EENG 415: Data Science for Electrical Engineering
Dr. Salman Mohagheghi
Spring 2023

Course Project 2: Customer Classification

In this project, you will be applying clustering and classification techniques to divide and classify a group of hypothetical utility customers based on their energy consumption patterns.

We will use the dataset available through the Research Support Facility (RSF) at the National Renewable Energy Laboratory (NREL). The dataset is provided on the course project's Canvas page but can also be accessed in [1]. RSF is equipped with multitude of sensors and actuators that allow for efficient energy management of the building. The data collected is also used to support research.

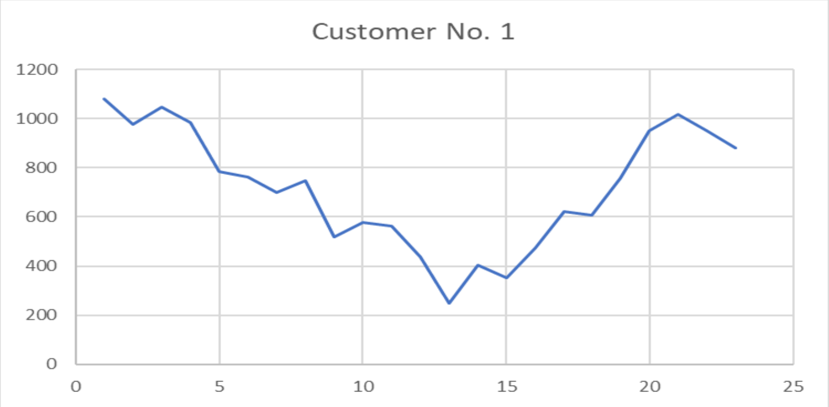
The dataset for this project includes the measured energy data from the RSF Systems Model and contains hourly data for total cooling (kW), total heating (kW), total mechanical (kW), total lighting (kW), total plug loads (kW), total data center (kW), total building (kW), PV (kW), and building net (kW). Hourly data is collected for an entire year, i.e., a total of 8,760 measurements.

A. Data Processing

For the purposes of this project, we will alter the dataset as follows:

- We will only consider the "building net" values and will assume that the units are in watts, not kW.
- We assume that measurement data for each day of the year, i.e., hour 12:00 am to 11:00 pm, represents the daily load profile of one "customer." This allows us to convert the annual dataset into 365 data instances, each representing a single "customer."

For instance, the following graph represents the hourly demand for Saturday, Jan. 1, 2011. In our analysis, this will be a data instance representing "customer" no. 1.



In practice, this data instance is a time series, and not a single data point. However, to make sure we can use the techniques learned in this class, you will need to convert it into a multi-dimensional data instance (if you would like to learn about data mining techniques for time series, please refer to chapter 14 in [2]). To do this, you would need to identify important attributes that explain the time series. As an example[1], one can replace the above time series with a 4-dimensional data instance that consists of:

- Peak daily demand (W)

[1] This is just an example and does not mean that you should use it as a template.

- Minimum daily demand (W)
- Difference between peak and minimum (W)
- Hour of the day when peak demand occurs

Work with your team to identify the best attributes that can explain the time series and provide justification for each one, i.e., why it is important. Your grade in this section will in part be based on the number of attributes you identify, the justification behind them, and how unique they are.

Once you have decided on the list of attributes, convert the time series dataset into a multi-dimensional dataset. If you have identified P attributes to explain the time series, you will obtain a 365×P matrix representing 365 "customers" with P attributes for each one. Provide this dataset (in csv format) along with your final report as one of the deliverables.

B. Customer Clustering

Your dataset consists of 365 daily load profiles, representing 365 "customers." Our goal in this part is to cluster this dataset into groups of customers that are similar to one another based on their energy consumption patterns.

The deliverables for this part are:

- Discussion on the clustering algorithm chosen, along with justification of why you think it's the best choice[2].
- If applicable, discussion on the parameters for the algorithm chosen[3].
- A visual representation of the clustering outcome.
- Performance metrics for validating the effectiveness of the clustering algorithm.
- Your MATLAB code.

C. Customer Classification

Now, consider the daily measurements from Jan. 1, 2011, to Nov. 30, 2011. Assume that this is your training data, i.e., combinations of individual customers with their attributes and the class (cluster) to which they belong. This will provide you with 334 data instances along with their class labels. Use this training dataset to develop a classifier. Once you have implemented the classifier, test it on the test set, which consists of the remaining 31 data instances[4].

The deliverables for this part are:

- Discussion on the classification model chosen, along with justification[5].
- If applicable, discussion on the parameters for the algorithm chosen.
- A visual representation of the classification outcome.
- Performance metrics for validating the effectiveness of the classifier.
- Your MATLAB code.

[2] You can also include a comparative analysis of different algorithms as part of your justification.

[3] For instance, if you have chosen *k*-means as your clustering algorithm, you would need to explain how you have decided on the parameter *k*.

[4] These are in fact the daily load profiles for the month of December.

[5] You can also include a comparative analysis of different algorithms as part of your justification.

Rubric

Category	Item	Max Points
Data Processing	Team has identified a reasonable number of attributes (at least 5) to represent the time series	2
	The choice of attributes is properly justified	3
	The choice of attributes is innovative and unique	2
	The modified dataset (in csv format) is submitted	1
Clustering	Team justifies the choice of the clustering algorithm chosen	3
	Clustering algorithm chosen is appropriate	1
	A visual representation of the clustering outcome is provided	1
	Appropriate validation metrics are used for assessing the effectiveness of the clustering algorithm	2
	MATLAB code is provided and runs with no issues	2
Classification	Team has correctly divided the dataset into training and test sets	2
	Team justifies the choice of the classification model chosen	3
	Classification model chosen is appropriate	1
	A visual representation of the classification outcome is provided	1
	Appropriate performance metrics are used for assessing the effectiveness of the classification model	2
	MATLAB code is provided and runs with no issues	2
Report	Report is well-written and professional	1
	Report (excluding the Appendix and any references) is less than 3 pages. Supplementary data or extra graphs may appear in the Appendix.	1
		30

References

[1] [Online]. Available at: https://data.openei.org/submissions/358.
[2] C.C. Aggarwal, *Data Mining – The Textbook*, Springer International Publishing, 2015.