Metody Odkrywania Wiedzy (MOW)
Projekt analityczny - klasyfikacja
Dokumentacja wstępna

Górski Michał
Napieralski Adam

15 kwietnia 2020

1 Cel projektu

Celem projektu jest wnikliwa i szeroko zakrojona analiza danych z wykorzystaniem wybranych pakietów i algorytmów zaimplementowanych w środowisku R. Za pomocą dostępnych w R narzędzi stworzone zostaną odpowiednie modele pozwalające na przydzielenie obserwacji do jednej z kategorii na podstawie wartości jej atrybutów. Działania jakie zostaną podjęte przy realizacji zadania projektowego to:

- wstępne przetworzenie/transformacja danych,
- podstawowy statystyczny opis danych,
- strojenie parametrów algorytmów,
- tworzenie i ocena jakości modeli,
- opis wyników i przedstawienie wniosków.

Skonkretyzowanym celem analizy jest zbudowanie modeli klasyfikujących typ wykonywanej aktywności fizycznej na postawie dostępnych danych pochodzących z wewnętrznych czujników smartfona.

2 Opis danych

Dane do projektu pochodzą ze strony UCI Machine Learning Repository [1]. Zawierają one wstępnie przetworzone dane oparte na surowych danych z 3-osiowego akcelerometru i żyroskopu (opisujących przyspieszenia liniowe i prędkości kątowe) wbudowanych w smartfon, które zebrane zostały w trakcie eksperymentu na próbie 30 wolontariuszy w wieku 19-48 lat. Przetworzone zostały przez zastosowanie filtrów i połączenie w próbki - odpowiadające przykładom, dla których wyznaczone zostały wektory 561 atrybutów.
Atrybuty te zawierają m.in. zestawy wartości opisujące: średnią, odchylenie standardowe, średnie odchylenie bezwzględne, wartość największą, wartość najmniejszą, obszar magnitudy sygnału, energię, rozstęp ćwiartkowy, entropię,

współczynniki autoregresji, współczynniki korelacji, parametry charakterystyki częstotliwościowej.
Wszystkie te atrybuty są numeryczne, a ich wartości zostały znormalizowane i ograniczone do przedziału $[-1, 1]$.
Atrybut dyskretny, pełniący rolę interesującego pojęcia domyślnego, to rodzaj wykonywanej aktywności, składający się z kategorii przedstawionych w Tablicy 1.

Tablica 1: Klasy pojęcia domyślnego

Klasa	Nazwa
1	Chodzenie
2	Wchodzenie po schodach
3	Schodzenie
4	Siedzenie
5	Stanie
6	Leżenie
7	Siadanie ze stania
8	Wstawanie z siedzenia
9	Położenie z siedzenia
10	Siedzenie z leżenia
11	Leżenie ze stania
12	Stanie z leżenia

Pełen zbiór zawiera 10 929 przykładów.

3 Wstępne przygotowanie danych

Wszystkie dane, kategoryczne (binarne) oraz nominalne, przedstawione są w postaci liczbowej. W zbiorze nie występują niepełne obserwacje. Wartości są znormalizowane. Analiza ograniczy się tylko do klas z aktywnościami nieprzejściowymi przedstawionymi w Tabeli 2, z uwagi na ich znacząco (ok. 20 razy) większą reprezentację w przykładach. Zbiór uwzględniający klasy z aktywnościami nieprzejściowymi zawiera 10411 przykładów.

Tablica 2: Wybrane klasy pojęcia docelowego

Klasa	Nazwa
1	Chodzenie
2	Wchodzenie po schodach
3	Schodzenie
4	Siedzenie
5	Stanie
6	Leżenie

Dodatkowo w celu ograniczenia znaczącej liczby atrybutów podjęta zostanie próba selekcji najistotniejszych z nich za pomocą metody *Boruta* lub innej.

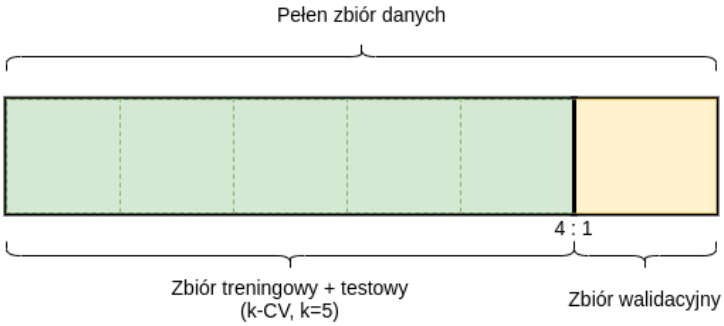
4 Wybór i strojenie algorytmów

Do analizy wstępnie wybrane zostały algorytmy klasyfikacji:

1. Naiwny Klasyfikator Bayesowski (z pakietu *e1071* lub "klaR").
   Naiwny klasyfikator Bayesowski nie wymaga strojenia parametrów, co jest wygodną cechą. Pomimo że założenia niezależności są często naruszane, sprawdza się on dobrze w modelach, w których wymagane jest uwzględnienie nieznacznego wpływu znacznej liczby atrybutów, czego możemy się spodziewać w tym zadaniu.

2. SVM - Support Vector Machines (z pakietu *e1071*).
   Przy strojeniu sprawdzone zostanie działanie algorytmu dla jądra typu *radial basis* przy różnych wartościach jego parametru σ. Dodatkowo możliwe jest w mniejszym zakresie sprawdzenie typów jądra *polynomial* czy *sigmoid*.

Zależnie od ocenianej na bieżąco sprawności prac, niewykluczona jest klasyfikacja z wykorzystaniem algorytmu lasu losowego *randomForest* i odpowiednia parametryzacja używanego w nim drzewa.
Algorytmy strojone będą z wykorzystaniem oceny pośredniej wyprowadzonej na wybranej z użyciem losowania warstwowego 4/5 części zbioru pełnego. Stosowana będzie tam k-krotna walidacja krzyżowa (k-CV) z $k = 5$ (Rys. 1).
Dodatkowe szczegóły dotyczące wyboru algorytmu i dokładnego określenia parametrów podjęte zostaną na etapie realizacji analizy, uzależniając ją od stopnia skomplikowania zadania projektowego.



Rysunek 1: Podział zbioru danych na podzbiory.

5 Ocena jakości modeli

Prowadzenie oceny pośredniej nastąpi za pomocą wspomnianej procedury k-krotnej walidacji krzyżowej (k-CV). Ocena końcowa wyprowadzona zostanie zgodnie z procedurą *holdout* na nieużywanej wcześniej części zbioru danych. Do oceny uwzględnione zostaną rozkłady pomyłek, wyznaczona zostanie macierz pomyłek na zbiorze oraz powiązane z nią współczynniki. Na ich podstawie możliwe będzie wykreślenie krzywej ROC oraz wyznaczenie AUC.

Literatura

[1] Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set `http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions`