Time Series: Defining a Search Engine

Philipp Beer

September 25, 2021

1 Thesis: Time Series Search Engine

Purpose

The purpose of this thesis is to explore the possibility of creating a time series series search engine.

Introduction

Time Series Definition

Time series is often described as "anything that is observed sequentially over time" which usually are observed at regular intervals of time [1]. They can be described as collection of observations that are considered together in chronological order rather than as individual values or a multiset of values. Their representation can be described as ordered pairs: $S = (s_1, s_2, \ldots, s_n)$ where $s_n = (t_n, v_n)$. t_n can be a date, timestamp or any other element that defines order. v_1 represents the observation at that position in the time series.

Fu in their work [2] categorized time series research into (1) representation, (2) indexing, (3) similarity measure, (4) segmentation, (5) visualization and (6) mining. Research in these different fields started taking off in the second half of the 20^{th} century. For example in [3] the authors worked on questions of representation via sampling the sampling of time series in 1969.

Time series share the same challenges as other high dimensional data in that it quickly requires high computational power to process them and they suffer from the "curse of dimensionality" [4].

Applications

Time series are encountered everywhere. Any metric that is captured over time can utilized as time series. Granularity can be used as descriptor for the sampling rate of a series or more general how often measurements for a particular metric are taken. This granularity has a tendency to increase as well. As example consumer electronics that capture health and fitness data can mentioned. Or sensors which are utilized in the automotive industry or heavy machinery where they are employed to capture information for predictive maintenance applictions.

In the financial industry time series are a very fundamental component of decision making, like the development of stock prices over time or financial metrics of interest. The same is true for macro economic information or metrics concerning social structures in society, etc.

In the medical field time series are also ubiquitous. Whether they relate to patient date like blood pressure. The bio statistics field utilizes electrographical data like electrocardiography, electroencephalography and many others. In more aggregate medical analysis like toxicology analysis of drug treatments for vaccine approvals they are utilized and in many forms of risk management, for example, population level obesity levels.

Time series data is paramount to a wide variety of areas, relating to many different fields.

What are they used for

Time series are utilized to analyze and gain insight from historic events/patterns with respect to the observational variable(s) and their interactions. A second area of application is forecasting. Here time series are utilized to predict the observations that occur in future under the assumption that the historic information can provide insight into the behavior of the observed variables.

TODO to be integrated

- refer to previous work on measures of similarity and outcome
- measure of similarity required
- challenges with time series (domains, granularity, length, outliers)
- area of signal processing interesting methods

M-Competition

The research in time series has been numerous and focused on various properties of them as well as finding methods to accurately predict them. Aside of forecasting ell researched areas are measures of similarity and retrieval of time series.

- Forecasting In the arena of forecasting the M-competition organized by Prof. Makridakis played a big role in the development of forecasting methods shortly after their inception in 1979.

  One of the aspects that has been correct up until the 5th installment of the M-competition is that statistical methods in forecasting have outperformed more complex machine learning methods. So learning algorithms did not benefit sufficiently from learning from multiple series to generate more accurate point predictions and prediction intervals compared to the statistics-based alternatives.

  One interesting question in this area is whether clustering of time series that have similar properties and training algorithms per cluster of "similar" series can help simplify the learning process for machine learning methods and in consequence improve their performance in future competitions.

  However, expressing similarity for time series is a challenging questions with respect to which metrics to utilize, computational complexity as well as limiting assumptions that need to be made for time series.

Focus

The focus for this work is to generate a method that allows to build a time series search engine that can:

- find similar time series in a database of available datasets
- be sufficiently generic to be useful for practical applications
- and rely on algorithms with sufficiently low computational complexity so that search results can retrieved quickly

Exisiting work

Measuring similarity

In order to be able to describe the closeness of time series or multiple time series to each a measure for similarity is required. In the literature various general measures and corresponding computation methods can be found. Wang et al. reviewed time series measures and categorized the similarity measures into 4 categories: (1) lock-step measures, (2) elastic measures, (3) threshold-based measures, and (4) pattern-based measures.

Lockstep-measures include the L_p-norms (Manhatten and Euclidean Distance) as well as Dissimilarity Measure (DISSIM). **Elastic measures** include metrics like Dynamic Time Warping (DTW) and edit distance based measures like Longest Common Subsequence (LCSS), Edit Sequence on Real Sequence (EDR), Swale and Edit Distance with Real Penalty. An example for **threshold-based measures** are threshold query based similarity search (TQuEST). And Spatial Assembling Distance (SpADe) is an example for pattern-based measures.

- Euclidean Distance Euclidean Distance is the most widely used distance metric in the research of time series. (add list of papers here)
  - explain advantages
  - mention shortcomings
    - same length period
    - handling of outliers and noise
    - handling of stretching of series
    - computational complexity
- Dynamic Time Warping
  - invented by [6] in 1994
  - warp series by computing the distance from one point to all other points in the other series and define a warped path that minimizes the distance

$$DTW(S_a, S_b) = min\{\sqrt{\sum_{k=1}^{P} \delta(\omega_k)} \tag{1}$$

  - advantages: handles distortions, does not require same length ts
  - disadvantages: outliers may create a false impression of similarity, computaional complexity of $O(n^2)$ makes utilization for very long time series impractical and comparison with large sets of time series is also very time intensive
- Similarity through decomposition
  - introduce time series decomposition (reference in [1])
  - trend and seasonality (mention assumptions about period)

Time series representation

- Principal Component Analysis
- SAX
- Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT)
  - mention origin in signal processing and ubiquitous use in engineering (image and audio compression)

Challenges when building a time series

- length of series
- trend
- seasonality
- computational complexity -> issue because of data size
- granularity or sampling rates
- noise
- data quality
- similarity is task dependent (level)
- usual need for preprocessing the time series data (denoising, detrending, amplitude scaling) -> any pre-processing does modify the series

Data Analysis

what does M4 data look like

Challenges

- How many frequencies to compare?
- priorities of frequencies (power spectrum)
- different length of time series (leading to different frequencies) - ranges solved with logs

Methodology

Main contribution of the thesis

- transformation into Fourier-space
- transfer frequencies into frequency range band with increasing range width (using log scale)
- computation of frequency energy levels (sort and keep top 5) -> ask Prof. how to name this parameter
- conversion of ordered frequencies into frequency range band

- for each series to compare -> compare whether the frequency matches on the ordered positions -> provide exponential value per position -> match on more powerful frequencies is valued higher

additional computations

- utilization of FFT utilizes only frequency space (future work should consider comparison of energy levels per frequency)
- additional simple statistics computed (mean, std, quantiles)
- ts decomposition for trend estimation (requires parameter for period) -> then best line fit for slope of the time series
- computation of deltas for each series to search with statistics and slope of all other time series (review computational complexlity)
- ranking of matching series based highest frequency range match and ONE statistic

Preprocessing

- M4 data wide format vs. long format

Parallelization

- computation times
- scalability
- Samples for results only (stratification vs. non-stratification)
- Threads vs. Processes

Technology (check with Prof. if required)

R vs. Python vs. Mathematica, Matlab

- load
- transform to FFT vector space
- compare most important frequencies
- compare candidates
- select winner (which criteria)

Exploratory Data Study

- what do results look like

Formal Evaluation

- (maybe) improvement in forecasting approach
- find dataset with ground truth and compare DTW to this approach
- Distance metrics
- computational complexity

Conclusion & future work

Successes

Failures

Flaws

- final computation

What is missing

- denoising of time series
- adjustment of number of frequencies used
-

Results & Discussion

References

[1] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014. ISBN: 9780987507105. URL: https://books.google.de/books?id=gDuRBAAAQBAJ.

[2] Tak-chung Fu. "A review on time series data mining". In: *Engineering Applications of Artificial Intelligence* 24.1 (Feb. 2011), pp. 164–181. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2010.09.007. URL: http://dx.doi.org/10.1016/j.engappai.2010.09.007.

[3] K.J. Åström. "On the choice of sampling rates in parametric identification of time series". In: *Information Sciences* 1.3 (July 1969), pp. 273–278. ISSN: 0020-0255. DOI: 10.1016/s0020-0255(69)80013-7. URL: http://dx.doi.org/10.1016/S0020-0255(69)80013-7.

[4] Yongqiang Tang, Yuan Xie, Xuebing Yang, et al. "Tensor Multi-Elastic Kernel Self-Paced Learning for Time Series Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* (2019), pp. 1–1. ISSN: 2326-3865. DOI: 10.1109/tkde.2019.2937027. URL: http://dx.doi.org/10.1109/TKDE.2019.2937027.

[5] Xiaoyue Wang, Abdullah Mueen, Hui Ding, et al. "Experimental comparison of representation methods and distance measures for time series data". In: *Data Mining and Knowledge Discovery* 26.2 (Feb. 2012), pp. 275–309. ISSN: 1573-756X. DOI: 10.1007/s10618-012-0250-5. URL: http://dx.doi.org/10.1007/s10618-012-0250-5.

[6] Donald J. Berndt and James Clifford. "Using Dynamic Time Warping to Find Patterns in Time Series". In: KDD Workshop. 1994, pp. 359–370.