

Metody Odkrywania Wiedzy

Dokumentacja końcowa projektu

„Predykcja zużycia energii na podstawie danych czujnikowych”

Krzysztof Belewicz
Paweł Pińczuk

26 stycznia 2020

1. Opis projektu

Celem projektu było wyznaczenie całkowitego zużycia energii dla zadanej chwili czasu, tzn. sumy poborów sprzętów AGD (kolumna *Appliances*) i oświetlenia (kolumna *lights*). Zbiór danych został pozyskany z archiwum dostępnego na stronie: <https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction>. Pojęciem docelowym jest wartość całkowitej pobieranej mocy przez gospodarstwo domowe. W ramach projektu zdecydowano się na oddzielne wykonania zadania regresji dla celu *Appliances* i celu *lights*, ze względu na hipotezę, że modele je wyznaczające mogą mieć inne właściwości.

Dokonano selekcji atrybutów za pomocą trzech algorytmów opisanych w rozdziale 3. Przeprowadzono procedurę oceny algorytmów liniowej regresji, drzew regresji oraz kawałkami liniowej regresji.

2. Opis danych

2.1. Charakterystyka danych

Dane wykorzystywane do eksperymentów zostały zebrane za pomocą sieci czujników w niewielkim domu w czasie 4.5 miesiąca. Składają się z:

- daty i godziny pomiaru,
- poboru energii sprzętów domowych [*Wh*],
- poboru energii oświetlenia [*Wh*],
- pomiarów temperatury i wilgotności dla 8 różnych pomieszczeń ($^{\circ}\text{C}$, [%]),
- pomiarów temperatury i wilgotności dla zewnętrznej, północnej strony budynku ($^{\circ}\text{C}$, [%]),
- danych z pobliskiej stacji pogodowej:
 - temperatura powietrza [$^{\circ}\text{C}$],
 - temperatura punktu rosy [$^{\circ}\text{C}$],
 - ciśnienie atmosferyczne [*mm Hg*],
 - wilgotność [%],
 - prędkość wiatru [*m/s*],
 - widoczność [*km*].

2.2. Przygotowanie danych

Każdy pomiar został uśredniony z 3 próbek wykonanych w równych odstępach co ok. 3,3 min. W ramach przygotowania danych, data i godzina pomiaru zostały rozdzielone na cztery oddzielne kolumny, zawierające miesiąc, dzień, godzinę i minutę pomiaru.

3. Selekcja atrybutów

Aby zapobiec nadmiernemu dopasowaniu, stosuje się selekcję atrybutów, która wybiera kilka najważniejszych atrybutów do późniejszego stworzenia modeli. Po zastosowaniu selekcji, modele oparte o ograniczoną liczbę atrybutów zwykle są lepsze od opartych o wszystkie atrybuty. Istnieje wiele metod selekcji atrybutów; w ramach projektu zostało sprawdzone kilka metod (w nawiasach umieszczono opcję type funkcji *feature_selection*):

- prosty filtr statystyczny („*simple*”) - pomiędzy każdym z atrybutów a celem regresji stosuje się miarę statystyczną, która określa zależność celu od danego atrybutu (dalej „miara zależności”). Następnie wybiera się kilka atrybutów o największej „mierze zależności”. W ramach regresji pomiędzy atrybutami ciągłymi zastosowano współczynnik korelacji (Pearsona);
- bazująca na drzewach losowych („*rf*”) - w tym celu wykorzystano pakiet *randomForest* i jego wbudowaną opcję zwracającą parametr *IMPORTANCE* (bazujący na mierze MSE), o możliwości konfiguracji ilości drzew;
- metoda *RRELIEF* („*relief*”) - wersja algorytmu *RELIEF* do zastosowań w zadaniu regresji. Algorytm *RELIEF*, początkowo zaprojektowany dla zadania klasyfikacji binarnej, polega na losowym wybraniu obserwacji (jednego rekordu klasy+atrybuty). Następnie wyszukuje się *k* najbardziej podobnych obserwacji tej samej klasy, oraz *k* klasy przeciwnej. Dla każdego atrybutu oblicza się wagę istotności. Po wykonaniu *trees_num* operacji, wykonuje się średnią wag istotności. Atrybuty segreguje się według wag istotności. W zadaniu regresji stosuje się inne funkcje obliczające wagę np. funkcję rozkładu.

W ramach projektu stosuje się następujące podejście: dla każdej wymienionej metody wykonuje się selekcję połowy atrybutów (*part=0.5*) atrybutów, następnie wyznaczoną formułę aplikuje się do stworzenia modelu *rpart()*, i procedurze oceny (10-krotnej walidacji krzyżowej *model_eval()*). Następnie największy współczynnik korelacji Pearsona wyznacza najlepszą metodę selekcji atrybutów oraz formułę do stworzenia modelu.

3.1. Wyniki selekcji atrybutów

Na Rys. przedstawiono wyniki każdej z selekcji. W Tabeli 3.1 przedstawiono porównanie wyników każdej z selekcji. Wynika z tego, że atrybuty wyznaczone metodą pozwalają na najlepsze tłumaczenie modelu. Dla porównania przedstawiono też wynik walidacji krzyżowej dla modelu opartego o wszystkie atrybuty. Wynika

Tablica 3.1: Wyniki selekcji atrybutów - współczynniki korelacji

Parametr	<i>randomForest</i>	<i>simple</i>	<i>RELIEF</i>	bez selekcji
Appliances	0,764	0,798	0,766	0,779
lights	0,764	0,798	0,766	0,779

4. Konstrukcja i ocena modeli

4.1. Metody konstrukcji modeli

Pojedyncze modele drzew regresji zazwyczaj cierpią z powodu wysokiej wariancji – jedną z metod jej redukcji jest tzw. Bagging (**B**ootstrap **a**ggregating). Metoda ta polega na łączeniu i uśrednianiu wielu modeli drzew, co zmniejsza wariancję i redukuje zbytnie dopasowanie.

Bagging może zostać zrealizowany za pomocą pakietu *ipred* lub *caret*. *Ipred* jest z reguły prostszy w realizacji, jednakże stosowanie *caret* niesie za sobą kilka zalet. Znacznie prościej jest weryfikować krzyżowo wyniki – pomimo możliwości wykorzystania błędu OOB (Out-of-Bag) w *ipred*, weryfikacja krzyżowa daje dużo lepsze zrozumienie spodziewanego błędu. Dodatkowo możliwy jest dostęp do zmiennej odpowiedniości w wygenerowanych drzewach.

Do konstrukcji modeli klasyfikacji został wykorzystany pakiet *rpart* budujący drzewo klasyfikacji oraz pakiet *e1071*, który umożliwia waidacje skrosna oraz automatyczny dobór parametrów w celu zminimalizowania błędu.

4.2. Budowa modelu klasyfikacji z pakietu *ipred*

TODO - pakiety R, parametry (np kryteria stopu, etc)

4.3. Miary jakości

Dla zbudowanych modeli oblicza się następujące miary jakości:

1. CC - współczynnik korelacji liniowej Pearsona

$$CC = \frac{cov(P, A)}{var(P) \cdot var(A)}$$

2. MSE - błąd średniokwadratowy

$$MSE = \frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}$$

3. RMSE - pierwiastek z błędu średniokwadratowego

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$$

4. MAE - średni błąd względny

$$MAE = \frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n}$$

5. RSE - względny błąd kwadratowy

$$RSE = \frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}$$

6. RRSE - pierwiastek ze względnego błędu kwadratowego

$$RRSE = \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$$

7. RAE - błąd względny

$$RAE = \frac{|p_1 - a_1| + \dots + |p_n - a_n|}{|a_1 - \bar{a}| + \dots + |a_n - \bar{a}|}$$

TODO tutaj ładnie się wpasuje co po kolei w kodzie poszło

4.4. Procedury oceny

Aby móc ocenić model pod względem przydatności zastosowano metodę k-krotnej walidacji krzyżowej. Kod zawarto w funkcji *model_eval()*, która stanowi . Zbiór testowy jest dzielony losowo na k podzbiorów równej wielkości. W kolejnych iteracjach każdy ze zbiorów jest traktowany jako zbiór testowy, podczas gdy na reszcie danych buduje się model. Następnie modele są uśredniane i następuje predykcja. Po predykcji modelu na zbiorze testowym wyznacza się miary jakości opisane w 4.3.

5. Wnioski

TODO