EENG 415: Data Science for Electrical Engineering
Dr. Salman Mohagheghi
Spring 2024

Course Project 2: Customer Classification

Learning Objectives:

- Identify an appropriate clustering algorithm for a multidimensional dataset
- Apply a clustering algorithm to a multidimensional dataset
- Identify an appropriate classification model for a multidimensional dataset
- Apply a classification model to a multidimensional dataset
- Assess the performance of clustering and classification models

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 consumption data (in kW) for total cooling, total heating, total lighting, total plug loads, building net, etc. Hourly data is collected for an entire year, i.e., a total of 8,760 measurements.

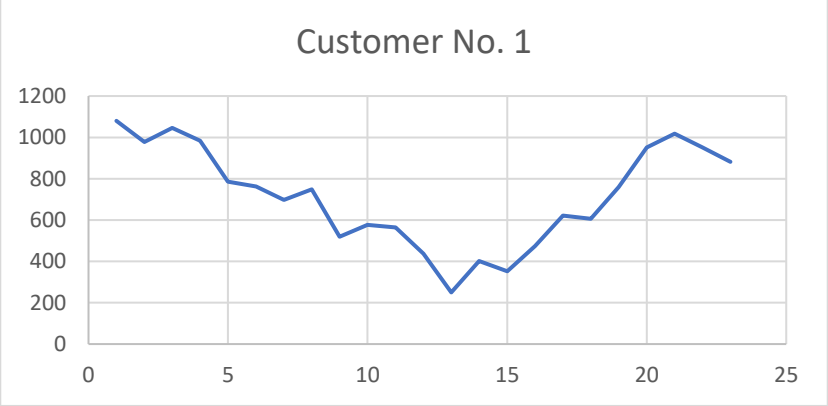
Your report should follow the exact headings listed below. Deliverables for each section are highlighted in blue. No additional deliverables or discussions are required.

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[1]. To do this, you would need to identify important attributes that explain the time series. As an example[2], one can replace the above time series with a 4-dimensional data instance that consists of:

- Peak daily demand (W)
- 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. 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.

Deliverables for this section are:

- List of attributes you have used to explain the time series, along with justification for each one,
- Multidimensional dataset in `xslx` or `csv` format.

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 (using the attributes you identified in the previous section).

The deliverables for this part are:

- Discussion on the clustering algorithm chosen, along with justification of why you think it's the best choice[3].
- If applicable, discussion on the parameters for the algorithm chosen[4].
- A visual representation of the clustering outcome.
- Performance metrics for validating the effectiveness of the clustering algorithm.

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 (that you obtained in the previous section). 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

[1] If you want to learn more about data mining techniques for time series, refer to chapter 14 in (Aggarwal, 2015).
[2] This is just an example and does not mean that you should use it as a template.
[3] You can also include a comparative analysis of different algorithms as part of your justification.
[4] 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*.

data instances[5].

The deliverables for this part are:

- Discussion on the classification model chosen, along with justification[6].
- 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.

Rubric

Category	Item	Max Points
Data Processing	The team has identified at least 5 attributes to represent the time series and has provided justification behind each one	5
	The choice of attributes is innovative and unique	2
	The modified dataset (in `xslx` or `csv` format) is submitted	1
Clustering	Clustering algorithm chosen is appropriate, and the team has justified its selection	3
	A visual representation of the clustering outcome is provided	2
	Validation metrics used for assessing the effectiveness of the clustering algorithm are appropriate and comprehensive	2
Classification	Classification model chosen is appropriate, and the team has justified its selection	3
	A visual representation of the classification outcome is provided	2
	Performance metrics used for assessing the effectiveness of the classification model are appropriate and comprehensive	2
Report	Report is well-written and professional	2
	Report (excluding the Appendix and any references) is less than 3 pages. Supplementary data or extra graphs may appear in the Appendix.	1
		25

References

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

[5] These are in fact the daily load profiles for the month of December.
[6] You can also include a comparative analysis of different algorithms as part of your justification.