10.12.2018

Błażej Kucman
ind. 238228
gr.1

Inteligencja Obliczeniowa – Projekt 2

Zgłębianie danych

Tematem zadania było przećwiczenie poznanych na laboratoriach technik zgłębiania danych na wybranym zbiorze danych (najlepiej z kolumną „class" z dwoma możliwymi wartościami).

1. Wstęp i omówienie bazy

Do zadania została wybrana baza z wyniki i danymi osób z problemami chorób serca. Na ich podstawie chcemy się dowiedzieć przy jakich wynikach oraz danych osoby jest możliwa choroba serca.

Baza składa się z kolumn:

Skrót ST – częstoskurcz zatokowy

age - wiek w latach,
sex – płeć(1=mężczyzna, 0= kobieta),
cp – rodzaj bólu w klatce piersiowej (wartości 1-4)
1.Typowa angina
2.Nietypowa angina
3.Nie anginowy ból
4.Bezobjawowy
trestbps - spoczynkowe ciśnienie krwi (w mm Hg przy przyjęciu do szpitala)
chol - poziom cholesterolu w surowicy w mg/dl
fbs – cukier we krwi(na czczo) > 120 mg/dl (1-tak, 0-nie)
restecg - spoczynkowe wyniki elektrokardiograficzne(wartości 0-3)
- Wartość 0: normalna
- Wartość 1: nieprawidłowość fali ST-T (odwrócenie załamka T i / lub uniesienie odcinka ST> 0,05 mV)
- Wartość 2: wykazanie prawdopodobnego lub określonego przerostu lewej komory według kryteriów Estesa.
thalach – osiągnięte maksymalne tętno
exang – angina wywołana wysiłkiem (1-tak , 0-nie)
oldpeak – spadek ST wywołany pod czas ćwiczeń względem stanu spoczynku
slope – nachylenie spadku ST pod czas ćwiczeń(0-3)
- Value 1: upsloping
- Value 2: flat
- Value 3: downsloping

ca - liczba głównych naczyń (wartości 0-3) zabarwionych za pomocą fluoroskopii
thal – 3 – normalny; 6 – wyleczony , 7- wada odwracalna
target – 0-zdrowy, (1-4) – chory

Jako klasa wybrana kolumna target została ona przekształcona tak aby przechowywała tylko wartości 0 i 1 gdzie 1 będzie znaczyło chory .

2. Przetwarzanie i obróbka danych

Przetworzona została kolumna target która przechowywała wartości od 0 do 4, na taką która będzie miała wartości 0 i 1. Kolumna ta będzie służyć jako kolumna klasy.

Wykorzystana została biblioteka deducorrect i funkcja editWithRules. Poniżej kod odpowidzialny za to.

```
setSickCorrection <- correctionRules(expression(
  if(is.finite(target)){
    if(target != 0)
    {
      if (target == 1 | target == 2 | target == 3 | target == 4) {
        target <- 1
      } else
      {
        target <- NA
      }
    }
  }else
  {
    target <- NA
  }
))
```

```
heart.disase.SickSet <- (correctWithRules(setSickCorrection,heart.disase))$corrected
```

Po edycji zostaje wyciągnięta tylko część kolumn z danymi (corrected) pomijając informacje o tym które zostały zmienione.

Funkcja zmieniała także każdą wartość która nie pasuje do wzorca na i w następnym kroku zostaną usunięte wiersze które zawierają NA w target, ponieważ interesują nas tylko te które mają wartość poprawną

```
heart.disase.Better <- subset(heart.disase.SickSet,is.finite(target))
```

Za pomocą biblioteki editrules zostało sprawdzone czy w bazie znajdują się jakiekolwiek dane nie pasujące do wytycznych czy zakresów.

```
E <- editset(c(
        "age >0.0","sex %in% c(0,1)","cp %in% c(1,2,3,4)","trestbps > 0",
        "chol > 0", "fbs %in% c(1,0)","restecg %in% c(0,1,2,3)","thalach > 0",
        "exang %in% c(0,1)", "oldpeak >=0","slope %in% c(1,2,3)","ca %in% c(0,1,2,3)",
        "thal %in% c(3,6,7)"
        ))

ve <- violatedEdits(E,heart.disase.Better)
```

Dzięki temu wiadomo dla których kolum należy pisać funkcje uzupełniające dane bądź usuwające wiersze.

Po przeanalizowaniu kolumn wybrana została metoda kNN z biblioteki VIM aby na podstawie najbliższych sąsiadów uzupełnić luki w bazi. Kolumny te przechowują dane ściśle określone np.(1 lub 0), więc średnia nie była by oczekiwana w tym wypadku.

```
heart.knn <- kNN(heart.disase.Better)
```

3. Klasyfikatory i ich ewaluacja

Kolumną klasy jest target. Przechowuje ona wartość 0 i 1 czyli zdrowy i chory.

a)

Baza została podzielona na treningową i testową w stosunku 67/33.

Wylosowanie gdzie trafi który wiersz.

```
ind <- sample(2, nrow(heart.disase.ready), replace=TRUE, prob=c(0.67, 0.33))
```

i podzielenie bazy

```
heart.disase.train <- heart.disase.ready[ind==1, 1:14]
heart.disase.test <- heart.disase.ready[ind==2, 1:14]
```

b)

Użycie poszczególnych klasyfikatorów