Database Establishment for Machine Learning in NILM

Sean Lai*, Mark Trayer,
Sudhir Ramakrishna, and Ying Li
Samsung Dallas Technology Lab
*email: s.lai@samsung.com

Data and Machine Learning

- Learning begins with observing.
- Success of machine learning depends heavily on availability of data sets.
- Reference Energy Disaggregation Data Set (REDD) is one open data set for NILM.
  - High frequency whole home data +
  - Low frequency individual circuit and plug data.

On NILM Database for Machine Learning

- Data for machine learning:
  - Training set: to tune machine learning methods
  - Validation set: to select, check generalization properties of methods
  - Test set: do methods really work or not?
- Different types of data are suitable for different types of machine learning methods.
  - Temporal resolution, duration and so on
  - Whole home data, single appliances, multiple appliances and so on.
- Key database properties: Informative, diverse, and scalable

Informative

- Is the kind of information that a machine learning method needs available in the data?
  - High frequency sampling? E.g. 1k, 10k 1MHz?
  - Long duration? E.g.:
    - Covering the duration of events?
    - Whole Operation period?
    - Multiple days → behavioral information
  - Detailed labeling → diverse and scalable

Diverse

- Single appliance
  - Brand/manufacture of the device, e.g. Samsung, LG, HP and so on
  - Operational mode, e.g. popcorn, high heat for microwave
  - Environment parameters: heat, humidity and so on.
  - Geolocations
  - “Age” or malfunction of device
- Aggregate appliance data
  - What kinds of appliances/manufacture/modes
  - Timing and duration
- Whole home data

Scalable

- Good book keeping is necessary
- Unified format
- Pointers of data: user can quick retrieve
  - Specific appliance across datasets
  - Specific combination of appliance aggregation
  - Recording specifications: sampling rate, duration and so on

Data Set Types

- Single appliance data sets:
  - Essential for training most existing NILM methods.
  - Revealing the strength and weakness of a method within and across appliance categories
  - Proxy upper bound of the accuracy when extending the method for whole home data sets.
- Whole home main circuit data sets:
  - Best test set and most realistic.
- Intermediate aggregate appliance data sets:
  - Specifically focus on scenarios related to the error, e.g. specific combinations of aggregate appliance categories.
  - Guiding the synthesis of aggregate appliance data sets from single appliance data sets.