UDC 681.51:519.876.5 Ambartsumian S.[1], Shevchenko Y.[2]

TIME-SERIES CATEGORICAL DATA CLUSTERING

Topic relevance. Clustering is a widely used technique in data analysis, primarily applied to numerical datasets where data points can be compared using distance metrics. However, applying traditional clustering methods to categorical and time-series data presents several challenges.

One major issue is the **lack of a natural distance metric**. Numerical clustering algorithms, such as k-means, rely on Euclidean distance to group similar points, but categorical values like character names or roles do not have an inherent numerical relationship. Methods like k-modes adapt k-means by using the mode instead of centroids, but they still fail to capture sequence patterns over time

Another problem is the **temporal nature of the data**. Standard clustering techniques group data based on static attributes, whereas time-series clustering must account for evolving trends. For instance, in gaming analytics, a player's character choices change over time, requiring an algorithm that detects shifts rather than fixed categories

Hierarchical clustering offer potential solutions, but they are computationally expensive and not optimized for large datasets with categorical values.

Therefore, a more efficient approach is needed to handle categorical time-series data while maintaining computational feasibility.

Problem investigation. To address these challenges, we decided to create a **sequential clustering algorithm** which designed specifically for categorical time-series data.

This algorithm groups data points based on evolving trends rather than static similarities, making it well-suited for applications such as player behavior analysis in games.

Key Parameters and Settings*:*

1) *Recent Games Window* (recent_games_window):

- defines the number of past matches considered when determining a player's dominant playstyle;

- a larger window smooths short-term fluctuations but may delay

[1] Student at the Department of System Analysis and Control, Dnipro University of Technology, Dnipro, Ukraine

[2] Assistant Lecturer at the Department of System Analysis and Control, Dnipro University of Technology, Dnipro, Ukraine

detection of playstyle shifts.

2) *Hero Dominance Threshold* (hero_dominance_threshold):

- represents the minimum proportion of recent matches where a hero must appear to be considered dominant;
- for example, if set to 0.5, the hero must be used in at least 50% of the recent games to trigger a cluster change.

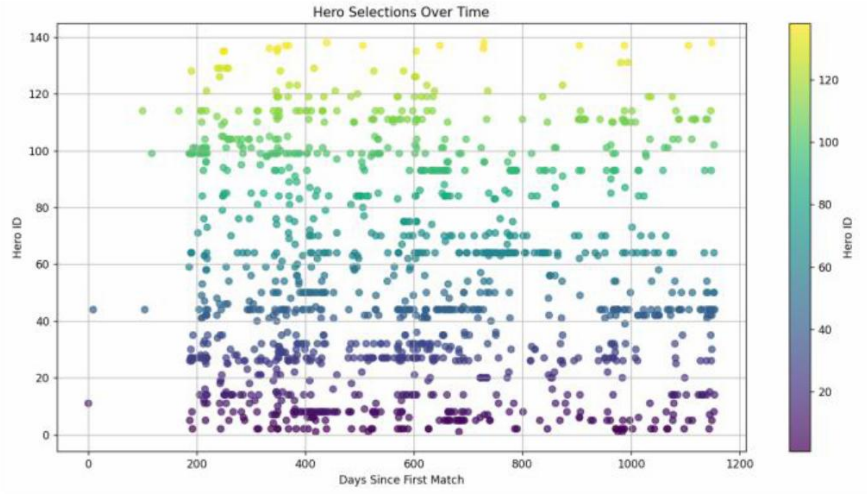
3) *Minimum Cluster Size* (min_cluster_size):

- prevents the formation of overly small clusters that may result from brief variations in hero selection;
- ensures clusters represent stable playstyle trends rather than one-off anomalies.

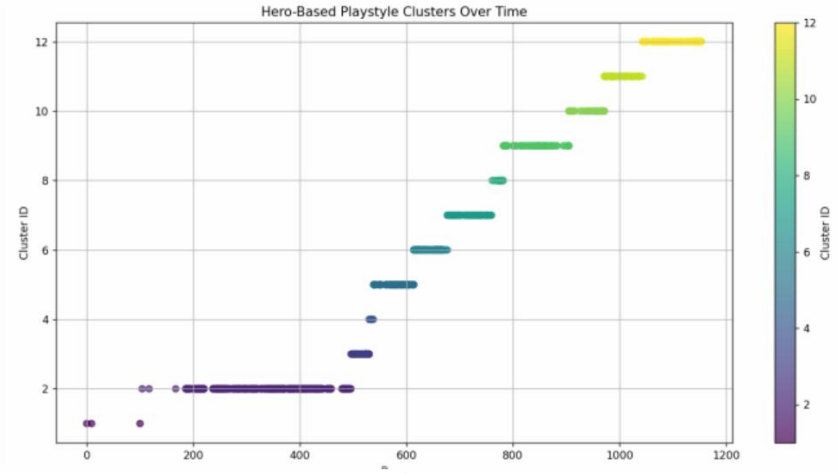
*How the **Algorithm for Categorical Time-Series Clustering** works:*

1. **Reprocessing:** Convert timestamps into a numerical format (days since first match) and sort matches chronologically.
2. **Tracking Hero Trends**: Maintain a sliding window of recent_games_window matches to detect dominant heroes.
3. **Cluster Formation**: If a hero dominates within this window beyond hero_dominance_threshold, a new cluster is created.
4. **Cluster Refinement**: Small clusters are merged or discarded based on min_cluster_size to ensure meaningful segmentation.

On the Figure 1 (*a,b,c*) you can example of usage algorithm, where 1315 games; recent_games_window = 10; hero_dominance_threshold = 0.5; min_cluster_size = 0.



a)



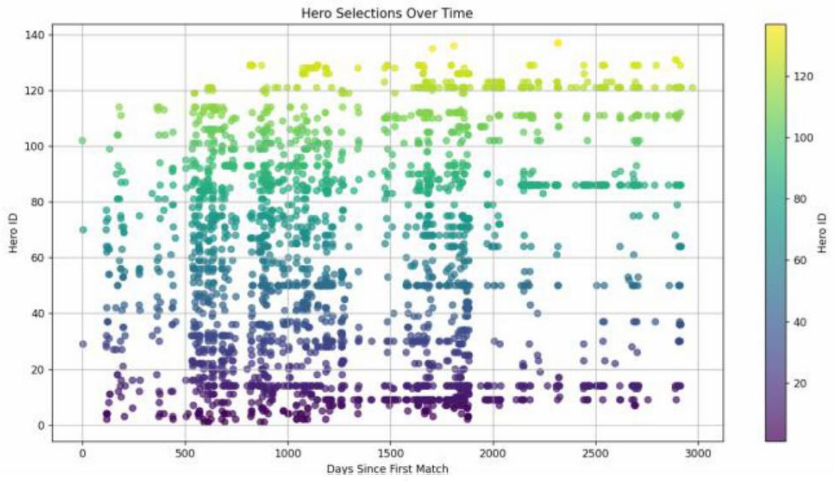
b)

	Cluster Number	Total Games	Most Popular Hero 1	Hero 1 Percentage
0	1	3	11.0	33.33%
1	2	520	8.0	7.12%
2	3	57	27.0	15.79%
3	4	10	99.0	50.00%
4	5	148	27.0	20.27%
5	6	113	44.0	13.27%
6	7	109	64.0	16.51%
7	8	26	64.0	30.77%
8	9	89	93.0	15.73%
9	10	63	111.0	14.29%
10	11	72	44.0	16.67%
11	12	104	44.0	17.31%

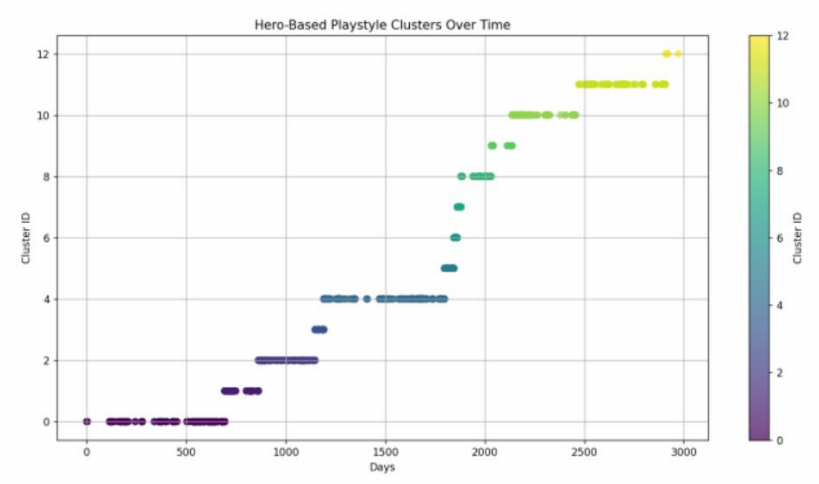
c)

Figure 1 - Example of usage algorithm with 1315 games

On the Figure 2 (*a, b, c*) you can example of usage algorithm with 3291 games; recent_games_window = 20; hero_dominance_threshold = 0.5; min_cluster_size = 20.



a)



b)

	Cluster Number	Total Games	Most Popular Hero 1	Hero 1 Percentage
0	0	572	32.0	5.07%
1	1	228	93.0	31.14%
2	2	573	50.0	6.98%
3	3	132	50.0	44.70%
4	4	710	9.0	21.13%
5	5	147	64.0	10.88%
6	6	122	9.0	38.52%
7	7	123	50.0	20.33%
8	8	124	9.0	20.16%
9	9	58	121.0	60.34%
10	10	200	86.0	40.50%
11	11	218	86.0	25.69%
12	12	26	121.0	50.00%

c)

Figure 2 - Example of usage algorithm with 3291 games

Practical significance. Unlike traditional clustering methods, it dynamically adapts to changes in categorical sequences, making it a practical solution for time-series categorical data processing.

REFERENCES

1. Aghabozorgi S., Shirkhorshidi A., Wah T. Time-series clustering. A DecadeReview Information systems. 2015. Vol. 53. P. 16-38.

2. Pandas : веб-сайт. URL: https://pandas.pydata.org/docs/ (дата звернення: 25.02.2025).

3. Matplotlib : веб-сайт. URL: https://matplotlib.org/stable/users/index.html (дата звернення: 25.02.2025).

4. Scikit-learn : веб-сайт. URL: https://scikit-learn.org/stable/user_guide.html (дата звернення: 25.02.2025).

5. Коряшкіна Л. С., Станіна О. Д., Шевченко Ю. О. Практикум з диференційних рівнянь. 2024. URL: