Podstawy Przetwarzania Danych

Laboratoria 14: Niezbalansowane zbiory danych

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.0
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Niezbalansowane zbiory danych

- W niezbalansowanym zbiorze danych mamy do czynienia z brakiem równowagi w liczności poszczególnych klas.
- W przypadku uczenia maszynowego, może to skutkować uzyskaniem wysokiej wartości miary Skuteczności przy jednoczesnym braku poprawnego rozpoznawana niedostatecznie reprezentowanych klas.
- Można przeciwdziałać występowaniu tego zjawiska przekształcając dane lub stosując dedykowane klasyfikatory.

Dane

```
1 require(mlbench)
2 data(Glass)
```

- Zbiór Glass z pakietu mlbench zawiera informacje o składzie chemicznym siedmiu rodzajów szkła.
- Poszczególne rodzaje szkła reprezentowane w różnym stopniu.

Zadanie 1

- Wczytaj zbiór Glass i zostaw w nim tylko przedstawicieli klas 2 i 3.
- Nadaj klasom oznaczenia 0 i 1.
- Podziel losowo zbiór na testowy i uczący w stosunku 1:1.
- Upewnij się, że rozkład klas w zbiorze uczącym i testowym są podobne table.
- Zbuduj klasyfikator, las losowy złożony z 30 drzew, który rozróżnia obie klasy.
- Sprawdź skuteczność i zbalansowaną skuteczność tego klasyfikatora.

ROSE

```
1 require(ROSE)
2 data.rose <- ROSE(form, data=learning)$data
```

- Dwie podstawowe metody równoważenia zbioru danych to oversampling i undersampling.
  - Oversampling zwiększa liczność mniejszej klasy poprzez zdublowanie losowych przedstawicieli.
  - Undersampling zmniejsza liczność większej klasy poprzez usuwanie losowych przedstawicieli.
- Algorytm ROSE tworzy sztucznych przedstawicieli mniejszej klasy w charakterystycznej dla niej przestrzeni cech.
- Wywołanie funkcji ROSE tworzy obiekt, z którego można odczytać nowe cech z pola data

Zadanie 2

- Przekształć zbiór uczący algorytmem ROSE.
- Utwórz las losowy zbudowany z 30 drzew, który rozróżnia obie klasy, ucząc go na nowych danych ale testując na niezmodyfikowanym zbiorze testowym.
- Sprawdź skuteczność i zbalansowaną skuteczność tego klasyfikatora.

RUSBoost

```
require(ebmc)
model <-rus(form,data=learning,size=size,alg='c50')
```

- Algorytm RUSBoost buduje kolekcję słabych binarnych klasyfikatorów używając do ich budowy metody undersampling.
- Kolekcja takich klasyfikatorów wzmacnia wybieranie mniejszej klasy poprzez klasyfikację przez głosowanie.
- Algorytm jest zaimplementowany jako funkcja `rus` w pakiecie `ebmc` (pakiet zawiera też implementacje podobnych algorytmów np. SMOTE).
- Rodzaj słabych klasyfikatorów określamy parametrem `alg` a ich liczbę parametrem `size`.

Zadanie 3

- Dla nieprzekształconego zbioru uczącego i testowego zastosuj algorytm RUSBoost.
- Do budowy klasyfikatora użyj 30 drzew decyzyjnych `c50` jako słabe klasyfikatory.
- Sprawdź skuteczność i zbalansowaną skuteczność tego klasyfikatora.