(Supplementary Material) SmartD: Smart Meter Data Analytics Dashboard *

Aylin Jarrah Nezhad, Tri Kurniawan Wijaya, Matteo Vasirani, and Karl Aberer
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{aylin.jarrahnezhad, tri-kurniawan.wijaya, matteo.vasirani, karl.aberer}@epfl.ch

1. SUPPLEMENT FOR SECTION 2.1

Consumer aggregations.

A simple example of a grammar which expresses consumer selection/aggregation: string "1; 2; 3, 4, 5" corresponds to the visualization of the energy data of consumer 1 alone, consumer 2 alone, and a cluster composed of consumers 3, 4, and 5.

2. SUPPLEMENT FOR SECTION 2.2

Estimating typical load profile.[1]

SmartD is able to estimate consumer typical load profile given her demographics and contextual information. Let D be the set of demographic information, and C be the context that we are interested in. The set of demographic information, D, can be, for example: a family with two children, live in 2000 sq. ft. apartment, and own a dishwasher. The context, C, can be, for example: weekdays in January, or Monday in the summer. In addition, let N be the set of $k \in \mathbb{N}$ consumers with the closest demographics to D. Thus, $|N| = k$. Then, the estimated load profile of consumers with demographics D on context C is the average of (hourly) load profile of consumers in N.

A question remains, however, to decide the best k. Should k be 1, 2, 3, or something else? To answer this, for each k under consideration, we perform *leave-one-out-cross-validation*. See Algorithm 2.1 for details. For load profiles L_i and L_j, function $dist(L_i, L_j)$ return the distance between L_i and L_j. It can be computed, for example, using the difference between the norm of L_i and L_j.

Algorithm 2.1: Find the best k

Input: a set of consumers $\mathcal{A}$, a set of k under consideration $\mathcal{K} = \{k_1, \dots k_n\}$, contextual information C
Output: the best $k \in \mathcal{K}$

foreach $k \in \mathcal{K}$ **do**
  $\delta_k \leftarrow 0$
  foreach $i \in \mathcal{A}$ **do**
    $\mathcal{A}' \leftarrow \mathcal{A} \setminus i$
    Let N be the set of k consumers in $\mathcal{A}'$ having the closest demographics to i
    $L_i \leftarrow$ (hourly) load profile of i
    $L_N \leftarrow$ average (hourly) load profile of consumers in N on context C
    $\delta_k \leftarrow \delta_k + dist(L_i, L_N)$
return $\arg\min_k(\delta_k)$

Discovering significant demographic characteristics.

SmartD is also able to infer demographic information that significantly influences energy consumption on a specific context, e.g., weekdays in January, or Monday in the summer. For this purpose, we use a supervised feature selection algorithm, namely *correlation-based feature selection*.[2] We refer to this algorithm as *cfs*.

Let an *instance* be a tuple (F, l), where $F = \{f_1, \dots, f_{|F|}\}$ is a feature set and l is a target attribute. Given a set of instances $\mathcal{I}$, applying *cfs* to $\mathcal{I}$ results in the set R of indexes of the features that are deemed to be relevant to the target attributes. Formally $cfs(\mathcal{I}) = R = \{r_1, \dots, r_{|R|}\}$ such that $r_i \subseteq \{1, \dots, |F|\}$.

Next, we explain how to infer top-q demographic characteristics which are relevant to the energy consumption for a context C. Let $D = \{d_1, \dots, d_{|D|}\}$ be the set of consumer demographics. We define F_i as the feature set of consumer i, where each of its element is consumer i's demographic information. Thus $|F| = |D|$. Let l_i^h be the average of hourly energy consumption of consumer i, on context C, at hour $1 \leq h \leq 24$. Further, let $\mathcal{A}$ be the set of consumers, and $\mathcal{I}^h$ be the set of instances for hour h, consist of tuples (F_i, l_i^h) for all consumers $i \in \mathcal{A}$.

For $1 \leq h \leq 24$, let $cfs(\mathcal{I}^h) = R^h$. Then, we define $score(r) = |\{R^h \mid r \in R^h, 1 \leq h \leq 24\}|$, for $1 \leq r \leq |F|$. The top-q demographic characteristics of the set of consumers $\mathcal{A}$ on context C is the q demographics $d_{r_1^*}, \dots, d_{r_q^*}$ with the highest scores. That is, the top-q demographics are $d_{r_1^*}, \dots, d_{r_q^*}$, where $score(r^*) \geq score(r)$ for all $r^* \in \{r_1^*, \dots, r_q^*\}$ and $r \in \{1, \dots, |F|\} \setminus \{r_1^*, \dots, r_q^*\}$.

[1] We use the terms *energy consumption* and *load* interchangeably.
[2] See the bibliographic information for this method in the main paper.