Introduction to Machine Learning Research on Time Series

Umaa Rebbapragada
Tufts University
Advisor: Carla Brodley
1/29/07

Machine Learning (ML)

- Originally a subfield of AI
- Extraction of rules and patterns from data sets
- Focused on:
    - Computational complexity
    - Memory

Machine Learning Tasks for Time Series

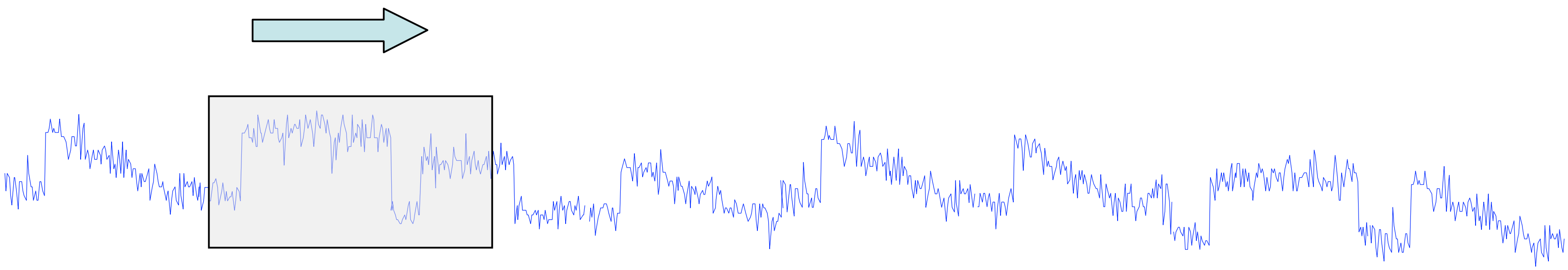
- Classification
- Clustering
- Semi-supervised learning
- Anomaly Detection

Assumptions

- Univariate time series
- Time series databases

Single Time Series

- A single long time series can be converted into a set of smaller time series by sliding a window incrementally across the time series :



- Window length is usually a user-specified parameter.

Challenges of Times Series Data

- High dimensional
- Voluminous
- Requires fast technique

Brute Force Similarity Search

- Given query time series Q, the best match by sequential scanning is found by:

$$\min_{1 \leq i \leq N} \sum_{t=1}^{d} (X_i(t) - Q(t))^2$$

- O(nd)
- Finding the nearest neighbor for each time series in the database is prohibitive.

Similarity Search

- Clustering and classification methods perform many similarity calculations
- Some require storage of the *k* nearest neighbors of each data instance
- Critical that these calculations be fast

Speeding up Similarity Search

- Alternate time series representations
- Search databases faster
- New similarity metrics

Data Mining Time Series Toolbox

- Indexing
- Dimensionality Reduction
- Segmentation
- Discretization
- Similarity metric

Indexing

- Faster than a sequential scan
- Insertions and deletions do not require rebuilding the entire index
- Partition the data into regions
- Search regions that contain a likely match
- Requires a similarity metric that obeys triangle inequality

Indexing

- R-trees
- kd-trees
- linear quad-trees
- grid-files

Indexing on Times Series Data

- High dimensionality slows down speed of computation
- Curse of dimensionality inhibits efficiency of of indexing

Dimensionality Reduction

- Reduces the size of the time series
- Distance on transformed data should lower bound the original distance

$$D_{trans}(F(P), F(Q)) \leq D_{orig}(P, Q)$$

- This guarantees no false dismissals (false negatives)

Dimensionality Reduction: DFT, DWT, SVD

- Represent time series using subsets of
  - Fourier coefficients
  - Wavelet coefficients
  - eigenvalue/vectors
- Euclidean-distance is lower-bounded on DFT[1], DWT[2], SVD[3]

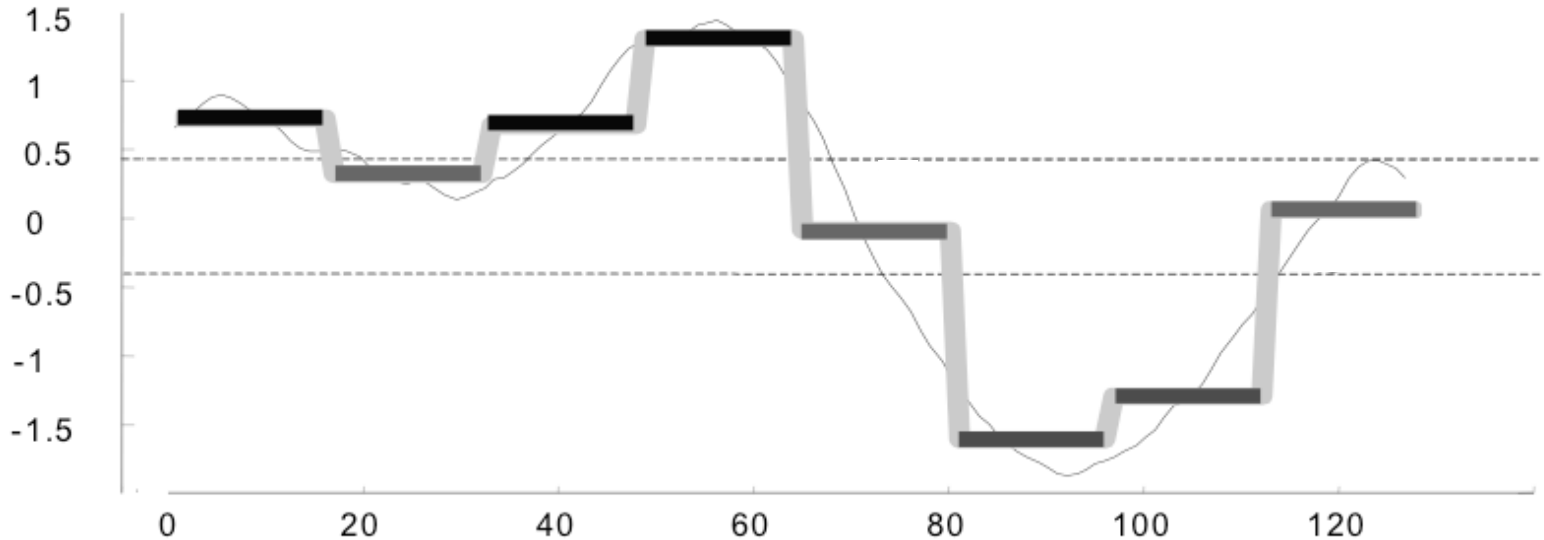
[1] C. Faloutsos et al.: Fast Subsequence Matching in Time-Series Databases. SIGMOD Conference 1994: 419-429
[2] K. Chan and A. Fu: Efficient Time Series Matching by Wavelets. ICDE 1999: 126-133
[3] F. Korn et al.: Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. SIGMOD Conference 1997: 289-300

Gemini Framework

- Faloutsos et al., 1994
- Map each time series to a lower dimension
- Store in multi-dimensional indexing structure

C. Faloutsos et al.: Fast Subsequence Matching in Time-Series Databases. SIGMOD Conference 1994: 419-429

Piecewise Aggregate Approximation (PAA)

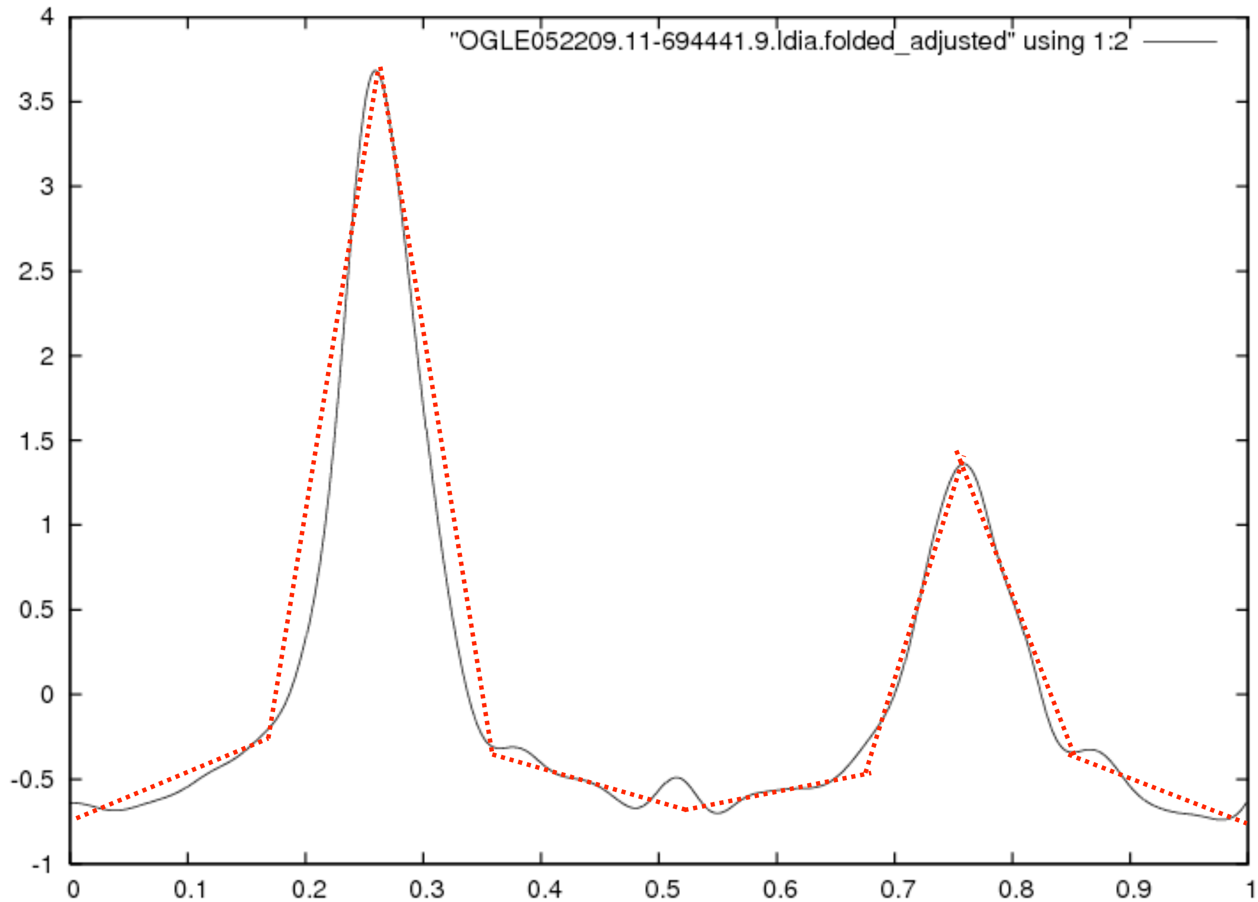


Eamonn J. Keogh, et al.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases.
Knowl. Inf. Syst. 3(3): 263-286 (2001)
Fig: Eamonn J. Keogh, et al.: HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. ICDM 2005: 226-233

Segmentation

- Represent the time series in smaller, less complex segments.
  - Piecewise Linear Approximation (PLA)
  - Minimum Bounding Rectangles (MBR)

Piecewise Linear Approximation (PLA)



Minimum-Bounding Rectangles (MBR)

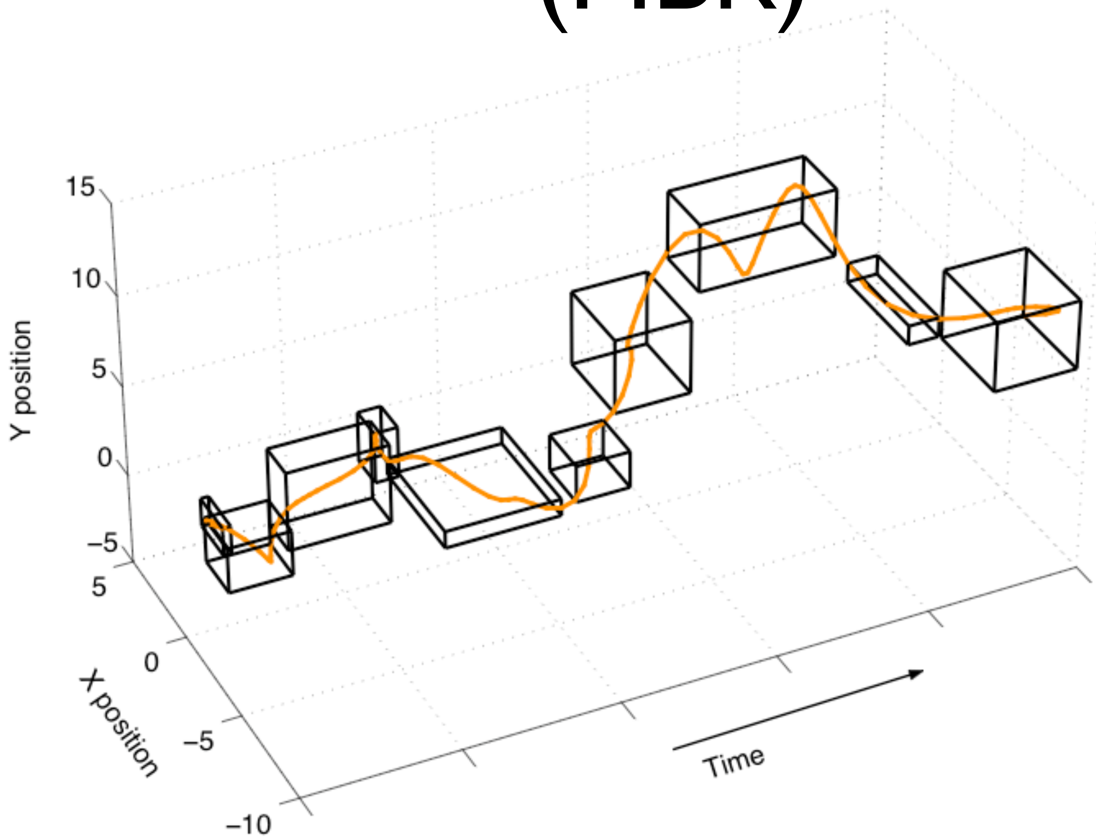
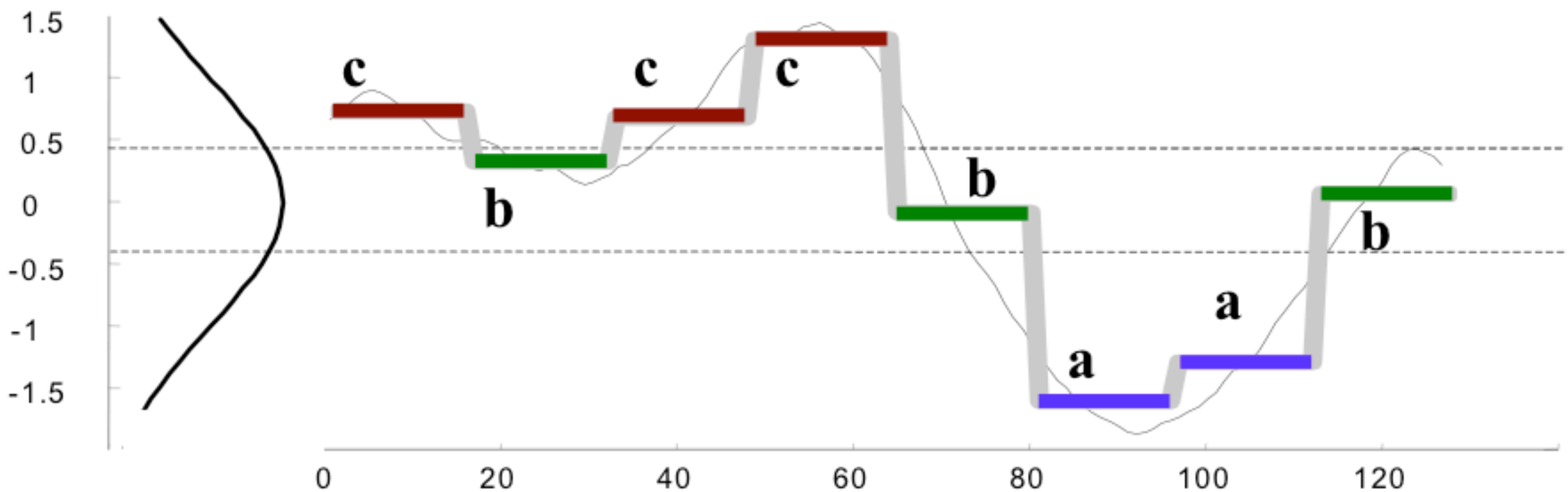


Fig: A. Anagnostopoulos et al: Global distance-based segmentation of trajectories. SIGKDD Conference 2006: 34-43

Discretization

- Transforms a real-valued time series into a sequence of characters from a discrete alphabet
- Dimensionality reduction implicit
- Allows use of string functions on time series

SAX



Jessica Lin et al. A symbolic representation of time series, with implications for streaming algorithms. DMKD 2003: 2-11
Fig: Eamonn J. Keogh, et al.: HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. ICDM 2005: 226-233

Is Euclidean Distance Best Metric?

- Everything discussed so far used ED as similarity metric
- Is it the best similarity metric for time series?

Drawbacks of Euclidean Distance

- Requires two time series to have same dimensionality
- 1-to-1 alignment of the time axis

Cross Correlation

- Cross correlation with convolution can find optimal phase shift to maximize similarity

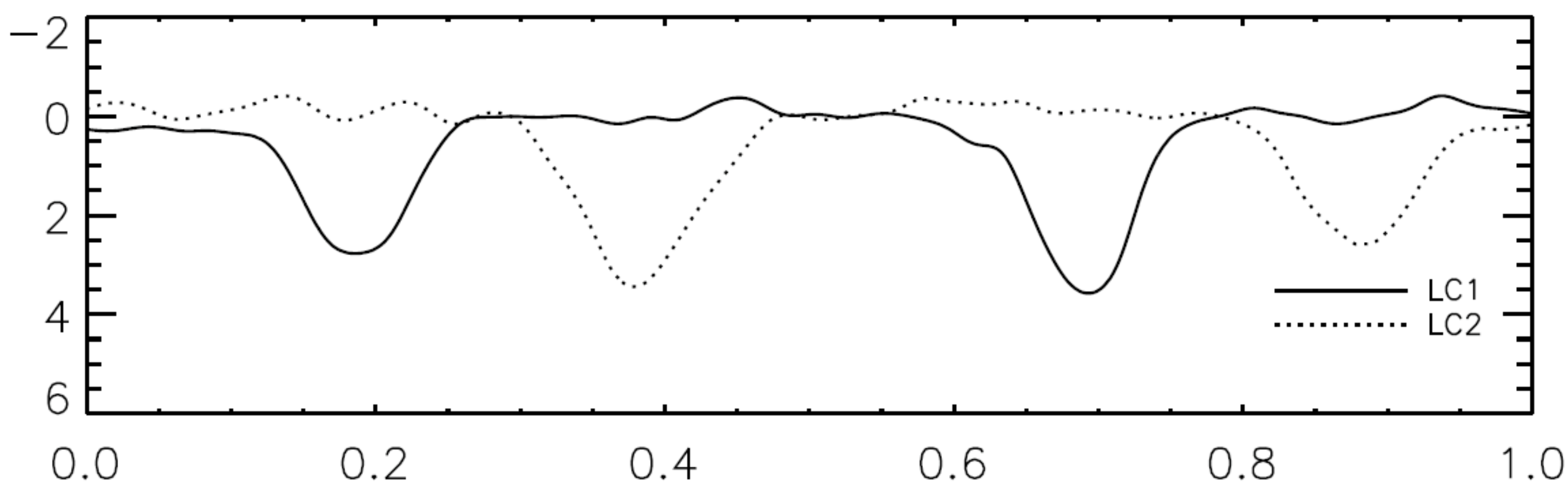


Fig: P. Protopapas et al.: Finding outlier light-curves in catalogs of periodic variable stars. Mon. Not. Roy. Astron. Soc. 369 (2006) 677-696

Cross Correlation

- Optimal phase shift (to left) of solid line is 0.3

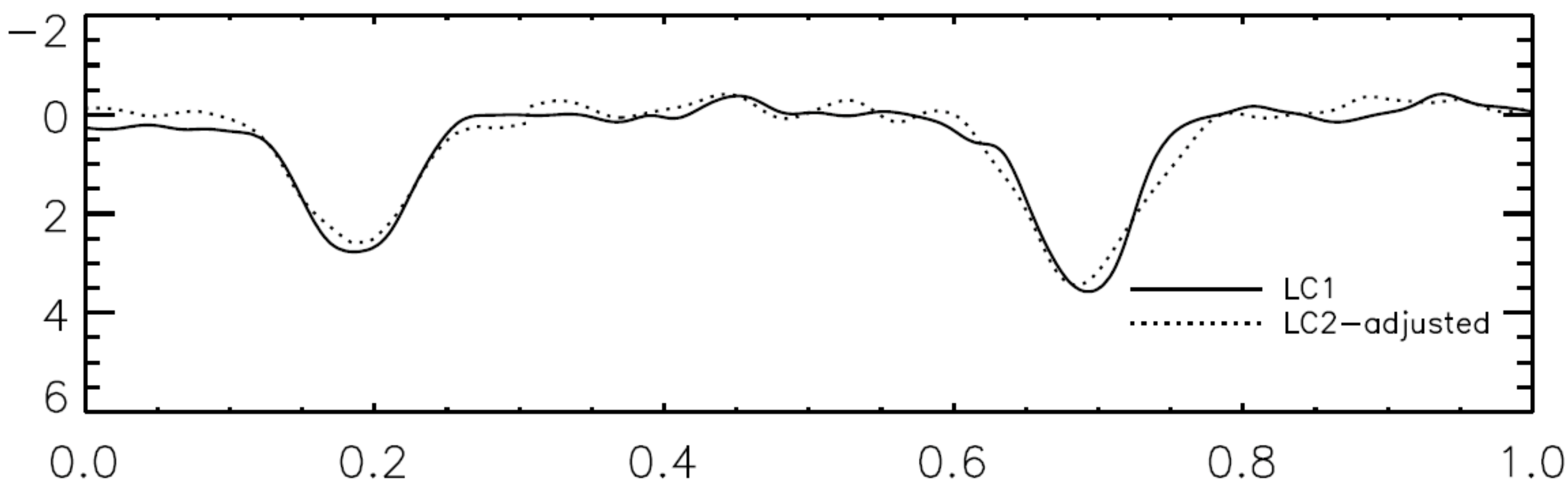
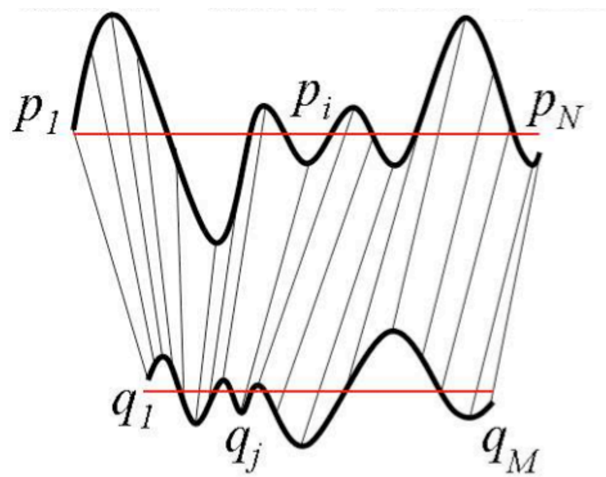


Fig: P. Protopapas et al.: Finding outlier light-curves in catalogs of periodic variable stars. Mon. Not. Roy. Astron. Soc. 369 (2006) 677-696

Dynamic Time Warping (DTW)

- DTW allows many-to-one alignment
- Time series need not be same size



"Warped" Time Axis

Fig: Y. Sakurai, et al.: FTW: fast similarity search under the time warping distance. PODS 2005: 326-337
D. J. Berndt, and J. Clifford: Finding Patterns in Time Series: A Dynamic Programming Approach. Advances in Knowledge Discovery and Data Mining 1996: 229-248

DTW Algorithm

$$1 \leq i \leq N, \quad 1 \leq j \leq M$$

$$f[i,j] = d(i,j) + \min \begin{cases} f(i, j-1) \\ f(i-1, j-1) \\ f(j-1, j) \end{cases}$$

$$D_{\mathrm{dtw}} = f(N, M)$$

DTW Algorithm

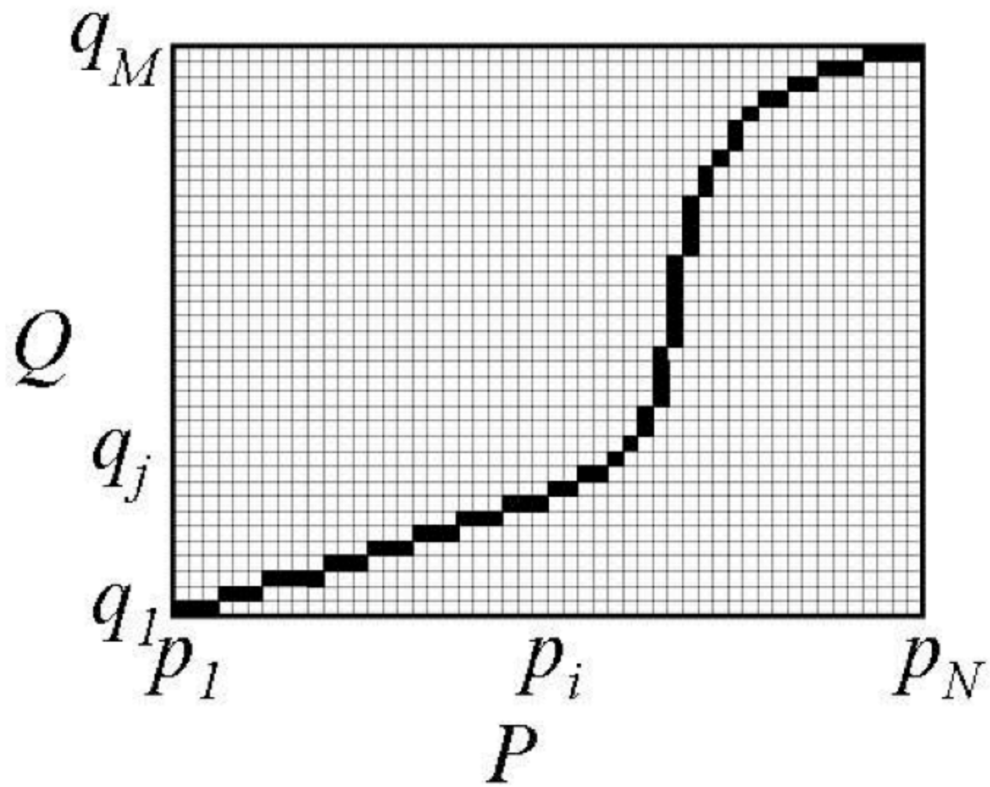


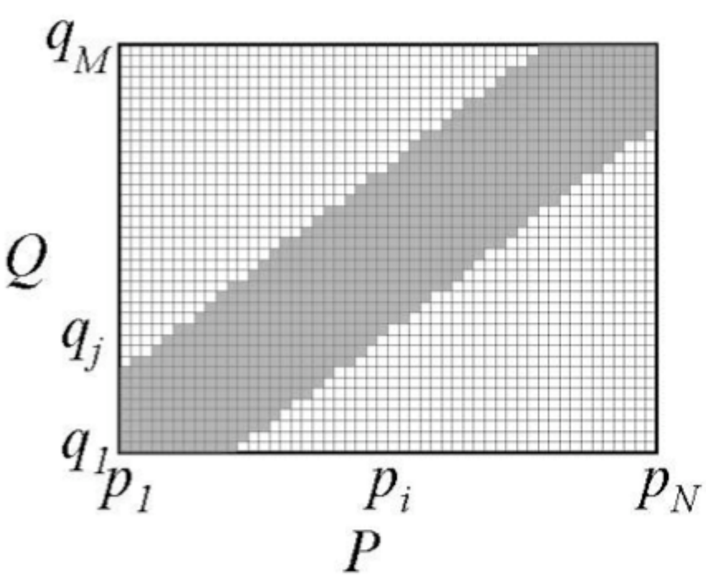
Fig: Y. Sakurai, et al.: FTW: fast similarity search under the time warping distance. PODS 2005: 326-337

Drawbacks of DTW

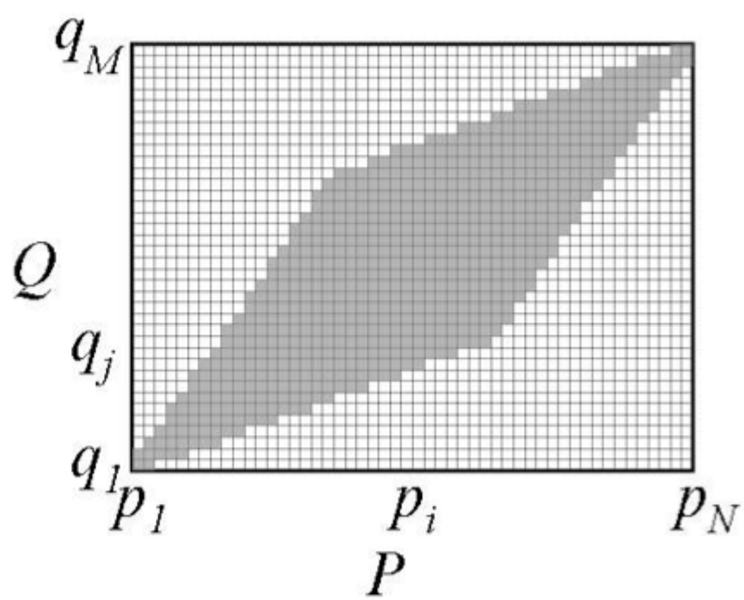
- Computationally expensive
- Does not adhere to triangle inequality => cannot use it for indexing

Making DTW Faster

- Global constraints:



Sakoe-Chiba Band



Itakura Parallelogram

Y. Sakurai, et al.: FTW: fast similarity search under the time warping distance. PODS 2005: 326-337

Making DTW Faster

- Y. Sakurai et al.: FTW: fast similarity search under the time warping distance. PODS 2005: 326-337
- E. Keogh and C. Ratanamahatana: Exact indexing of dynamic time warping. Knowl. Inf. Syst. 7(3): 358-386 (2005)
- Y. Zhu and D. Shasha: Warping Indexes with Envelope Transforms for Query by Humming. SIGMOD Conference 2003: 181-192
- E. Keogh and M. Pazzani: Scaling up dynamic time warping for datamining applications. KDD 2000: 285-289
- B.-K. Yi et al.: Efficient Retrieval of Similar Time Sequences Under Time Warping. ICDE 1998: 201-208

Other Areas of Research

- Anomaly Detection
- Change Point Detection

Thesis Research

- Anomaly detection methods
  - fast
  - preserve interesting features

Thank You