Klasyfikacja urządzeń domoych po poborze prądu

Kamil Zaborniak, Tomasz Gładki

June 2024

Spis treści

1 Wprowadzenie

Jest to drugi projekt z przedmioty Metody Klasyfikacji i Redukcji Wymiaru dotyczący poboru prądu urządzeń domowych. Celem tego pojektu jest wytrenowanie modeli do rozpoznawania urządzeń domowych w celu właściwej klasyfikacji. Metodą używaną w tym celu będą Ukryte Modele Markowa.

2 Dane

2.1 Pochodzenie danych

Dane pochodzą z pliku **house3_5devices_train.csv**, w którym podany jest pobór prądu (tzw. "active power") zmierzony dla 5 urządzeń domoych w odstępach około 20 sekundowych. Dane są fragmentem zbioru REDD http://redd.csail.mit.edu/, były mierzone w jednym domu w okresie od 16.04.2011. godz 5:11:43 do 21.04.2011 godz. 7:21:44 (ok. 15000 pomiarów).

2.2 Format danych

Dane mają następujący format:

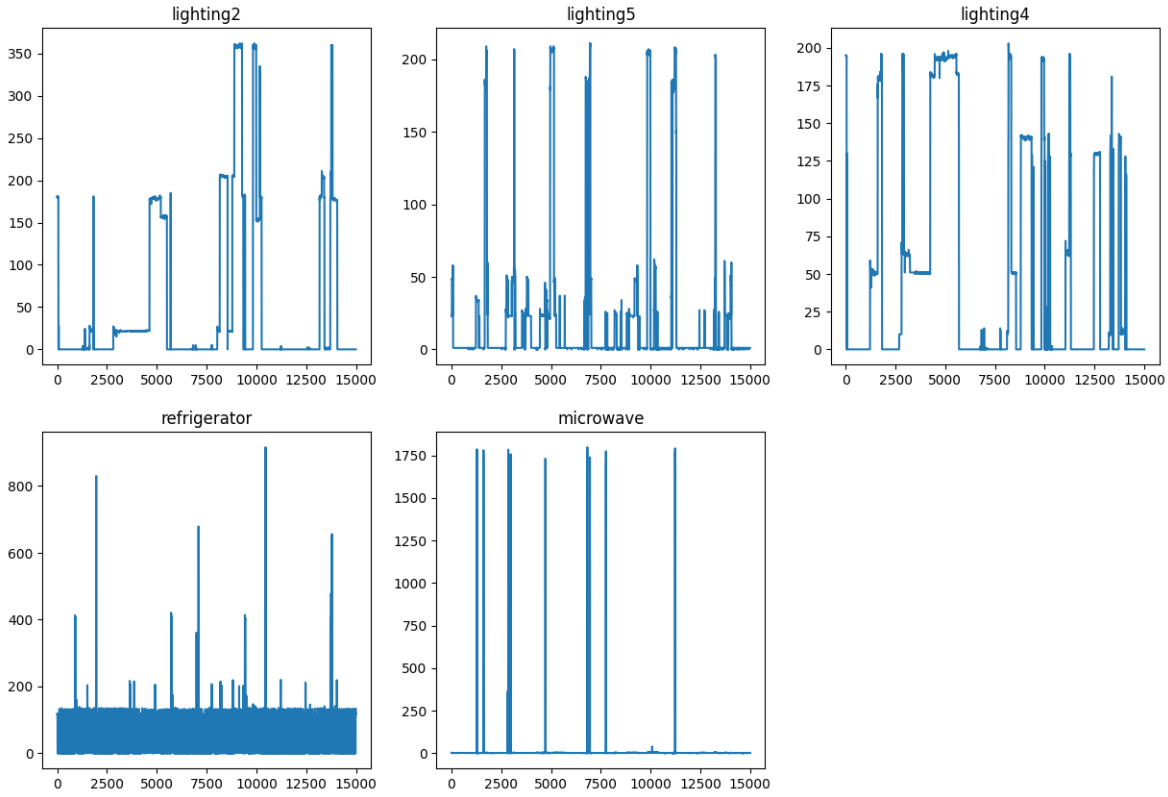
Tabela 1: Format poboru prądu dla urządzeń

Time	Lighting2	Lighting5	Lighting4	Refrigerator	Microwave
1302930703	180	23	195	117	2
1302930721	181	23	195	119	2
1302930738	180	23	195	117	2
1302930765	181	23	195	117	2
1302930782	180	23	195	118	2
...	...	...	...	...	...
1303370423	0	1	1	119	1
...	...	...	...	...	...

Pierwsza kolumna to oznaczenie czasu, dla nas nie będzie ono grało żadnej roli w trenowaniu modeli, jedynie będziemy go traktować jako kolejne punkty czasowe. Pozostałe 5 kolumn to wartości poboru mocy podane dla 3 żarówek, lodówki i mikrofalówki.

2.3 Wizualizacja danych

Powyższe dane będą bardziej czytelne po przedstawieniu ich na wykresach:



Rysunek 1: Wykres poboru mocy każdego z urządzeń.

Widać, że urządzenia dość różnią się między sobą w poborze prądu, najbardziej wyróżniają się lodówka i mikrofalówka, których zakres jest większy, niż żarówek. Każde z urządzeń ma pewnego rodzaju skoki, a najbliżej stałego poboru mocy jest lodówka, choć nie jest on stały w zupełności.

3 Metody matematyczne

3.1 Łańcuchy Markowa

Łańcuch Markowa jest ciągiem $X_1, X_2, X_3, \ldots$ zmiennych losowych. Dziedzinę tych zmiennych nazywamy przestrzenią stanów, a realizację X_n to stany w czasie n.

Łańcuch Markowa jest charakteryzowany przez stan początkowy X_0 oraz macierz przejść. Macierz przejść to macierz stochastyczna, której wyraz (i, j)

wyraża się wzorem:

$$p_{i,j} = P(X_{n+1} = j | X_n = i),$$

co oznacza prawdopodobieństwo, że w czasie $n+1$ łańcuch markowa przyjmie stan j, wiedząć, że w chwili n przyjmował stan i.

Aby zrozumieć zastosowanie łańcuchów, posłużymy się następującym przykładem.

Weźmy wektor $\mu = (S, D)$ jako wektor możliwych stanów pogody, gdzie S oznacza pogodę słoneczną, a D pogodę deszczową. Przyjmijmy następującą macierz przejścia:

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{pmatrix}$$

Pierwszy wiersz to prawdopodobieństwo zmiany pogody po dniu słonecznym, jest to kolejny dzień słoneczny z prawdopodobieństwem 0.9 lub dzień deszczowy z prawdopodobieństwem 0.1. Z kolei drugi wiersz oznacza pogodę po dniu deszczowym: z prawdopodobieństwem 0.8 wystąpi dzień słoneczny, a z 0.2 kolejny dzień deszczowy.

W ten sposób za pomocą łańcucha markowa można opisać pogodę jako pewien proces losowy.

3.2 Ukryte Łańcuchy Markowa

Załóżmy, że do powyższego przykładu dokładamy jescze jedną zmienną - tylko ona jest przez nas obserwowana. Będą te stany aktywności pewnego mieszkańca opisywane jako $\nu = (B, Z, T)$, gdzie B oznacza bieganie, Z zakupy, a T oglądanie telewizji. Załóżmy, że znamy prawdopodobieństwa z jakimi ten mieszkaniec podejmuje te aktywności w zależności od pogody (z oznaczeniami jak z poprzedniego przykładu):

$$\begin{aligned} P(B|S) &= 0.8, \\ P(Z|S) &= 0.1, \\ P(T|S) &= 0.1, \\ P(B|D) &= 0.15, \\ P(Z|D) &= 0.35, \\ P(T|D) &= 0.5. \end{aligned}$$

Na podstawie ciągu obserwacji czynności mieszkańca $C = \{BBBZTTBZ \dots\}$ będziemy próbowali powiedzieć jaka pogoda miała miejsce w poszczególne dni.

Jest to bardzo optymistyczny przykład ukrytych łańcuchów markowa, jednak w rzeczywistości bardzo często dany jest pewien ciąg obserwacji bez podanej liczby stanów, możliwych do przyjmowania. Wtedy cały ciężar leży w wyznaczeniu ilości tych stanów i w dalszej analizie pozwalającej na określenie kolejności ich występowania. W takiej właśnie sytuacji jesteśmy w przypadku naszego projektu - mamy podany wyłącznie pobór prądu w kolejnych chwilach, ale nic poza tym nie jest nam znane.

4 Stosowane sposoby klasyfikowania urządzeń

W naszych danych możemy zaobserwować, że za pobór prądu muszą odpowiadać jakieś stany, dodatkowo nie ogarniczają się one do włączony/wyłączony, lecz jest ich więcej, najwyraźniej musi stać za tym więcej zmiennych, których nie znamy wejściowo, a jedynie możemy się domyślać.

4.1 Kryteria wyboru modelu

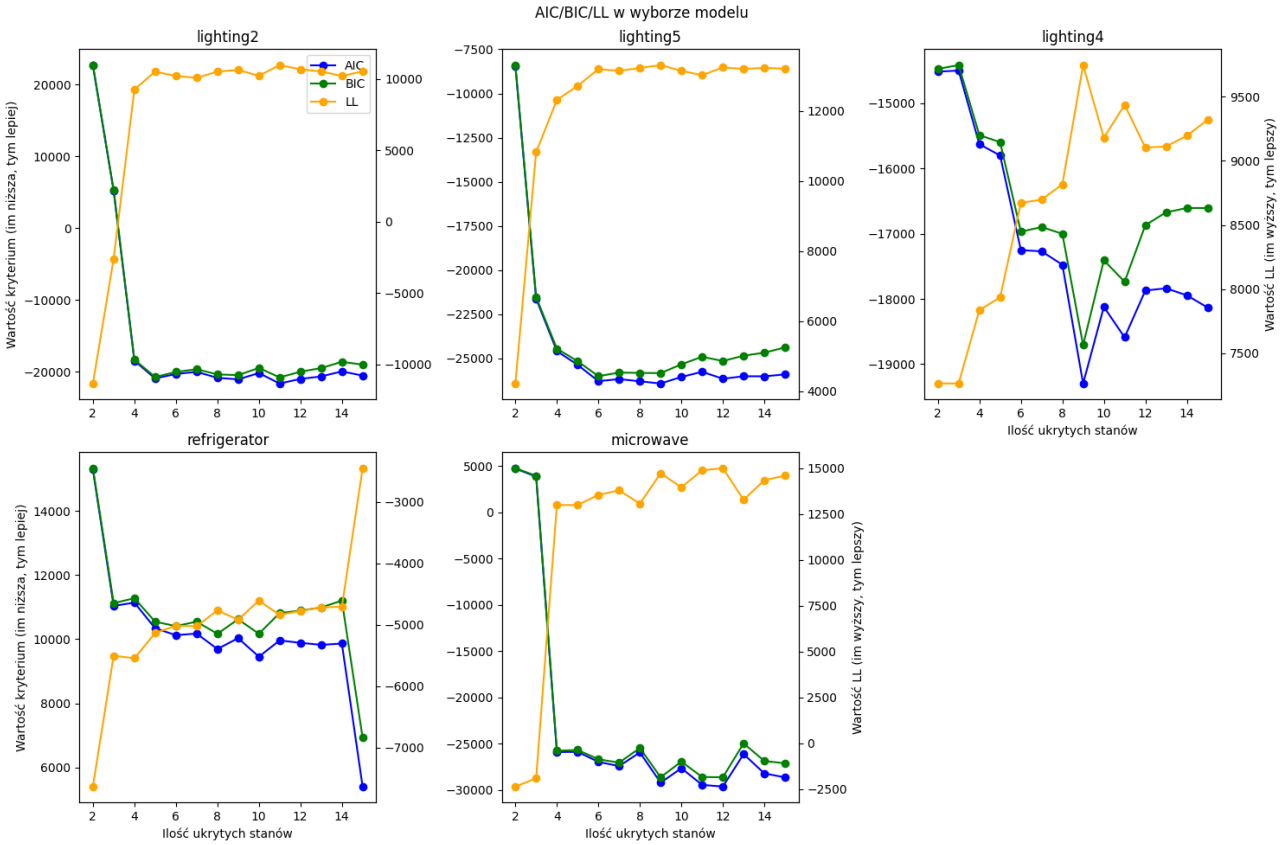
W dalszej części będziemy korzystać z trzech kryteriów doboru modeli:

- logLikelyhood: logarytm z funkcji wiarogodności, spodziewamy się, że wytrenowany model zwróci wysokie wyniki dla danych testowych podobnych do treningowych.
- AIC (Akaike Information Criterion): $\mathrm{AIC} = 2k - \log L$, jest to kryterium informacyjne, którego wartość rośnie zależnie od liczby parametrów k, a maleje wraz z wartością logarytmu wiarogodności. To kryterium faworyzuje możliwie maksymalne objaśnianie modelu, więc jest skłonne do brania dużej liczby parametrów. Dobór dobrego modelu jest oznaczony przez niską wartość.
- BIC (Bayesian Information Criterion): $\mathrm{BIC} = k \log N - 2 \log L$, to również jest kryterium informacyjne jak AIC, jednak ono ma dodatkową karę, która zależnie od liczby obserwacji będzie wzmacniała wagę liczby parametrów. To kryterium będzie faworyzowało dobre objaśnienie, przy możliwie małej liczbie parametrów. Dobór dobrego modelu będzie oznaczony przez niską wartość tego kryterium.

4.2 Wpływ ilości ukrytych stanów na jakość klasyfikacji

W naszej funkcji dane - cały zbió r- dzielimy na dwa - treningowy i testowy w stosunku 12:3, następnie dla liczby stanów od 2 do 15 powtarzamy dziesięć razy tworzenie modelu z funkcji GaussianHMM dla różnych ziaren i wybieramy najlepszy model, porównując wynik LL, AIC i BIC na zbiorze testowym.

Otrzymane wyniki przedstawiamy na wykresie:



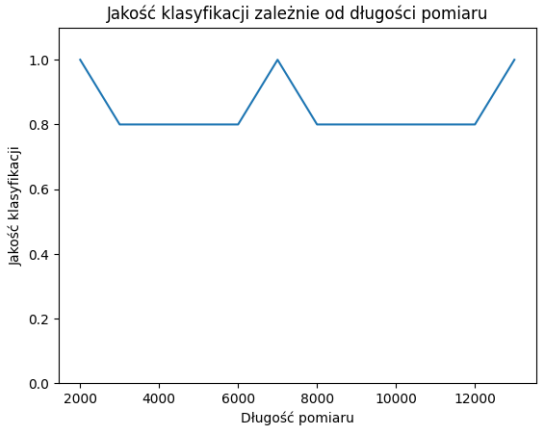
Rysunek 2: Wykres wpływu ilości ukrytych stanów.

Otrzymane wykresy pokazują, że dobre wyniki otrzymujemy:

- lighting2: już od 5 stanów ukrytych,
- lighting5: już od 6 stanów ukrytych,
- ligthing4: tutaj wykres jest bardzo rozrzucony, ale najlepszy wynik otrzymujemy dla 9 stanów ukrytych,
- refrigerator: najlepszy wynik otrzymujemy dopiero dla 15 stanów, choć to może nie być wystarczająco, obserwując ostatnią tendencję,
- microwave: już od 9 stanów ukrytych,
- dodatkowo widać, że kryterium BIC ma większe wartości od AIC, co wynika ze zwiększania liczby parametrów, bez uzyskiwania skoków na jakości dopasowania.

4.3 Wpływ długości pomiaru na jakość klasyfikacji

W tej części bierzemy początkowe n obserwacji ze zbioru jako zbiór treningowy, a następnie $\frac{n}{5}$ kolejnych obserwacji jako zbiór testowy i uruchamiamy na nich program, aby przydzielił klastyfikację. Następnie prezentujemy to na wykresie:



Rysunek 3: Wykres jakości klasyfikacji w zależności od długości pomiaru treningowego.

W większości przypadków klasyfikator poradził sobie dobrze lub bardzo dobrze. Największym problemem jest poprawne sklasyfikowanie ligthing2, ponieważ jego wykres poboru mocy na niektórych odcinkach jest identyczny do lighting4, stąd pojawiają się błędy i klasyfikacja nie jest idealna. Pozostałem urządzenia nasz model klasyfikuje bardzo dobrze.

5 Podsumowanie

Problem klasyfikacji urządzeń z naszych danych okazał się być możliwy do rozwiązania w całkiem dobry sposób - skorzystaliśmy z ukrytych łańcuchów markowa, które pozwoliły dokonać właściwej klasyfikacji. Nasza metoda działa dobrze i można ją stosować w podobnych problemach.

Jednocześnie ma ona wadę, ponieważ nie jest w stanie rozróżnić dość podobnych sygnałów z różnych urządzeń, więc może wymagać pewnych poprawek.