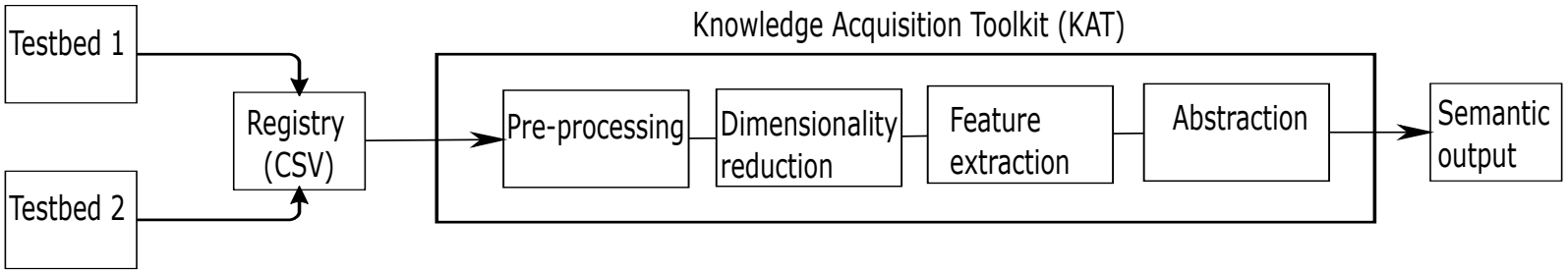
Data Analysis Tools For FIESTA

KAT Tool

University of Surrey

December 1, 2015

Overview



The proposed KAT toolkit

1. Pre-processing
2. Dimensionality reduction
3. Feature extraction
4. Abstraction

Pre-processing

- **Data selection:**
  - Select a desired category when there are several categories inside a data source, e.g. select temperature among the available information (e.g. wind speed, humidity, temperature, pressure) recorded at a weather station
  - Select a desired time-period[1], for instance days 3 to 10
- **Cleaning:** Check for the missing data, e.g. those occurred due to the sudden power cut-off or sensor malfunction
- **Resample:** To bring data from different sources into a unique shape for further processing[1]
- **Noise removal:** Remove transmission cost and filter unwanted data, such as noise and outliers. This can be done using different algorithms including Min-Max, Mean-Median, Variance or Bandpass filters
- **Normalisation:** To compare datasets with different offsets and amplitudes

[1] This feature would require a common label for 'time' in the registry output

Dimensionality reduction I

To reduce the size and length of the data while keeping the key features and patterns

- Non-data adaptive:
  - **Discrete Fourier Transform (DFT):**
    - ✗ DFT loses the time information of the data and transforms the data globally, which causes a slow calculation
  - **Discrete Wavelet Transform (DWT):**
    - ✓ DWT preserves the time dimension and transforms the data locally which leads to a faster calculation.
    - ✓ Wavelets have the useful multiresolution property, but are only defined for time series that are an integer power of two in length.
- Data adaptive:
  - **Piecewise Aggregation Approximation (PAA):**
    - ✗ PAA works with a fixed window length. In times of low data activity, the same event is aggregated over and over. In contrast, if there is a lot of activity, aggregation can lead to information loss.

Dimensionality reduction II

- **Symbolic Aggregate Approximation (SAX):** Transforms a time-series into a discretised series of letters referred to as a word.
  - ✓ The DWT,DFT, and PAA representations are real valued, and this limits the available algorithms. This limitation is addressed using symbolic representation methods such as SAX.
  - ✓ SAX algorithm provides both dimensionality and storage reduction (as fewer bits are required for letters and repetitive sequences can refer to a single store position). While general symbolic methods provide storage reduction, but not dimensionality reduction.
  - ✓ SAX can be also referred as a pattern creator technique to create abstract patterns and human/machine interpretable observations from the sensory data
  - ✓ If the standard deviation of the sequence before normalisation is below an epsilon, the entire word can be set to the middle-ranged alphabet

Feature extraction

To extract representative features from the sensor data and provide low-level abstractions in a local sensor data

- **Clustering (k-means):** Group samples with similar attributes into the same class. k-means algorithm only requires the expected number of groups

Abstraction

To provide higher level abstractions using the generate low-level abstraction to get the global picture about occurring events

- **Markov Chains:** Construct a model which represents likelihood of temporal relations between values

Modification of KAT tool

- Implemented
  - Selecting a desired time-period
  - Saving the data after every step of analysis
  - Additional normalisation method in the pre-processing step
- In progress
  - Perform cleaning and resampling[2] algorithms on all input data
  - Apply consecutive pre-processing algorithms

[2] This feature would require a common label for 'time' in the registry output

Main References

- Ganz, F., Puschmann, D., Barnaghi, P., & Carrez, F. (2015). A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things. IEEE Internet of Things Journal.
- Ganz, F., Barnaghi, P., & Carrez, F. (2014). Automated semantic knowledge acquisition from sensor data. IEEE Systems Journal.
- Ganz, F., Barnaghi, P., & Carrez, F. (2013). Information abstraction for heterogeneous real world internet data. IEEE Sensors Journal, 13(10), 3793-3805.

The End