# Oversampling i Undersampling

Michał Głowacki
Grzegorz Holewa
Anna Romik

# Problem klasyfikacji niezrównoważonych danych

- Niezrównoważone liczebnie klasy decyzyjne
- Słaba reprezentacja klasy mniejszczościowej
- Uczenie klasyfikatorów ukierunkowane w stronę klasy większościowej
- Duże znaczenie rozpoznawania klas mniejszczościowych

# Zastosowanie

- medycyna
- detekcja awarii
- przewidywanie katastrof

# Oversampling

- technika analizy danych wykorzystywana do regulacji podziału klas zbioru danych
- wybiera więcej próbek z klasy mniejszościowej (duplikuje dane)
- replikacja zwiększa wagę próbek populacji mniejszościowej
- nie zwiększa ilości informacji

# Synthetic minority oversampling technique (SMOTE)

- tworzenie nowych przykładów pomiędzy istniejącymi obserwacjami mniejszościowymi położonymi blisko siebie
- nadlosowywanie do uzyskania odpowiedniego poziomu

# SMOTE - schemat działania

1. Określenie globalnego poziomu nadlosowania - ***j***
2. Dla każdego przykładu z klasy mniejszościowej:

a) znaleźć k najbliższych sąsiadów (z klasy mniejszościowej)

b) spośród sąsiadów wylosować ***j*** z nich

c) utworzyć ***j*** syntetycznych przykładów pomiędzy danym przykładem a wylosowanymi sąsiadami

# Generowanie syntetycznych przykładów

- atrybuty ilościowe - losowanie liczby z zakresu wartości danego przykładu i wylosowanego sąsiada
- atrybuty jakościowe - wybranie wartości najczęściej występującej w klasie mniejszościowej

# HVDM - Heterogeneous Value Dierence Metric

$$D_{HDM}(x,y) = \sqrt{\sum_{i=1}^{N} (dh_i(x_i, y_i))^2}$$

$$Nvdm(x_i, y_i) = \sqrt{\sum_{j=1}^{C} \left( \frac{N_j(x_i)}{N(x_i)} - \frac{N_j(y_i)}{N(y_i)} \right)^2}$$

$$dh_i(x_i, y_i) = \begin{cases} 1 & (1) \\ Nvdm(x_i, y_i) & (2) \\ Ndif(x_i, y_i) & (3) \end{cases}$$

$$Ndif(x_i, y_i) = \frac{|x_i - y_i|}{4\sigma_i}$$

# Problemy

- Wzmacnianie klasy mniejszościowej w sposób losowy
- Nie uwzględnianie przykładów z klasy większościowej
- Niewielka odporność na szum
- Problem obliczania odległości dla atrybutów nominalnych

# Undersampling

- technika analizy danych wykorzystywana do regulacji podziału klas zbioru danych
- odrzuca część próbek z klasy większościowej

# Metody undersamplingu

- Random undersampling
- Tomek links
- Edited Nearest Neighbourhood
- Nearest Cleaning Rule

# Tomek links

- Usuwanie przykładów granicznych i szumu z klasy większościowej

Dla każdego przykładu z klasy większościowej i klasy mniejszościowej:

Para (Ei, Ej) nazywana jest “Tomek link”, jeżeli nie istnieje przykład El, dla którego jest spełniona zależność:

$$d(Ei, El) < d(Ei,Ej) \text{ lub } d(Ej, El) < d(Ei, Ej)$$

Przykłady będące w “Tomek link” są usuwane ze zbioru

# Edited nearest neighbourhood

Dla każdego przykładu z klasy większościowej znajdowanych jest trzech najbliższych sąsiadów.

Jeżeli klasy dwóch sąsiadów są różne od klasy przykładu to przykład jest usuwany.

# Nearest Cleaning Rule

Dla każdego przykładu ze zbioru znajdowanych jest trzech najbliższych sąsiadów.

Jeżeli przykład należy do klasy większościowej i sąsiedzi klasyfikują go inaczej to jest usuwany.

Jeżeli przykład należy do klasy mniejszościowej i sąsiedzi klasyfikują go inaczej to usuwani są sąsiedzi z klasy większościowej

# Problemy

- Usunięcie zbyt wielu przykładów klasy większościowej
- Doprowadzenie do gwałtownego pogorszenia rozpoznawania klas większościowych

# Przykład

Przykład:

Mając zbiór 1000 osób w którym 66% to mężczyźni. Wiemy także, że ogólny rozkład populacji to 50% mężczyzn i 50% kobiet.

Oversampling w tym przypadku policzy każdą kobietę 2 razy. Wyprodukuje zbiór zbalansowany mający 1333 elementy i 50% populacji kobiet.

Undersampling losowo usunie część mężczyzn z tego zbioru tworząc zbiór 667 elementów, również z 50% kobiet.