Discovering similar multidimensional trajectories

Ritu Patel, Sarthak Siddhpura, Moin Vinchhi, Vrunda Patel, Vishv Boda
Ahmedabad University, Ahmedabad, Gujarat 380009, India

I. Research Contributions by Paper

- Introduced novel non-metric similarity functions ($S1$ and $S2$) based on the Longest Common Subsequence (LCSS) model, designed to measure similarity between multidimensional trajectories, robust to noise, varying speeds, and spatial translations.
- Developed an efficient exact algorithm to compute the $S2$ similarity function in $O((n+m)^3\delta^3)$ time, accounting for translations in 2D or 3D space, where n and m are trajectory lengths and δ is the time-stretching parameter.
- Proposed a faster approximate algorithm for $S2$ with a bounded error β, achieving a time complexity of $O((n+m)\delta^3/\beta^2)$, enhancing scalability for large trajectory datasets.
- Established a weaker form of the triangle inequality for the LCSS-based distance function ($D2$), enabling pruning in an indexing structure for nearest neighbor queries despite its non-metric nature.
- Designed an indexing scheme using hierarchical clustering with medoids, optimizing similarity retrieval for trajectories of varying lengths, tested on real datasets like SEALS and ASL.
- Demonstrated superior clustering performance of LCSS over Euclidean and Dynamic Time Warping (DTW) distances under noisy conditions, validated through experiments on video tracking and sign language datasets.

II. Key Insights From Paper

The paper by Vlachos et al. offers valuable insights for our research direction of identifying abnormal driving behavior using trajectory datasets, leveraging probabilistic and statistical methods. The introduction of the LCSS-based similarity functions ($S1$ and $S2$) is particularly relevant, as they address key challenges in trajectory analysis—noise, varying sampling rates, and spatial offsets—that are prevalent in real-world driving data, such as GPS traces from vehicles on a road patch. Our assumption that drivers behave similarly on the same road segment aligns with the LCSS's ability to focus on common subsequences, ignoring outliers like sudden stops or sensor errors. For instance, the $S1$ function, which allows time stretching within a parameter δ, can match trajectories of drivers moving at different speeds but following similar paths, enabling us to cluster normal driving patterns. The $S2$ function extends this by incorporating spatial translations, crucial for handling parallel movements (e.g., lane changes) that might occur across a road patch. This robustness to noise and flexibility in alignment could serve as a preprocessing step to group trajectories into a "normal" cluster, from which a binary classifier could then identify anomalies based on deviations in spatial-temporal features like position or velocity.

A significant insight is the paper's approach to handling noisy and variable-length trajectories, which directly supports our goal of developing a robust binary classifier. The LCSS model's ability to exclude dissimilar portions (e.g., erratic maneuvers or GPS glitches) ensures that clustering focuses on consistent driving behavior, such as steady lane-following or typical speed profiles. The experimental results, particularly with the SEALS dataset (marine animal trajectories) and noisy ASL data, highlight LCSS's superiority over Euclidean and DTW distances in noisy conditions, achieving up to 14 correct clusterings versus 5 and 7, respectively, in the noisy ASL experiment. For our project, this suggests that LCSS could effectively cluster trajectories from a road patch—despite inconsistencies in sampling or outliers from aggressive driving—providing a statistical foundation for normal behavior. The approximate algorithm for $S2$, with its $O((n+m)\delta^3/\beta^2)$ complexity, offers a practical trade-off between accuracy and computation time, enabling scalability for large datasets like urban traffic logs. By defining normalcy as trajectories with high LCSS similarity to cluster medoids, we could use probabilistic methods (e.g., assigning anomaly scores based on distance thresholds) to classify trajectories as normal or abnormal, aligning with our prob-stat focus.

Furthermore, the indexing structure and weaker triangle inequality provide a framework for efficient nearest neighbor queries, which could enhance our anomaly detection pipeline. The hierarchical clustering approach, partitioning trajectories by length and using medoids, suggests a way to organize driving data spatially and temporally within a road patch, facilitating quick retrieval of similar trajectories. This is critical for real-time applications, where we might need to compare a new trajectory against a database of normal behaviors rapidly. The weaker triangle inequality allows pruning of dissimilar clusters, reducing computational overhead when searching for anomalies. For our binary classifier, we could adapt this by first clustering trajectories with LCSS, then training a model (e.g., logistic regression or SVM) on features derived from LCSS distances to medoids, such as the proportion of matched points or translation parameters (c, d). The paper's emphasis on handling multidimensional data (2D or 3D) also fits our use of spatial-temporal features, enabling us to incorporate both position (x, y) and time into the analysis, capturing the full context of driving behavior. This comprehensive approach—combining robust similarity measurement, efficient computation, and indexing—positions the LCSS model as a powerful tool for our research objectives.

III. Unaddressed Issues and Assumptions

- **Unaddressed Issues:**
  - The paper does not explore integration with supervised learning techniques, such as binary classification, limiting its direct applicability to distinguishing normal versus abnormal trajectories.
  - Real-time processing capabilities are not evaluated, which is crucial for dynamic anomaly detection in driving scenarios.
  - The impact of contextual factors (e.g., traffic conditions, road type) on similarity measures is not addressed, potentially affecting clustering accuracy for driving data.
  - The scalability of the indexing structure for very large datasets (e.g., millions of trajectories) remains untested, a concern for urban traffic analysis.
- **Assumptions Made:**
  - Assumes trajectories have sufficient common subsequences to define similarity, which may not hold if driving behaviors are highly erratic or diverse on a road patch.
  - Assumes noise is sporadic and can be ignored by LCSS, potentially overlooking systematic errors in GPS data that affect all points.
  - Assumes the parameters δ and ϵ can be universally tuned, whereas optimal values may vary significantly across different road patches or driving contexts.
  - Assumes translations alone suffice for spatial alignment, neglecting other transformations (e.g., rotation, scaling) that might occur in complex driving maneuvers.

IV. Motivation Behind Choosing This Paper

The motivation for selecting this paper stems from its innovative LCSS-based similarity measures, which directly address the challenges of analyzing noisy, multidimensional trajectory data—a core requirement for our research on detecting abnormal driving behavior. The robustness of LCSS to noise and its flexibility with varying speeds and spatial offsets make it an ideal candidate for clustering vehicle trajectories on a road patch, supporting our assumption of similar normal behaviors with potential anomalies. The paper's focus on efficient computation and indexing aligns with the need to process large-scale driving datasets, while its probabilistic underpinnings (e.g., approximate algorithms, clustering tendencies) complement our prob-stat approach. By providing a method to define normal driving patterns statistically, it lays the groundwork for a binary classifier to identify anomalies, making it a compelling choice to advance our project's objectives.

References