University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, 7. März 2016

MSc Project: Implementing Correlation Measures for Streaming Time Series

A streaming time series s_i is a sequence of data points that receives a new value every time unit (e.g. every 10 minutes). Such data appears in many applications, e.g., the financial stock market, meteorology, sensor networks and network monitoring to name only a few. Time series data is often incomplete, though, due to sensor failures or other issues. In our research we developed the Top-k Case Matching (TKCM) algorithm [2] to *impute* missing values in streams of meteorological time series, i.e. to replace the missing values with good estimates of what the values could have been.

Since streaming time series are by definition unbounded, TKCM can only keep a portion of a time series in main memory. Let $W = [\underline{t}, \bar{t}]$ be a time window of length $|W|$, where time $\bar{t}$ represents the most recent time point for which the streams produced a new value and $\underline{t}$ represents the oldest time point that still fits into the sliding window. TKCM imputes a missing value at time $\bar{t}$ of (base) time series s_i by looking for similar values to those seen at a set of correlated reference time series at time $\bar{t}$. To choose the set of reference time series TKCM ranks all time series s_j ($s_j \neq s_i$) according to their similarity (or correlation) with s_i. A popular correlation measure is the Pearson Correlation Coefficient (PCC) [1], defined as follows:

$$\text{PCC}(s_i, s_j) = \frac{\sum_{t=\underline{t}}^{\bar{t}}(s_i(t) - \bar{s}_i)(s_j(t) - \bar{s}_j)}{\sqrt{\sum_{t=\underline{t}}^{\bar{t}}(s_i(t) - \bar{s}_i)^2}\sqrt{\sum_{t=\underline{t}}^{\bar{t}}(s_j(t) - \bar{s}_j)^2}}$$

where $\bar{s}_i$ and $\bar{s}_j$ is the mean of time series s_i and s_j, respectively. Since PCC measures only the linear correlation between two time series and TKCM also is used if the time series are non linearly correlated, we propose the Case Matching Similarity (CMS) to measure even non-linear correlation between time series.

CMS splits the *range* of time series s_j into equal-sized sub-ranges, called buckets. Each bucket contains the values of $s_i(t)$ such that $s_j(t)$ is within the bucket limits. Then CMS computes the

amount of variation within each bucket. The smaller the variation is, the more similar are the values $s_i(t)$ in a bucket and hence the stronger is the co-occurrence.

Let $w \in \mathbb{R}_{>0}$ denote the bucket width and $z \in \mathbb{Z}$ the ID of a bucket. Then a bucket b_z a list of the values $s_i(t)$ such that $s_j(t)$ is in the range $[zw, (z+1)w)$. More formally,

$$b_z = \{s_i(t) \mid t \in W \ \wedge \ zw \leq s_j(t) < (z+1)w\}$$

Moreover, we define the bucket mean μ_z and the bucket standard deviation σ_z as follows

$$\mu_z = \frac{1}{|b_z|} \sum_{s_i(t) \in b_z} s_i(t) \qquad \sigma_z = \sqrt{\frac{1}{|b_z|} \sum_{s_i(t) \in b_z} (s_i(t) - \mu_z)^2}$$

Each bucket b_z has a certain standard deviation σ_z, which is a measure of variation (or dispersion) inside the bucket. The smaller σ_z is, the closer are the values $s_i(t) \in b_z$ to the bucket mean μ_z.

Let $B = \{b_z \mid \forall z \in \mathbb{Z} : \ b_z \neq \emptyset\}$ be the set of all non-empty buckets. Notice that the buckets in B cover the entire range of $s_j(t)$. Then CMS is defined as the average bucket standard deviation, where each term is weighted by the number of elements in the corresponding bucket.

$$\text{CMS}(s_j, s_i) = \frac{1}{|B|} \sum_{b_z \in B} \frac{|b_z|}{|s_j|} \sigma_z$$

Tasks

- Compile a database of weather data from MeteoSwiss. The data set should comprise at least 50 time series, each 10 years long. If available, not only temperature data should be collected, but also humidity and other parameters.
- Get to know the data set and look for at least two interesting weather phenomena like e.g. the Föhn, which is a warm wind that can cause sudden and very noticeable changes in temperature. Ideally try to find weather phenomena where you think that either PCC or CMS have problems with.
- To deepen your understanding of PCC, formally show the connection between PCC and linear regression. More specifically, start with the equation of linear regression and show that PCC can be interpreted as quality measure for the obtained regression line.
- Implement PCC and CMS, preferably in the C programming language.
- Adapt PCC and CMS such that they can be incrementally computed when the sliding window advances.
- Test your implementations on the time series database that you compiled before, especially on weather phenomena like the Föhn. Find situations where PCC and CMS work well and when they fail.
- Conduct experiments to study the impact of the size of the sliding window on the ranking of the time series.
- Analyze the runtime complexity and memory consumption of PCC and CMS.
- Summarize your findings in a detailed report.

Optional Tasks

- Integrate your algorithms into TKCM [2] for finding the reference time series.

References

[1] A. Mueen, S. Nath, and J. Liu. Fast approximate correlation for massive time-series data. In *SIGMOD'10*, pages 171–182, New York, NY, USA, 2010. ACM.

[2] K. Wellenzohn, M. Böhlen, A. Dignös, J. Gamper, and H. Mitterer. Continuous imputation of missing values in highly correlated streams of time series data. Unpublished, 2016.

Supervisor: Kevin Wellenzohn (wellenzohn@ifi.uzh.ch)

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor