Improving Performance of Similarity Measures for Uncertain Time Series using Preprocessing Techniques

Mahsa Orang
Nematollaah Shiri

27th International Conference on Scientific and Statistical Database Management
Introduction

\[ u = \langle u_1, \ldots, u_n \rangle \]

\[ U = \langle U_1, \ldots, U_n \rangle \]
Introduction

Uncertain Time Series

\[ U = \langle U_1, \ldots, U_n \rangle \]

Multiple readings

Privacy concerns

Forecasting techniques

Data collection error

\[ \forall i, U_i = t_i + E_i \]

- Wireless sensor network
- Medical data analysis
- Location-based services
Introduction: Similarity Search Approaches

Traditional Similarity Measures

Uncertain Similarity Measures

Values

Values+
Statistical Information
Introduction: Similarity Search Approaches

Traditional Similarity Measures OUTPERFORM Uncertain Similarity Measures

[DNMP,12]

Why? Using a Preprocessing Step
Introduction: Preprocessing

Filtering:
- Smoother but farther from the exact values.

Filtering + Normalization:
- Closer to the exact values.

Uncertain Time Series

Standard Time Series
Motivation

- Can preprocessing improve the performance of uncertain similarity measures?
- Will uncertain similarity measures outperform traditional ones using preprocessing techniques?
Outline

• Uncertain Similarity Measures
• Preprocessing Techniques
• Experimental Evaluation
Uncertain Similarity Measures

- Orang & Shiri [OS,14]

$\text{Uncertain Similarity Measures}$
Given a dataset of uncertain time series $D$, an uncertain time series $Q$, a similarity threshold $s$, and a probability threshold $p$, a **probabilistic query** searches for uncertain time series $X$ in $D$ such that similarity between $X$ and $Q$ is higher than $s$, with a confidence of at least $p$. 
Preprocessing Techniques

- Moving Average Filters
- Normalization
$x =< x_1, \ldots, x_n >$

**MA**

Simple Moving Average

$x^MA =< x_1^MA, \ldots, x_m^MA >$

Each value is substituted by average of adjacent values.

Each value is substituted by weighted average of adjacent values, weights decrease exponentially.
Normalization makes similarity measures invariant to scaling and shifting and hence helps better capture the similarity.
Experiment Setup

- CPU: 2.66 GHz, RAM: 4GB
- 16 UCR datasets [KZHHXWR]
- Set up similar to [YWY, 09], [SM, 10]:
- Data Parameters:
  - The standard deviation of UTS: \( r \times \sigma \)
    - \( r \): Standard deviation ratio (SDR), varied from 0.01 to 4
    - \( \sigma \): The standard deviation of the given standard time series
  - Error distribution function: Exponential, Normal, Uniform
- Query Parameters
  - Probability threshold, varied from 0.1 to 0.9.
  - Similarity threshold,
    - Correlation threshold \( c \), varied from 0.1 to 0.9
    - Euclidean threshold \( d \), \( d = 2(n - 1)(1 - c) \)
- Filtering Parameters [DNMP, 12]
  - \( w = 2 \) and \( \lambda = 1 \)
- Performance measure: Classification error, F1 Score
  - Average over 10 runs
1) Deterministic Similarity Measures

- Comparison approach: \textit{1NN-classification with K-fold cross-validation} [DTSWK,08]

No Preprocessing

Simple Moving Average

DUST Error < Euclidean Error

Uncertain Moving Average

Uncertain Exponential moving Average

The weighted filters help the Euclidean distance achieve similar performance as a weighted similarity measure such as DUST.

DUST Error \approx= \text{Euclidean Error}
2) Probabilistic Similarity Measures: Uncertain Correlation

No Preprocessing

Simple Moving Average

More results

Uncertain Moving Average

Uncertain Exponential Moving Average
2) Probabilistic Similarity Measures: Uncertain Correlation

- Simple Moving Average
- Uncertain Moving Average
- Uncertain Exponential moving Average

For low probability thresholds, the F1 score is higher than simple moving average.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Filter</th>
<th>None</th>
<th>MA</th>
<th>UMA</th>
<th>UEMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>50words</td>
<td>None</td>
<td>0.38</td>
<td>0.53 ( +39%)</td>
<td>0.5 ( +32%)</td>
<td>0.48 ( +26%)</td>
</tr>
<tr>
<td>Adiac</td>
<td>None</td>
<td>0.51</td>
<td>0.72 ( +41%)</td>
<td>0.74 ( +45%)</td>
<td>0.71 ( +39%)</td>
</tr>
<tr>
<td>Beef</td>
<td>None</td>
<td>0.49</td>
<td>0.69 ( +41%)</td>
<td>0.69 ( +41%)</td>
<td>0.62 ( +27%)</td>
</tr>
<tr>
<td>CBF</td>
<td>None</td>
<td>0.38</td>
<td>0.48 ( +26%)</td>
<td>0.45 ( +18%)</td>
<td>0.45 ( +18%)</td>
</tr>
<tr>
<td>Coffee</td>
<td>None</td>
<td>0.51</td>
<td>0.73 ( +43%)</td>
<td>0.69 ( +35%)</td>
<td>0.68 ( +33%)</td>
</tr>
<tr>
<td>ECG200</td>
<td>None</td>
<td>0.46</td>
<td>0.62 ( +35%)</td>
<td>0.62 ( +35%)</td>
<td>0.61 ( +33%)</td>
</tr>
<tr>
<td>FISH</td>
<td>None</td>
<td>0.5</td>
<td>0.71 ( +42%)</td>
<td>0.74 ( +48%)</td>
<td>0.69 ( +38%)</td>
</tr>
<tr>
<td>FaceFour</td>
<td>None</td>
<td>0.39</td>
<td>0.53 ( +36%)</td>
<td>0.53 ( +36%)</td>
<td>0.5 ( +28%)</td>
</tr>
<tr>
<td>Gun-Point</td>
<td>None</td>
<td>0.48</td>
<td>0.68 ( +42%)</td>
<td>0.69 ( +44%)</td>
<td>0.66 ( +38%)</td>
</tr>
<tr>
<td>Lighting2</td>
<td>None</td>
<td>0.31</td>
<td>0.42 ( +35%)</td>
<td>0.43 ( +39%)</td>
<td>0.41 ( +32%)</td>
</tr>
<tr>
<td>Lighting7</td>
<td>None</td>
<td>0.35</td>
<td>0.47 ( +34%)</td>
<td>0.47 (34%)</td>
<td>0.45 ( +29%)</td>
</tr>
<tr>
<td>OSULeaf</td>
<td>None</td>
<td>0.34</td>
<td>0.47 ( +38%)</td>
<td>0.46 ( +35%)</td>
<td>0.43 ( +26%)</td>
</tr>
<tr>
<td>OliveOil</td>
<td>None</td>
<td>0.51</td>
<td>0.73 ( +43%)</td>
<td>0.73 ( +43%)</td>
<td>0.66 ( +29%)</td>
</tr>
<tr>
<td>Swedish</td>
<td>None</td>
<td>0.46</td>
<td>0.62 ( +35%)</td>
<td>0.63 ( +37%)</td>
<td>0.60 ( +30%)</td>
</tr>
<tr>
<td>Synthetic</td>
<td>None</td>
<td>0.33</td>
<td>0.33 ( 0%)</td>
<td>0.32 ( -3%)</td>
<td>0.33 ( 0%)</td>
</tr>
<tr>
<td>Trace</td>
<td>None</td>
<td>0.47</td>
<td>0.66 ( +40%)</td>
<td>0.66 ( +40%)</td>
<td>0.62 ( +32%)</td>
</tr>
</tbody>
</table>
2) Probabilistic Similarity Measures: Uncertain Correlation

Pearson Correlation
2) Probabilistic Similarity Measures: PROUD

- No Preprocessing
- Simple Moving Average
- Uncertain Moving Average
- Uncertain Exponential moving Average
2) Probabilistic Similarity Measures: PROUD

\[ Eucl(x, y) \leq 100 \]

\[ P(Eucl(X, Y) \leq 100) \geq 0.8 \]
2) Probabilistic Similarity Measures: PROUD

\[ Eucl(X, Y) = \sum_{i=1}^{n} (X_i - Y_i)^2 \]

\[ E(Eucl(X, Y)) = Eucl(E(X), E(Y)) + 300 \]

\[ E(X) = <E(X_1), \ldots, E(X_n)> \]
**PROUDS**: An Enhanced Version of PROUD

\[
Eucl(X, Y) = \sum_{i=1}^{n} (E(X_i)^2 + E(Y_i)^2 + 2X_iY_i)
\]

\[
E(Eucl(X, Y)) = Eucl(E(X), E(Y))
\]

**Lemma 1.** Given uncertain time series \(X\) and \(Y\) with normal forms \(\hat{X}\) and \(\hat{Y}\), the following holds.

\[
Eucl(\hat{X}, \hat{Y}) = 2(n - 1)(1 - Corr(X, Y))
\]
2) Probabilistic Similarity Measures: Uncertain Correlation
Conclusion

- Uncertain similarity measures can outperform the traditional similarity measures with and without data preprocessing.
  - This indicates the effectiveness of uncertain similarity measures in practice.
  - Uncertain similarity measures utilize all the available information to better quantify the similarity.
  - Probabilistic similarity measures provide the users with more information about reliability of the result.

- Preprocessing is necessary for similarity search in uncertain time series.
  - Simple and uncertain moving average filters improve the performance of the probabilistic measures more than uncertain exponential moving average filter.

- We propose an enhancement for the PROUD similarity measure, which improves its performance with and without data preprocessing.
  - We found a linear relationship between probabilistic similarity measures.
Future Work

• Research on uncertain time series is new, mostly focused on modeling and similarity search.

• More work is required in this field:
  – e.g., pattern discovery, indexing, prediction
References


Sorry we could not attend the conference.

For any questions or comments, please contact us at:

m_oran@encs.concordia.ca