, X. (2024). Fostc3net: A Lightweight YOLOv5 Based on the Network Structure Optimization. arXiv preprint arXiv:2403.13703.

[12] Ao, X., Qi, Z., Wang, H., Yang, Q., & Ma, D. (2024). A Multimodal Fusion Network for Student Emotion Recognition Based on Transformer and Tensor Product. arXiv preprint arXiv:2403.08511\

DOI: 10.54254/2753-8818/39/20240611

[13] Dai, W., Jiang, Y., Mou, C., & Zhang, C. (2023, September). An Integrative Paradigm for Enhanced Stroke Prediction: Synergizing XGBoost and xDeepFM Algorithms. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 28-32).