STATISTICAL LEARNING SYSTEMS

LECTURE 15: MINING MASSIVE DATA SETS

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Time series

In this lecture we confine ourselves to some issues which pertain to mining, say, terabyte sized time series or, rather, to naming some problems and hinting to their (or the existence of their) solutions.

The importance of retrieval and similarity search in time series data can hardly be overemphasized. The first issue which comes to mind is how to measure this similarity: apparently better by dynamic time warping (DTW) than by an L_p (Euclidean in particular) distance, but already the latter, let alone the former seems computationally demanding.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Time series

A remark on DTW: To align two sequences using DTW, an n-by-m matrix is constructed for two series, Q and C of lengths n and m respectively, with the (i-th, j-th) element of the matrix being the Euclidean distance $d(q_i, c_j)$ between the points q_i and c_j. A warping path P is a contiguous set of matrix elements that defines a mapping between Q and C. The t-th element of P is defined as $p_t = (i, j)_t$, so we have (clearly, **our goal is to find the shortest path**):

$$P = p_1, p_2, \cdots, p_t, \cdots, p_T, \quad \max(m, n) \leqslant T \leqslant m + n - 1,$$

with additional condition that

$$p_1 = (1, 1) \ \textit{and} \ p_T = (m, n).$$

And, however hard the problem seems, Rakthanmanon, Campana, Mueen, Batista, Westover, Zhu, Zakaria and Keogh show how to perform *Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping* (see SIGKDD 2012).

Time series

It has been emphasized in the paper mentioned that arbitrary query lengths cannot be inedexed. If, however, we know the length of queries ahead of time, we have, e.g., the following way out of the trouble:

iSAX: Indexing and mining terabyte sized time series (Shieh i Keogh, SIGKDD 2008)
and
iSAX 2.0: Indexing and Mining One Billion Time Series (Camerra, Palpanas, Shieh and Keogh, ICDM 2010),

where SAX (or Symbolic Aggregate approXimation) rests on firstly replacing the series by its piecewise constant approximation, secondly taking this approximation as an input and discretizing it into a small alphabet of symbols with a cardinality of size **a**, and finally building on this foundation a tree-structured index.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Time series

One more issue regarding long time series: whether we aim at clustering or classification, we can be better off using only some local patterns and deliberately ignoring the rest of the data; see
Clustering Time Series using Unsupervised-Shapelets by Zakaria, Mueen and Keogh (ICDM 2012)
and
Time Series Classification under More Realistic Assumtions by Hu, Chen and Keogh (SDM 2012).

Much more valuable research on DM for terabyte size time series is being done at UCR (Keogh's group) and elsewhere. At this juncture let us, however, mention two now classical warnings by Eamonn Keogh and Jessica Lin, and Shruti Kasetty, respectively:
Clustering of Time Series Subsequences is Meaningless: Implications for Past and Future Research (2003)
and
On the On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration (2002)

Mining massive data sets: topics which should be covered

Anand Rajaraman and Jeff Ullman, who have taught a course on Web mining at Stanford for several years, have produced an introductory textbook titled *Mining of Massive Datasets*. The topics covered by the textbook are perfectly chosen and they include:

- Distributed file systems and map-reduce as a tool for creating parallel algorithms that succeed on very large amounts of data.
- Similarity search, including the key techniques of minhashing and localitysensitive hashing.
- Data-stream processing and specialized algorithms for dealing with data that arrives so fast it must be processed immediately or lost.
- The technology of search engines, including Google's PageRank, link-spam detection, and the hubs-and-authorities approach.
- Frequent-itemset mining, including association rules, market-baskets, the A-Priori Algorithm and its improvements.
- Algorithms for clustering very large, high-dimensional datasets.

In lieu of a coda (my favorite title for a section which concludes a survey or a course)

- Much of most important and interesting material which belongs to the field of statistical learning has been covered by Szymon Jaroszewicz's course on Advanced Topics in ML (including graphical models broadly understood, in particular with CRF's also dealt with, and statistical relational learning).
- Due to lack of time, some important subfields of statistical learning have only been hinted to or treated in a very selective way; in particular, the following most immediate omissions have been made:
  - within unsupervised learning, random projections for dimension reduction;
  - within the area of semi-supervised learning, co-training paradigm and many more;
  - within the area of supervised learning, the issue (particularly important when dealing with data streams) of incremental learning (e.g., Very Fast Decision Trees such as Hoeffding trees), polycategorical learning, learning from imbalanced data, orthogonal decision trees.

Thank you

THANK YOU AND GOOD LUCK!

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund