A Symbolic Representation for Trajectory Data

Nguyen Huy Thach[*1] Einoshin Suzuki[*2]

[*1] thachnh@i.kyushu-u.ac.jp [*2] suzuki@inf.kyushu-u.ac.jp

In this paper, we propose a novel symbolic representation for trajectory data. A symbolic representation allows us to represent the original data into smaller and less complex symbol components that are beneficial for storage and computation. Previous works on trajectory representation only focus on trajectory shapes and ignore the position features. Therefore, they may fail to distinguish trajectories that have similar shapes but different positions, which are necessary in many trajectory data mining applications such as delivery, transportation and weather forecast. In this work, the positions of trajectories are considered and play important roles in our representation. Experimental results on four real data sets have shown the effectiveness of our method.

1. Introduction

With the improvement of satellite, GIS, RFID, sensor and wireless technologies, trajectory data has appeared in various real world applications. For example, in Starkey Project [Starkey], an automated radio telemetry system can automatically track locations of deers, elks and cattle to generate trajectories of these species. In weather forecast, the path of a hurricane can be tracked by using satellite photos and radar [Hurricane]. In transportation, tracks of buses and trucks can be collected via a mobile device or a wireless network infrastructure [Rtreeportal]. More generally, if there is a moving object, there will have a trajectory data, thus trajectory data occurs virtually in every science, transportation and zoology. Therefore, there is a need for analyzing trajectory data effectively to extract useful information.

Although there is a dramatical improvement of computing and storage capability, the rapid increase of data volume is still a challenge for the performance of computer networks and internet systems, which necessitates to design an efficient representation and an efficient storage mechanism. In this work, we tackle the problem of trajectory representation where almost all trajectories are long and complex. Despite the fact that there is an explosion of representations in time series data, which are a series of one-dimensional real numbers, there are few representations of trajectory data [Shatkay 95], [Vlachos 02], [Lee 07]. This may be surprising since trajectory data can be considered as a multi-dimensional time series data. A possible reason is that working with trajectory data is more complex than working with time series data.

Most trajectory data mining methods in the literature only focus on trajectory shape similarity search which is the task of grouping similar trajectory shapes together. These methods lead to several applications such as interactive generation of motions and discovery of subtle patterns during cellular mitoses in biological sciences [Vlachos 02]. However, they cannot be used in applications such as navigation systems and location-based information systems. For example, the prediction of elk, deer and cattle spatial distribution is helpful in land planning, stocking allocation and population management [Starkey]. Although these species live in the same region, they have different location distributions. If we do not consider the position of trajectories in the representation, we may lose this important feature. In hurricane forecast [Hurricane], an accurate prediction of the landfall location is of primary importance in hurricane forecast that helps the public to complete preparations and evacuations. Therefore, the position of hurricane trajectory must be considered in analyzing hurricane trajectories.

In this paper, we propose a trajectory representation, called *Trajectory Symbolic Aggregate approXimation* (TraSAX), which is an extension of SAX [Lin 03]. TraSAX represents the original trajectory into smaller and less complex symbol components where the position of trajectory is considered.

The rest of this paper is organized as follows. In section 2, we describe our proposal to represent trajectory data. Section 3 shows experiments of our proposal in a clustering task on four real trajectory datasets. Finally, conclusions and future directions are given in section 4.

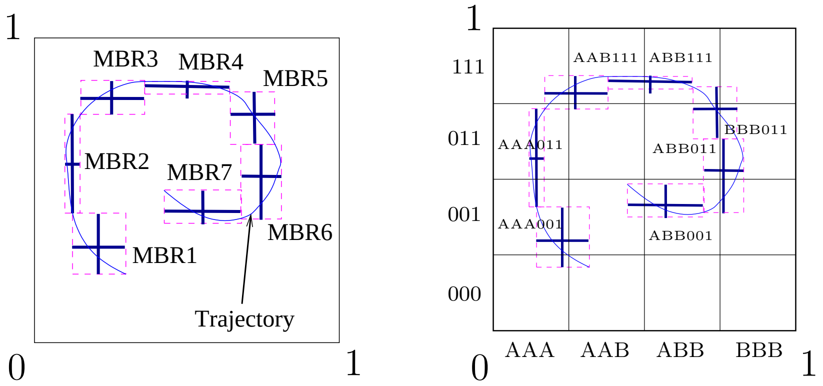
2. TraSAX representation

TraSAX allows to represent arbitrary length trajectories into equal length strings. This process includes three steps: TraSAX first normalizes the region size of the input trajectory set into a fixed region size, then Piecewise Aggregate Approximation (PAA) [Keogh 01] is employed to transform each trajectory into a PAA representation. Finally, each PAA representation is symbolized into one string.

2.1 Region Size Normalization

The region size of a trajectory set is defined as the Minimum Bounding Rectangle (MBR) [Vlachos 02] that contains all trajectories of the trajectory set. Our method can be utilized for multi-dimensional trajectory representation, however, for the ease of description, we assume that the dimensions of trajectories are two. In an x-y coordinate system, the region size of a trajectory set $\mathbb{T}$ is represented by two endpoints L_T (low point) and H_T (high point) of its major diagonal. It will be:

Contact: Nguyen Huy Thach, Dept. of Informatics, Graduate School of Systems Life Sciences, Kyushu University, Fukuoka 819-0395, Japan, tel: +81-92-802-7921, thachnh@i.kyushu-u.ac.jp



(a) MBR and PAA, w=7

(b) Map from PAA to symbols

Figure 1: TraSAX representation. a) MBRs and their corresponding PAA representations, b) Mapping from PAA representation into symbolic representation, in this example, TraSAX representation for the input trajectory: **AAA001AAA011-AAB111ABB111ABB011BBB011ABB001**.

$$Region\_size(\mathbb{T}) = (L_T, H_T) \qquad (1)$$

where $L_T = (x_{L_T}, y_{L_T})$, $H = (x_{H_T}, y_{H_T})$, $x_{L_T} \leq x_{H_T}$ and $y_{L_T} \leq y_{H_T}$.

In different applications, trajectory sets have different region sizes, i.e., the region size of a house robot trajectory could be some meters, a campus bus could have a region size of kilometers, and the trajectory region size of hurricanes could be in the order of hundred of kilometers. Therefore, we should normalize the trajectory region size being analyzed into a fixed region size. In our experiment, this fixed region size is [0, 1] × [0, 1]. A trajectory $Tr_i = \{(x_{i_1}, y_{i_1}),$ $(x_{i_2}, y_{i_2}), \ldots, (x_{i_{lenTri}}, y_{i_{lenTri}})\}$ belonging to a trajectory set $\mathbb{T}$ is represented in the new region as $TR_i = \{(X_{i_1},$ $Y_{i_1}), (X_{i_2}, Y_{i_2}), \ldots, (X_{i_{lenTri}}, Y_{i_{lenTri}})\}$, where (X_{i_k}, Y_{i_k}) is calculated as follows:

$$\begin{cases} X_{i_k} & = \frac{x_{i_k} - x_{L_T}}{x_{H_T} - x_{L_T}} \\ Y_{i_k} & = \frac{y_{i_k} - y_{L_T}}{y_{H_T} - y_{L_T}} \end{cases}$$

2.2 PAA representation for trajectory data

After normalizing the input trajectory set into the new region of [0, 1] × [0, 1], each trajectory will be converted into the PAA representation. Given a trajectory $TR_i =$ $\{(X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2}), \ldots, (X_{i_{lenTri}}, Y_{i_{lenTri}})\}$, we divide it into w equal sub-trajectories ($w < len_{Tri}$, typically $w \ll len_{Tri}$). In order to reduce the dimension of trajectories, each sub-trajectory is represented by one MBR, where the mathematical mean of each MBR becomes the representing point of that MBR and is calculated as follows:

$$\begin{cases} \overline{X}_{i_k} & = \frac{w}{lenTri} \sum_{X_{i_j} \in MBR_k} X_{i_j} \\ \overline{Y}_{i_k} & = \frac{w}{lenTri} \sum_{Y_{i_j} \in MBR_k} Y_{i_j} \end{cases}$$

By this process, the PAA approximation of trajectory Tr_i is represented by a two dimensional vector: $\overline{TR}_i =$ $\{(\overline{X}_{i_1}, \overline{Y}_{i_1}), (\overline{X}_{i_2}, \overline{Y}_{i_2}), \ldots, (\overline{X}_{i_{lenTri}}, \overline{Y}_{i_{lenTri}})\}$. Fig. 1a illustrates the idea of this step.

2.3 Transforming PAA into a Symbolic Representation

We employ "breakpoints" defined in [Lin 03] to transform PAA representation into a symbolic representation. In [Lin 03], the "breakpoints" are defined with an objective to produce symbols with equiprobability. However, in this work, our proposal is to have a symbolic representation where the position of a trajectory is considered, hence we have invented a discretization technique which defines a uniform grid. The breakpoints divide the normalized region into a × a equal-sized cells. In our experiments, we use the same breakpoint set for both X and Y dimensions.

Definition 1. *Breakpoints for trajectory* are the sorted list of number $B = \beta_0, \beta_1, \ldots, \beta a$ where $\beta_i = \frac{i}{a}$ $(0 \leq i \leq a)$.

Breakpoints divide the region of [0, 1] × [0, 1] into a × a equal-sized cells and each cell is represented by a unique symbol. We define symbols for each axis then combine them together to determine a symbol for each cell. Mathematically, in each axis, the i^{th} segment between two conjunctive breakpoints β_{i-1} and β_i is defined as either an alphabet: $alphabet(i)$ for X-axis or a number: $number(i)$ for Y-axis, where $alphabet(i)$ contains (a-i) characters 'A' and (i-1) characters 'B' while $number(i)$ contains (a-i) characters '0' and (i-1) characters '1':

$$\begin{cases} alphabet(i) = \underbrace{AA\ldots A}_{(a-i)}\underbrace{BB\ldots B}_{(i-1)} \\ number(i) = \underbrace{00\ldots 0}_{(a-i)}\underbrace{11\ldots 1}_{(i-1)} \end{cases}$$

For example, let a=4, $alphabet(1)$ = 'AAA', $alphabet(2)$ = 'AAB', $number(1)$ = '000', $number(2)$ = '001'. We use this symbol set to encode cells of the trajectory region instead of using a unique character per one cell to guarantee that symbol-based distance functions applied in our symbolic space will have a close correlation with the corresponding distance measures defined on the original space. In [Shieh 08], binary bits have also been used to represent time series where distance value is obtained from a lookup table. In contrast, the distance value of TraSAX is contained in the representation itself. For example, we achieve distance between '00' and '11' is greater than the distance between '00' and '01' in almost compression-based distance functions.

Once symbol sets of X-axis and Y-axis have been defined, the PAA representation $\overline{TR}_i = \{(\overline{X}_{i_1}, \overline{Y}_{i_1}), (\overline{X}_{i_2}, \overline{Y}_{i_2}), \ldots, (\overline{X}_{i_{lenTri}}, \overline{Y}_{i_{lenTri}})\}$ is represented as a string $\widehat{TR}_i = \widehat{X}_{i_1}, \widehat{Y}_{i_1}, \widehat{X}_{i_2}, \widehat{Y}_{i_2}, \ldots, \widehat{X}_{i_{lenTri}}, \widehat{Y}_{i_{lenTri}}$, where

$$\begin{cases} \widehat{X}_{ij} = alphabet(k) \quad if \quad \beta_{k-1} \leq \overline{X}_{ij} < \beta_k \\ \widehat{Y}_{ij} = number(l) \quad if \quad \beta_{l-1} \leq \overline{Y}_{ij} < \beta_l \end{cases}$$

Fig. 1b shows an example of this definition where a=4 and the TraSAX representation of the input trajectory is **AAA001AAA011AAB111ABB111ABB011-BBB011ABB001**.

3. Experimental Results

Our representation can be applied to various trajectory data mining tasks including clustering, classification, indexing, finding motif of subsequence and anomaly subsequence detection. In this paper, we verify our proposal by using the hierarchical clustering with Compression-Based Dissimilarity Measure (CDM) [Keogh 04] in four real trajectory datasets. We have chosen the hierarchical clustering task since it is one of the most commonly used clusterings [Berkhin 02] and it provides a visual dendrogram to clearly understand the clustering process. Beside that, CDM is a parameter-free algorithm and [Keogh 04] shows that CDM performs significantly better than almost other distance/dissimilarity measures. In addition, the results of the original data, a shape-based [Lin 03], the 2 Dimensional Discrete Fourier Transform (2D DFT) [Shatkay 95] representations are also obtained to be compared with our proposal.

3.1 Data sets

Four real trajectory datasets: *ASLclean*, *trackingPoor*, *cameraMouse* [Keogh 06] and *ourRobot* are employed to verify our proposal. *OurRobot* data contains trajectories of one robot in one of our projects [Suzuki 09]. In the task of *ourRobot*, the objective of each robot is to visit as many cells in the workspace as possible in a limited time. Interested readers are referred to [Suzuki 09]. Basic statistics and a plot of the four trajectory data are shown in Table 1 and Fig. 2, respectively.

3.2 Clustering Results and Comparisons

The results of *Original* data, *Shape-based*, 2D DFT and TraSAX representations are shown in Fig. 3. We have four datasets, so we would hope to get a clustering result of four clusters where each cluster contains trajectories of only one dataset. In hierarchical clustering, to achieve k clusters, we just have to cut the k-1 [Berkhin 02] longest links from the root, in our case k=4. Fig. 3d shows a perfect result of our proposal where each sub-tree after the cutting contains trajectories of only one dataset. None of *Original*, *Shape-based* or 2D DFT representations obtain the perfect result.

We use three metrics Q [Keogh 04], *Normalized Mutual Information* (*NMI*) and *Accuracy* (*Ac*) [Amig 09] to evaluate the quality of clustering results (due to lack of space, interested readers of these metrics are referred to [Keogh 04] and [Amig 09]). The results in Table 2 and Fig. 3 confirm the superiority of our proposal. Firstly, our representation outperforms 2D DFT representation by 25%. Secondly, Fig. 3b shows that if we only consider trajectory shape property, we lose useful information. Finally, TraSAX shows a better performance than the original data representation as shown in Fig. 3a.

Table 1: Properties of real trajectory datasets (# traj. represents the number of trajectories, avg. length represents the average length of trajectories)

No.	Dataset Name	# traj.	avg. length
1	ASLclean (AC)	23	69
2	trackingPoor (TP)	23	543
3	cameraMouse (CM)	15	1151
4	ourRobot (OR)	4	6400

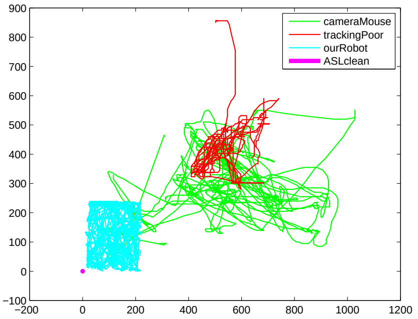


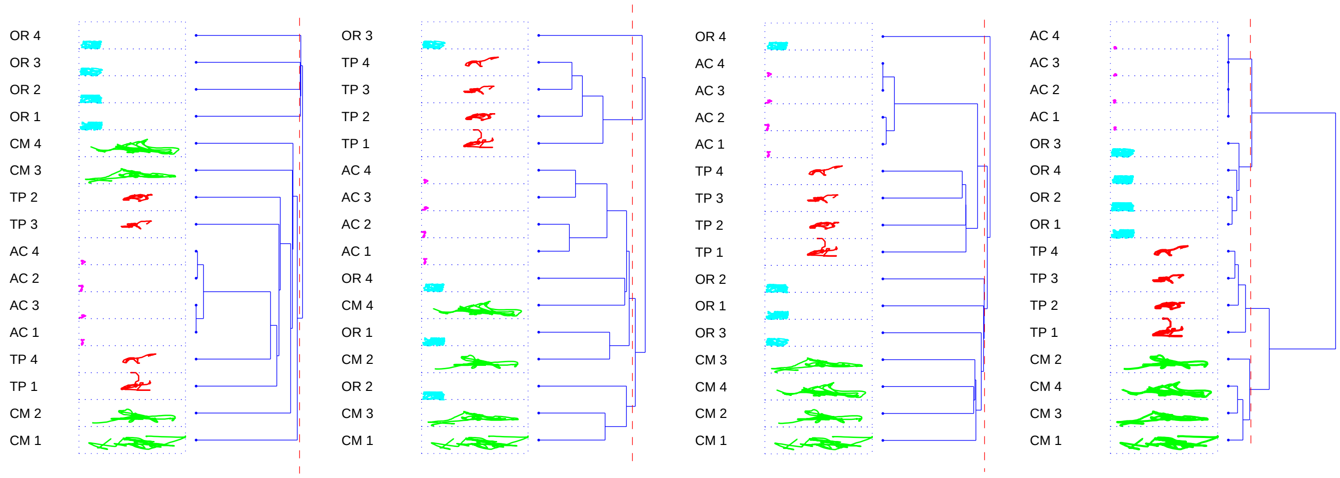
Figure 2: Shape and distribution of four real datasets

Table 2: Q, NMI and Ac results of Original data, Shape-based, 2D DFT and TraSAX representations.

	Original	Shape-based	DFT	TraSAX
Q	0.5	0.5	0.75	1
AC	0.75	0.688	0.813	1
NMI	0.75	0.508	0.696	1

4. Conclusions and Future Works

In this paper, we have proposed a novel symbolic representation for trajectory data. Our representation has three main advantages: it reduces dimensions of the input trajectory during the discretization process by symbolic representation, and it allows to increase accuracy compared to the original data. In addition, our representation gives an empirical evidence that the position property of trajectory



(a) Original data

(b) Shape-based representation

(c) 2D DFT representation

(d) TraSAX representation

Figure 3: Experimental results of four trajectory datasets where *AC*, *OR*, *TP* and *CM* stand for *ASLclean*, *ourRobot*, *trackingPoor* and *cameraMouse*, respectively. The vertical dashed line in each plot is the cutting line to achieve four clusters.

is important in trajectory representation. The clustering results on three benchmark data sets and one trajectory data set from our project have shown the effectiveness of our proposal.

As future works, we plan to employ our proposal in other trajectory data mining tasks including classification, indexing, finding motifs of sub-trajectory and anomaly sub-trajectory detection. We also plan to study effects of injecting various degrees and types of noises to our proposal. Another direction of research is to improve the performance in speed when the size of input trajectory datasets increases, e.g., to terabytes.

Acknowledgments

A part of this research was supported by Strategic International Cooperative Program funded by Japan Science and Technology Agency (JST).

References

[Amig 09] E. Amig, J. Gonzalo, J. Artiles and F. Verdejo. *"A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints"*. Inf. Ret., Kluwer Academic Publishers, pp. 461-486, ISSN:1386-4564.

[Berkhin 02] P. Berkhin. *"Survey of Clustering Data Mining Techniques"*. Tech. Rep., Accrue Software, San Jose.

[Hurricane] http://weather.unisys.com/hurricane/atlantic/

[Keogh 01] Keogh, E., Chakrabarti, K., Pazzani, M. and Mehrotra, S. *"Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Database"*. In Proc. 2001 ACM SIGMOD, May 21-24, pp. 151-162.

[Keogh 04] Keogh, E., Lonardi, S. and Ratanamahatana, C. *"Towards Parameter-Free Data Mining"*. In Proc. KDD 2004, pp. 206-215.

[Keogh 06] http://www.cs.ucr.edu/˜eamonn/Keogh_Time_Series_CDrom.zip

[Lee 07] J. Lee, J. Han, K. Whang. *"Trajectory Clustering: a Partition-and-Group Framework"*. Proc. 2007 ACM SIGMOD, pp. 593-604.

[Lin 03] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. *"A Symbolic Representation of Time Series, with Implications for Streaming Algorithms"*. In Proc. Eighth ACM SIGMOD Workshop. June 13. pp. 2-11.

[Rtreeportal] http://www.rtreeportal.org/

[Shatkay 95] Shatkay, H. *"The Fourier Transform - a Primer"*. Tech. Rep. CS-95-37, Dep. of Computer Science, Brown University.

[Shieh 08] J. Shieh and E. Keogh. *"iSAX: Indexing and Mining Terabyte Sized Time Series"*. In KDD '08, pp. 623-631, NY, USA, 2008, ACM.

[Starkey] http://www.fs.fed.us/pnw/starkey/

[Suzuki 09] E. Suzuki, S. Takano, H. Hirai. *"Toward Using Symbolic Discovery in Designing Controllers of Autonomous Swarm Robots."* In Proc. First LEMIR, pp. 1-10, Sept. 2009.

[Vlachos 02] M. Vlachos, G. Kollios, and D. Gunopulos. *"Discovering Similar Multidimensional Trajectories"*. In Proc. 18th ICDE, pp. 673-684.