(Supplementary Material)
Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data *

Tri Kurniawan Wijaya † Tanuja Ganu ‡ Dipanjan Chakraborty‡ Karl Aberer†

Deva P. Seetharam‡

1 Supplement for Automatic Cluster Configuration Selection in Section 3.2.5

We give a brief description about the Silhouette [3], Dunn [2], and Davies-Bouldin [1] indices. They provide us a way to compare a cluster configuration from one to another. However, there are some differences.

Let x be a consumer, C be a cluster configuration (set of clusters), and $C(x) \in C$ be the cluster of x. In addition, let $dist(x, x')$ be the distance between two consumers x and x'.

The Silhouette index This index determines how well an object is clustered, based on the difference in the dissimilarity of the object to its cluster and to the other clusters.

Let $dist(x, c)$ be the average distance between x and all consumers in c, i.e.,

$$dist(x, c) = \frac{1}{|c|} \sum_{x' \in c} dist(x, x').$$

Let $a(x)$ be the average dissimilarity of consumer x to all other fellow cluster members in $C(x)$, i.e,

$$a(x) = \frac{1}{|C(x)| - 1} \sum_{\substack{x' \in C(x) \\ x' \neq x}} dist(x, x').$$

Assuming that $dist(x, x) = 0$, then we can also rewrite the equation above into:

$$a(x) = \frac{dist(x, C(x))}{|C(x)| - 1}.$$

Let $b(x)$ be the minimum average dissimilarity between x and other clusters, i.e.,

$$b(x) = \min_{c \neq C(x)} \frac{dist(x, c)}{|c|}.$$

Then, we define the Silhouette value of x as:

$$silh(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

The Silhouette index of a cluster configuration is the average of the Silhouette index of all consumers (in the configuration):

$$silh(C) = \frac{1}{|C|} \sum_{c \in C} \Big(\frac{1}{|c|} \sum_{x \in c} silh(x) \Big)$$

Silhouette index range from -1 to +1. The closer it is to 1, the better.

The Dunn index This index seeks the largest inter-cluster distance and the lowest intra-cluster distance. The Dunn index is computed based on the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance.

Let us define the inter-cluster distance between two clusters, c_1 and c_2, as the minimum distance between any two points in c_1 and c_2, i.e.,

$$interdst(c_1, c_2) = \min_{\substack{x_1 \in c_1 \\ x_2 \in c_2}} dist(x_1, x_2),$$

In addition, we define the intra-cluster distance (or *diameter*) of a cluster c, as the maximum distance between any two points in c, i.e.,

$$dia(c) = \max_{x_1, x_2 \in c} dist(x_1, x_2).$$

Then, we define the Dunn index of a configuration C as:

$$dunn(C) = \frac{\displaystyle\min_{\substack{c_1, c_2 \in C \\ c_1 \neq c_2}} interdst(c_1, c_2)}{\displaystyle\max_{c \in C} dia(c)}.$$

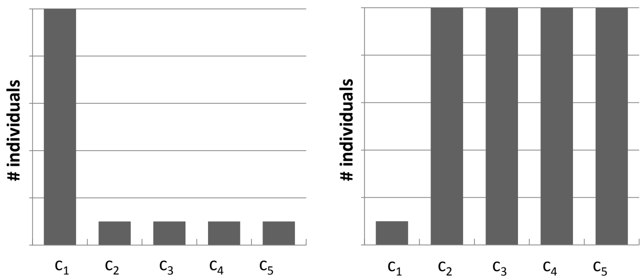
The larger the Dunn index, the better.

The Davies-Bouldin index This index is similar to the Dunn index, i.e., it aims to indentify a cluster configuration which has the largest inter-cluster distance and the lowest intra-cluster distance. The Davies-Bouldin index is computed

*Supported by European Union's Seventh Framework Programme (FP7/2007-2013) 288322, Wattalyst.

†School of Computer and Communication Sciences, EPFL, Switzerland. {tri-kurniawan.wijaya,karl.aberer}@epfl.ch

‡IBM Research India. {tanuja.ganu,cdipanjan,dseetharam}@in.ibm.com



(a) Feature f_1 (b) Feature f_2

Figure 1: Feature f_1 is discriminative positive for cluster c_1, whereas f_2 is discriminative negative for c_1. While entropy measure is able to recognize only discriminative positive features, our *discriminative index* is able to recognize both, discriminative positive and negative features.

based on the sum of diameter between two clusters divided by their inter-cluster distance:

$$daviesBouldin(C) = \frac{1}{|C|} \sum_{c_1 \in C} \max_{\substack{c_2 \in C \\ c_2 \neq c_1}} \frac{dia(c_1) + dia(c_2)}{interdst(c_1, c_2)}$$

In this case, we define the intra-cluster distance of a cluster c as the average distance of the cluster members to its centroid, i.e.,

$$dia(c) = \frac{1}{|c|} \sum_{x \in c} dist(x, \zeta^c),$$

where ζ^c is the centroid of cluster c. We define the inter-cluster distance to be similar with the one used for computing the Dunn index. Note that we define the Davies-Bouldin index here a little bit different compared to its original version [1]. However, as long as $dist$ is a proper distance metric, our definition satisfies Definition 1 to 5 in [1]. The lower the Davies-Bouldin index, the better.

2 Supplement for Section 5

An alternative to discriminative index Entropy can be used as an alternative to our *discriminative index* for determining whether a certain consumer characteristics is discriminative or not, using the same idea as in the decision tree learning. However, there is a subtle difference.

Using entropy, a feature is said to be discriminative for a particular class (or cluster, in our case) when it has low entropy. In Figure 1, f_1 has low entropy, and hence it is discriminative. That is, f_1 is an appropriate feature to distinguish cluster c_1 from others. Moreover, f_1 as an example of what we called as a *discriminative positive* feature. Feature f_2 in Figure 1, has high entropy. Thus, according to the entropy measure, f_2 is not discriminative. However, we can see that f_2 is actually also a discriminative feature, i.e., it characterizes an individual which does not belong to c_1 (might belong to any other clusters). Feature f_2 is an example of what we called as a *discriminative negative* feature.

While entropy is useful measure to recognize discriminative positive feature, it does not recognize discriminative negative feature. Our *discriminative index*, on the other hand, is able to distinguish both, discriminative positive and negative features.

3 Supplement for Section 6.3

Compared to appliance usage, information about appliance ownership is simpler and cheaper to obtain. Using questionnaire is enough to obtain the information whether a consumer own a certain appliance. Detailed appliance usage information, however, is more expensive to obtain because it involves sensor measurement.[1] Thus, knowing whether ownership of a particular appliance determines consumer's energy consumption is a valuable insight.

In our dataset, we have a set of question/answer whether a consumer own these appliances:

- washing machine,
- tumble dryer,
- dishwasher,
- electric shower,
- electric cooker,
- stand alone freezer,
- water pump,
- immersion,
- TV less than 21 inch,
- TV greater than 21 inch,
- desktop computer,
- laptop computer, and
- games consoles.

In Table 1 and 2, we show customer characteristics which related to appliance ownership only. Both shows how discriminative is an ownership of a particular appliance for different clusters, based on absolute consumption and consumption variability. Let $support$ be Z_c in case of discriminative positive and $Z_{\neg c}$ in case of discriminative negative. We show only characteristics with DI ≥ 0.6 (highly discriminative) and $support \geq 0.4$ (highly evident).

[1] Typical appliance usage, however, as in our dataset, can be obtained through questionnaire.

Table 1: Discriminative appliances' ownership for different clusters based on their absolute consumption. We show only for DI $\geq$ 0.60 and support $\geq$ 0.40. A minus (-) sign denotes discriminative negative.

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.76
2	games consoles	low	(-) yes	-0.70
3	tumble dryer	low	no	0.68
4	dishwasher	low	no	0.67
5	games consoles	high	yes	0.61

Table 2: Discriminative appliances' ownership for different clusters based on their consumption variability. We show only for DI $\geq$ 0.60 and support $\geq$ 0.40. A minus (-) sign denotes discriminative negative.

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.72
2	tumble dryer	high	(-) no	-0.72
3	tumble dryer	low	no	0.71
4	games consoles	low	(-) yes	-0.69
5	dishwasher	low	no	0.67
6	games consoles	high	yes	0.60

Over all appliances, we found that, only the ownership of big (power consuming) appliances (dishwasher and tumble dryer), which are highly discriminative. That is, the owner of these appliances are more likely to consume more energy and have higher consumption variability. The ownership of other appliances, which are not shown in Table 1 and 2, are less discriminative.[2]

The consistent presence of games consoles in both tables, however, is rather interesting since they are not big appliances (their power consumption is comparable to other electronic devices such as TV or desktop computer). We conjecture that the ownership of games consoles is highly correlated with family type, e.g., families with children are more likely to have games consoles at home compared to singles. Because family type is a highly discriminative characteristics for households' energy consumption behavior (see Table 1 and 2 in the main paper), then its correlation with games consoles ownership explains why games consoles ownership is also discriminative. Our conjecture is then confirmed in Figure 2, where it shows that, indeed, families with children are the most likely to own games consoles, followed by adults only families, and then by singles, who are the least likely to own games consoles.

[2]However, their usage pattern might be highly discriminative (such as washing machine, electric shower, water pump – see Table 1 and 2 in the main paper).

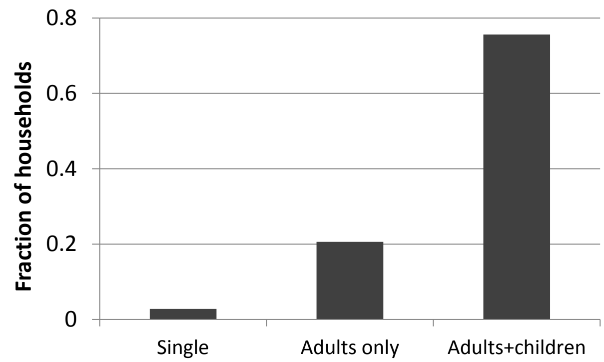


Figure 2: Fraction of households which own games consoles for different family types.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 288322 (Wattalyst). We also would like to the anonymous reviewers for their helpful comments.

References

[1] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[2] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[3] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.