Spatio-Temporal Data Mining (Part II)

Donato Malerba and Annalisa Appice
Goals

• We will
  • Formulate the problem of learning with spatio-temporal data acquired in real-time through a number of (wireless) remote sensors
  • Clarify the role of various forms of correlation
    • Spatial-, temporal, multivariate correlation
  • Establish links with descriptive and predictive tasks
    • Summarization, interpolation, forecasting, anomaly detection)
Goals

• We will:
  • Point to important results
    • Algorithmic/experimental
    • Applications
  • Describe algorithms and techniques
  • Present open problems
Assumed Background

• We assume a basic knowledge of machine learning methods for clustering, classification and regression tasks
  • For background, please see Mitchell (1997)
• Basic knowledge of spatial statistics, time series analysis
Outline

• Data and tasks

• Issues and challenges

• Univariate learning
  • Summarization, interpolation, anomaly/change detection, forecasting

• Beyond the univariate case – multivariate case
  • Summarization vs interpolation

• Open challenges
Sensor networks

• A set of (wireless) sensor stations $K$ which monitor an environment by collecting geophysical data (temperature, humidity, light,...)

• Each node in a sensor network can be imagined as a small computer, equipped with the basic capacity to sense, process, and act

• Sensors act in dynamic environments, often under adverse conditions
From sensor networks to geophysical time series

- A sensor network is scattered in a (possibly large) region where it is meant to collect data through its sensor nodes.
- Every sensor $k \in K$ measures a space of geophysical fields $\mathbf{Z} = (Z_1, \ldots, Z_m)$ repeatedly at the time points of $T$.
- It feeds a time series of data snapshots $(\mathbf{Z}, t_1, K)$, $(\mathbf{Z}, t_2, K)$, ...
Applications

• Typical applications of sensor networks include monitoring, tracking, and controlling
• Some of the specific applications are photovoltaic plant controlling, habitat monitoring, traffic monitoring, and ecological surveillance
Online data

• Sensor networks for climate data
  • National Oceanic and Atmospheric Administration (NOAA) data (http://www.ncdc.noaa.gov/)

• Sensor networks for ecology
  • NatureServe data (http://www.natureserve.org/biodiversity-science/conservation-topics/data)

• Sensor networks for energy markets
  • National Renewable Energy Lab data (http://www.nrel.gov/)
  • PVGIS (http://photovoltaic-software.com/pvgis.php)
Data scenario

• Time series of data that are measured repeatedly over a set of sensor stations.
  • The spatial location of a sensor station is modeled by means of point coordinates (e.g., latitude and longitude).
  • The spatial locations of the sensors are known, distinct and invariant, while the number of recording sensors may change in time: a sensor may be inactive and transmit no data for a time interval.
  • Active sensors transmit measurements for a number of numeric variables (multi-variate data) and they are synchronized in the transmission time.
Tasks

- Training on (incomplete) past data snapshots \((Z,t_1,K), (Z,t_2,K), ..., (Z,t_n,K)\), in order to
  - Predict on missing data in some data snapshot \((Z,t_i,K)\) with \(1 \leq i \leq n\) -- INTERPOLATION
  - Forecast some next data snapshot \((Z,t_i,K)\) with \(i > n\) -- FORECASTING
  - Perform anomaly/change detection in the last data snapshot \((Z,t_n,K)\) -- ANOMALY/CHANGE DETECTION
Issues & challenges

• Spatial dimension:
  → Inferences on spatial correlation: how data taken at a relatively close location behave similarly to each other (proximity relation, network structure, local & global indexes, non stationarity)

• Temporal dimension:
  → Inferences on the temporal correlation: how many future observations can be predicted from past behavior (stationarity vs variation, concept drift, anomalies)

• Multivariate data:
  → Inferences on cross-correlation of variables measured at the same site, as well as at close sites
Additional issues & challenges

• Huge volume of data which cannot be entirely recorded for future analysis.
  • Computing data aggregates, discarding real data, using data aggregates in future analysis.

• Sensed data must be processed on-line (patterns are computed in (near) real time).

• Computation can be distributed in-network
Tasks

- Summarization
- Interpolation
- Forecasting
- Anomaly & change detection

Descriptive task + Predictive task


- Predictive clustering (H. Blockeel et al., ICML 1998; D. Stojanova et al., Ecological Informatics 2013; S. Dzeroski et al., KDID 2006)
Summarization

“Derive a compact representation of data for storage”
Summarization

• Sampling (uniform vs stratified): (S. Acharya et al., SIGMOD 2000)

• Discrete Fourier Transform: signal processing technique (Y. Zhu & D. Shasha, VLDB 2002)

• Histograms (optimal, equal-width, end-biased): summary structures used to capture the distribution of values (A.C. Gilbert, STOC 2002; M. Greenwald & S. Khanna, ACM SIGMOD Rec 2001; F. Furfaro et al., Knowl Inf Syst 2008)
Summarization

• Sketches: approximation algorithms which allow the estimation of frequency moments and aggregates over joins (N. Alon et al., STOC 1996; J. Hershberger et al., Algorithmica 2006)

• Wavelets: projection of a sequence of data onto an orthogonal set of basis vectors (N. Alon, STOV 1996, Y. Matias, VLDB 2000)

• SAX: reduction of a numeric time series to a string of arbitrary length (J. Lin et al., Data Min Knowl Discov, 2007)

Summarization in sensor network analysis

- Centric summarization: deployed on the server station of networks by aggregating (spatial and/or temporal) correlated data.
  - Spatial cluster analysis snapshot by snapshot (X. Ma et al., APWeb/WAIM 2007)
  - Temporal cluster analysis sensor by sensor (P.P. Rodrigues et al., ECMLPKDD 2008)
Summarization in sensor network analysis

- In-network summarization: sensor on-board summarization of data, only the summary is transferred to the centralized station → data communication and energy usage can be minimized
  - sampling, k-means or wavelet computed on the sensor (R. Chiky and G. Hébrail, DaWaK 2008)
  - simple aggregates (sum, count, histogram), computed along a tree-coordinating schema (Z. Chen, WTS 2010)
  - Spatio-temporal clustering (data which are autocorrelated both in space and time computed along a tree-coordinating schema (S. Yoon and C. Shahabi, ACM Trans Sens Netw 2007)
A trend cluster (Ciampi et al, KES 2010) is a triple:
\[ [ W;C; T ] \]

where:

1. \( W \) is a **time horizon** along which field data were collected;

2. \( C \) is a **cluster** of spatially close sensors which transmitted values whose temporal variation was similar along the time horizon of the window;

3. \( T \) is the **time series** which represents the trend of the clustered measures as they were collected at the transmission time points comprised in \( W \).
Trend cluster discovery
Trend cluster discovery

- Count-based window model - SUMATRA (Appice et al., DAMI 2015)

http://www.di.uniba.it/~appice/software/SUMATRATRECI/index.htm
South-America air climate

- Monthly mean temperature: 6477 sensors
Intel Berkeley Lab

- Temperature every 31 secs: 54 sensors
- $W=256$, $\delta=2.5$
Further work

- Trend compression (A. Ciampi et al., CIDM 2011) by:
  - Discrete Fourier Transform
  - Haar Wavelets
  - Sampling
  - Least Square regression

- On-line selection of a trend compression technique (A. Appice et al., DAMI 2015)

Intel Berkeley Lab, $W=256$, $\delta=2.5$
In-network trend cluster discovery

- SUMATRA on the bottom-level sink nodes of a tree-based WSN
- metaSUMATRA on the top-level sink nodes
- A top-level sink:
  1. receives trend clusters from its child sinks
  2. gathers together received trend clusters
  3. propagates (meta) trend clusters to the parent sink

(A. Appice et al., DAMI 2015)
### South-America air climate

<table>
<thead>
<tr>
<th>Tree level</th>
<th>Role</th>
<th>in-SUMATRA</th>
<th>SUMATRA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg n. clusters</td>
<td>kbytes</td>
</tr>
<tr>
<td>5</td>
<td>Sensing device</td>
<td>9,833</td>
<td>9,833</td>
</tr>
<tr>
<td>4</td>
<td>Bottom sink</td>
<td>136.7</td>
<td>601</td>
</tr>
<tr>
<td>3</td>
<td>Top sink</td>
<td>65.5</td>
<td>493</td>
</tr>
<tr>
<td>2</td>
<td>Top sink</td>
<td>44.5</td>
<td>461</td>
</tr>
<tr>
<td>1</td>
<td>Server root</td>
<td>31.8</td>
<td>441</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11,384</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>39,824</td>
</tr>
</tbody>
</table>
Interpolation

“Supplement, smooth and standardize observational data”
Interpolation - spatial

• Inverse distance weighting: to calculate an unknown field value in a geographic location based on the degree of similarity in a neighbourhood (D. Shepard, ACM 1968)


• Kriging (N. Cressie, 1993): to calculate an unknown field value in a geographic location based on a linear combination of data in a neighbourhoods. Weights are based on the computation of a variogram. The variogram represents an approximate measure of the spatial dissimilarity of the observed data
Interpolation - spatial

• Kriging is more complex than IDW,
  • The variogram computation cost scales as the cube of the number of observed data
  • Kriging is highly dependent on a reliable estimation of the variogram
• but it has the undeniable advantage of computing the best linear unbiased estimator of the correlation model
• However, the accuracy of an IDW interpolator often approaches the accuracy of a Kriging interpolator, especially for smooth fields (G.Y. Lu & D.W. Wong, Journal of Computers and Geosciences, 2008)
Interpolation – spatio-temporal

• First performing spatial interpolation and then reducing temporal interpolation to the application of simple methods (such as linear or spline interpolation) to the sequence of snapshots of spatially interpolated data (L. Li et al., SARA 2011)

• First interpolating time series of data for each relevant location and then using them as sampled observations for the application of a traditional spatial interpolator (L. Li, GIS: Exploring Data for Decision Making, 2009)
Interpolation – spatio-temporal

- The true integration of the spatial and temporal data component essentially based on the application of a dynamic model, like the Kalman filter or the Markov Random field, to consecutive snapshots of data
  - e.g. Kriging + Kalman Filter (W.S. Kerwin & J.L. Prince, IEEE Transactions on Signal Processing, 1999)
- Non-stationary time series analysis (trend and armonic component + spatiotemporal model of log-transformed data are computed. The model consists of trend and noise and represents the spatiotemporal variations (R. Romanowicz et al., Environmental Modeling Software, 2006)
Interpolation – trend cluster

• For each trend cluster (A. Appice et al., J. Spatial Information Science 2013),
  • Extract a shape-dependent (quadtree-based) sample of clustered sensors (key sensors)
  • Determine a (polynomial) regression model of the time law underlying the trend time series

• Key sensors and regression coefficients (trend) are stored
Interpolation – trend cluster

- IDW applied to regression polyline-based predictions determined for the key sensors, at the time t

http://www.di.uniba.it/~appice/software/SUMATRATRECI/index.htm
Interpolation – trend cluster

- Trend cluster + Kriging (Guccione et al., MSM/MUSE 2011)
  - Trend cluster discovery to reduce the amount of data to mine for the variogram estimation
  - Trend cluster discovery + transfer learning to adapt the variogram learned at a time along the trend time series
### South America air climate

<table>
<thead>
<tr>
<th>SUMATRA</th>
<th>Sumatra + Sampling</th>
<th>TRECI</th>
</tr>
</thead>
<tbody>
<tr>
<td>average D</td>
<td>548.6</td>
<td>200.1</td>
</tr>
<tr>
<td>size (Kbytes)</td>
<td>548.6</td>
<td>200.1</td>
</tr>
<tr>
<td>rmse</td>
<td>1.25</td>
<td>1.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline</th>
<th>50% Sensors switching-off</th>
<th>50% Time points jumping-on</th>
<th>50% Sensors switching-off and 50% time points jumping-on</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC+IDW</td>
<td>1.97</td>
<td>2.48</td>
<td>2.72</td>
</tr>
<tr>
<td>TC+Kriging</td>
<td>1.94</td>
<td>2.08</td>
<td>-</td>
</tr>
</tbody>
</table>
Anomaly/Change detection

“Detect exceptional (anomaly) or stable (change) abnormal behaviour in data”
Anomaly detection

Change detection

• Gradual changes (drift) vs and abrupt changes (shift) in the data distribution
Anomaly and change detection

• A local polynomial fitting method + forward and backward prediction errors, (Z. Li et al., PAKDD 2007)

• Model fitting + outlier detection + quarantene to identify changes (M. Pechenizkiy et al., SIGKDD Explor. 2009)

• A change is alerted in the presence of outliers detected simultaneously in a snapshot (Bakker et al., KDD 2009)
Anomaly and change detection - trend cluster

- Performing an incremental modeling phase of a geophysical data stream by accounting for spatial and temporal autocorrelation of data
- Detecting outliers (data which do not conform the model)
- Classifying outliers in anomalies and change points
  1. Correcting anomalies
  2. Changing the data model when change points are met

(Appice et al., Information Science 2014)

http://www.di.uniba.it/~appice/software/SWOD/index.htm
Anomaly and change detection - trend cluster
Anomaly and change detection - trend cluster

South American Climate network (SACN): monthly-mean temperature ([-7.6 to 32.9]°C) form 6477 sensors between 1960 and 1990.


<table>
<thead>
<tr>
<th>Nr.TC</th>
<th>SWOD</th>
<th>TSA (ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B%</td>
<td>H%</td>
</tr>
<tr>
<td>SACN</td>
<td>311.8</td>
<td>3.914</td>
</tr>
<tr>
<td>IBLN (w = 32)</td>
<td>8.5</td>
<td>13.572</td>
</tr>
<tr>
<td>IBLN (w = 64)</td>
<td>6.0</td>
<td>16.687</td>
</tr>
<tr>
<td>IBLN (w = 128)</td>
<td>7.1</td>
<td>23.840</td>
</tr>
</tbody>
</table>
Forecasting

“Predict the future”
Time series analysis

- Exponential smoothing model: averages a time series up to the current sample by Brown (more weight to recent data), Holt (correcting a linear tendency in the trend part), Winters (assuming the seasonality) models (A.C. Harvey, 1989)

- ARIMA – family model: linear combinations are determined through the estimation of the autocorrelation function of the time series (ar, arma, arima, auto.arima) (G. E. P. Box and G. M. Jenkins, 1994)

- Multi-variate AR model: linear model of multiple time series used to forecast the target (var) (Lutkerpohl, 2005)
Spatial-aware forecasting

- Pokrajac and Obradovic (2001) have extended ARIMA by adding a term of auto-regressive disturbance, in order to model spatial-temporal correlation of residuals over defined neighborhood structures.

- Kamarianakis et al. (2005) have extended ARIMA, in order to account for the property of spatial correlation by expressing each data point at the time point t and the location (x;y) as a linearly weighted combination of data lagged both in space and time.

- Saengseedam and Kantanantha (2014) have used linear mixed models (LMMs) with spatial effects under a Bayesian framework.
Spatial-aware forecasting

• Luna and Genton (2005), as well as Barbosa et al. (2006) have explored the possibility of learning a VAR model from a vector of variables composed of the same variable observed at neighboring sites.
  • Sites are grouped in neighborhoods according to user-defined specifications.

• Pravilovic et al (2013, 2014) addressed the forecasting task with ARIMA-family models by dealing with spatial and temporal correlation when choosing parameters $p,d$ and $q$ and determining coefficients $\phi$ and $\sigma$

• Pravilovic et al (2014) integrated the spatio-temporal clustering analysis and forecasting in the same learning process
## Spatial-aware ARIMA

<table>
<thead>
<tr>
<th>Data title</th>
<th>Phenomenon</th>
<th>auto.Arima</th>
<th>auto.Arima$\pi$</th>
<th>sARIMA$\pi$</th>
<th>cARIMA$\pi$</th>
<th>cARIMA$\pi$</th>
<th>cARIMA$\pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha = 0.05$</td>
<td>$\alpha = \text{est.}$</td>
<td>$\alpha = \text{local}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCEQ</td>
<td>Wind Speed</td>
<td>0.32</td>
<td>0.31</td>
<td>0.34</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Air Temperature</td>
<td>0.48</td>
<td>0.41</td>
<td>0.40</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Ozone Concentration</td>
<td>0.69</td>
<td>0.58</td>
<td>0.58</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>MESA</td>
<td>NO$_x$ Concentration</td>
<td>0.21</td>
<td>0.18</td>
<td>0.18</td>
<td>0.20</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>NREL</td>
<td>Wind Speed</td>
<td>0.39</td>
<td>0.39</td>
<td>0.41</td>
<td>0.41</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>SAC</td>
<td>Air Temperature</td>
<td>0.20</td>
<td>0.21</td>
<td>0.16</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>NREL/NSRDB</td>
<td>Global Solar Radiation</td>
<td>0.34</td>
<td>0.26</td>
<td>0.35</td>
<td>0.42</td>
<td>0.35</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Direct Solar Radiation</td>
<td>0.51</td>
<td>0.45</td>
<td>0.52</td>
<td>0.55</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Diffuse Solar Radiation</td>
<td>0.47</td>
<td>0.43</td>
<td>0.48</td>
<td>0.45</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>NCDC</td>
<td>Air Temperature</td>
<td>0.19</td>
<td>0.24</td>
<td>0.19</td>
<td>0.16</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Precipitation</td>
<td>0.26</td>
<td>0.26</td>
<td>0.27</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Solar Energy</td>
<td>0.19</td>
<td>0.22</td>
<td>0.16</td>
<td>0.15</td>
<td>0.23</td>
<td>0.19</td>
</tr>
<tr>
<td><strong>Overall Mean</strong></td>
<td></td>
<td>0.35</td>
<td>0.33</td>
<td>0.34</td>
<td>0.35</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Overall Median</strong></td>
<td></td>
<td>0.33</td>
<td>0.29</td>
<td>0.34</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Spatial-aware forecasting

- Time-series clustering + Predictive clustering
  - Temporal-based power wind forecasting (S. Pravilovic et al., DS 2014)

- Spatial neighborhood + ARIMA
  - Spatio-temporal power wind forecasting (V. Almeida & J. Gama, ICAIS 2014)

- Spatio-temporal adaptive neighborhood + Knn
  - Spatio-temporal based power wind forecasting (A. Appice et al., DS 2015)
Spatial-aware forecasting

• A distributed system for storing huge amounts of data, gathered from energy production plants and weather prediction services
  • HBase over Hadoop framework

• One-day ahead forecast of PV energy production based on Artificial Neural Networks (with both structured and non-structured output prediction) (M. Ceci et al., IDEAS 2015)
Multivariate case

“Dealing with cross-correlation of various geophysical fields”
State of the art - spatial

- Stojanova et al. (2013) propose computing the mean of global measures (Moran I and global Getis C), computed for distinct variables of a vector, as a global indicator of spatial autocorrelation of the vector by blurring cross-correlations between separate variables.

- Dray et al. (2006) explore the theory of the principal coordinates of neighbour matrices and develop the framework of Moran’s eigenvector maps.
  - They demonstrate that their framework can be linked to spatial autocorrelation structure functions also in multivariate domains.
State of the art - spatial

• Dray and Jombart (2011) propose:
  • a two-step procedure, where data are first summarized with PCA. In a second step, any univariate (either global or local) spatial measure can be applied to PCA scores for each axis separately
  • An approach that finds coefficients to obtain a linear combination of variables, which maximizes the product between the variance and the global Moran measure of the scores

• Appice & Malerba (2014): interpolative clustering
  • Model the spatial autocorrelation when collecting the data records of multiple geophysical variables in a sensor network
  • Use this model to compute compact tree-based summaries of actual data that are discarded
  • Inject computed summaries into predictive (IDW-based) inferences to yield accurate estimations of geophysical data at any space location
State of the art – spatio-temporal

• Time-evolving interpolative clustering (A. Appice & D. Malerba, DAMI 2014)
  • We look for the interpolative clusters, which manifest change in the property (mean and variance) of the spatial autocorrelation of the clustered data
  • We build again only sub-tree of the existence tree, which do not cluster the new records appropriately (i.e. leaf conditions are not satisfied on the new data snapshot)
### Time-evolving interpolative clustering

<table>
<thead>
<tr>
<th>Sensors on %</th>
<th><strong>TICT</strong></th>
<th><strong>TRECI</strong></th>
<th><strong>Wilcoxon test</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td><strong>RRMSE Time</strong></td>
<td><strong>Size</strong></td>
<td><strong>RRMSE Time</strong></td>
</tr>
<tr>
<td>GHCN 20</td>
<td>825.14 0.55</td>
<td>8,040.90</td>
<td>902.34 0.51</td>
</tr>
<tr>
<td>40</td>
<td>1,993.8 0.49</td>
<td>20,129.00</td>
<td>1,566.5 0.48</td>
</tr>
<tr>
<td>50</td>
<td>1,988.7 0.46</td>
<td>31,698.00</td>
<td>1,798.2 0.47</td>
</tr>
<tr>
<td>60</td>
<td>2,303.4 0.46</td>
<td>37,778.00</td>
<td>2,161.9 0.47</td>
</tr>
<tr>
<td>80</td>
<td>3,093.6 0.41</td>
<td>60,223.00</td>
<td>2,654.3 0.46</td>
</tr>
<tr>
<td>NDBC WAVE 20</td>
<td>5,544.9 0.90</td>
<td>233.20</td>
<td>7,900.4 0.67</td>
</tr>
<tr>
<td>40</td>
<td>9,443.4 0.68</td>
<td>408.00</td>
<td>12,719.0 0.70</td>
</tr>
<tr>
<td>50</td>
<td>11,625 0.67</td>
<td>486.40</td>
<td>15,621.0 0.76</td>
</tr>
<tr>
<td>60</td>
<td>13,98 0.67</td>
<td>507.20</td>
<td>16,799.0 0.67</td>
</tr>
<tr>
<td>80</td>
<td>18,766 0.71</td>
<td>691.00</td>
<td>25,043.0 0.64</td>
</tr>
<tr>
<td>NDBC WIND 20</td>
<td>6,330.1 0.96</td>
<td>280.40</td>
<td>24,309.0 0.96</td>
</tr>
<tr>
<td>40</td>
<td>10,307 0.97</td>
<td>495.20</td>
<td>31.4 1.01</td>
</tr>
<tr>
<td>50</td>
<td>12,668 0.94</td>
<td>476.80</td>
<td>47,201.0 1.05</td>
</tr>
<tr>
<td>60</td>
<td>14,879 0.94</td>
<td>612.80</td>
<td>41,598.0 0.99</td>
</tr>
<tr>
<td>80</td>
<td>19,574 0.95</td>
<td>806.80</td>
<td>61,375.0 0.95</td>
</tr>
<tr>
<td>SAC 20</td>
<td>407.38 0.20</td>
<td>4,813.90</td>
<td>154.0 0.26</td>
</tr>
<tr>
<td>40</td>
<td>816.8 0.15</td>
<td>12,074.00</td>
<td>209.77 0.26</td>
</tr>
<tr>
<td>50</td>
<td>1,036.7 0.14</td>
<td>17,500.00</td>
<td>256.46 0.24</td>
</tr>
<tr>
<td>60</td>
<td>1,204.5 0.13</td>
<td>21,997.00</td>
<td>225.55 0.28</td>
</tr>
<tr>
<td>80</td>
<td>1,600.2 0.12</td>
<td>35,003.00</td>
<td>284.76 0.27</td>
</tr>
<tr>
<td>SensorKDD 20</td>
<td>1,133.4 0.54</td>
<td>32,503.00</td>
<td>2,044.5 0.53</td>
</tr>
<tr>
<td>40</td>
<td>2,287.7 0.48</td>
<td>109,050.00</td>
<td>3,124.3 0.46</td>
</tr>
<tr>
<td>50</td>
<td>2,797.1 0.46</td>
<td>145,200.00</td>
<td>3,655.2 0.47</td>
</tr>
<tr>
<td>60</td>
<td>3,350.3 0.44</td>
<td>189,610.00</td>
<td>3,575.0 0.51</td>
</tr>
<tr>
<td>80</td>
<td>4,467.5 0.43</td>
<td>338,210.00</td>
<td>4,262.0 0.48</td>
</tr>
<tr>
<td>SensorScope 20</td>
<td>72,041 3.09</td>
<td>1,004.90</td>
<td>361.78 3.58</td>
</tr>
<tr>
<td>40</td>
<td>205.74 5.13</td>
<td>2,825.70</td>
<td>679.98 4.12</td>
</tr>
<tr>
<td>50</td>
<td>251.4 4.98</td>
<td>3,422.60</td>
<td>653.57 4.52</td>
</tr>
<tr>
<td>60</td>
<td>575.99 3.72</td>
<td>3,492.30</td>
<td>771.3 4.55</td>
</tr>
<tr>
<td>80</td>
<td>799.48 7.64</td>
<td>5,207.60</td>
<td>1,023.7 24.43</td>
</tr>
</tbody>
</table>
Open challenges

• Designing in-network spatio-temporal algorithms
• Integrating big data technologies
• Dealing with the covariance in autocorrelation measures of several variables
• Dealing with multivariate case in spatio-temporal forecasting, as well as in anomaly/change detection task
• Completing the bridge between time series analysis and stream data mining
Thank you for the attention
References


• V. Weber, Smart sensor networks: technologies and applications for green growth, OECD Conference on ICTs, the environment and climate change, 2009

• F. Fumarola, A. Appice, D. Malerba, A Business Intelligence Solution for Monitoring Efficiency of Photovoltaic Power Plants, ISMIS 2014, 518-523, 2014

• A. Appice, S. Pravilovic, A. Lanza, D. Malerba, Very Short-Term Wind Speed Forecasting Using Spatio-Temporal Lazy Learning, Discovery Science 2015, 9-16, 2015


• J. Yuan, Z. Wang, M. Han, Y. Sun, A lazy associative classifier for time series, Intelligent Data Analysis, vol. 19, no. 5, pp. 983-1002, Ios Press 2015


• M. Greenwald, S. Khanna, Space-efficient online computation of quantile summaries. ACM SIGMOD Rec 30(2):58–66, 2001

References


- A. Ciampi, A. Appice, D. Malerba, Summarization for Geographically Distributed Data Streams. KES (3) 2010: 339-348, 2010


References

• N. Cressie, Statistics for spatial data. Wiley, New York, 1993


• R. Klinkenberg, Learning drifting concepts: example selection vs. example weighting, Intell. Data Anal. 8, 281–300, 2004
References

- M. Pechenizkiy, J. Bakker, I. Zliobaite, A. Ivannikov, T. Karkkainen, Online mass flow prediction in cfb boilers with explicit detection of sudden concept drift, SIGKDD Explor. 11, 109–116
- Y. Kamarianakis, P. Prastacos, Space–time modeling of traffic flow, Computers & Geosciences 31 (2), 119–133, 2005
References


