SMOTE cd / krzywa ROC

Paweł Szopa

SMOTE cd

Zbiory danych powinny być zbalansowane, czyli liczność jednaj klasy powinna być w przybliżeniu równa liczności drugiej klasy.

Często zdarza się jednak że zbiory danych są niezbalansowane, np:

Przykłady niezbalansowanych zb.

- monitorowanie usterek skrzyni biegów w helikopterach
- rozróżnienie pomiędzy trzęsieniem ziemi a wybuchem jądrowym
- filtrowanie dokumentów
- wykrywanie wycieków ropy
- wykrywanie fałszywych/nieuczciwych rozmów telefonicznych
- obrazy mammograficzne

Obrazy mammograficzne

Typowy mammograficzny zbiór danych może zawierać 98% prawidłowych oraz 2% nieprawidłowych pikseli.

Zadaniem metody SMOTE jest to aby klasifikator nie pominął badania nieprawidłowych pikseli.

Ocena jakości klasyfikatora

Krzywa **ROC** (Receiver Operating Characteristic), jest narzędziem do oceny poprawności klasyfikatora:

- zapewnia ona łączny opis jego czułości i specyficzności.
- jest szeroko wykorzystywana, np w diagnostyce medycznej (EKG, USG)

Zdefiniowanie problemu

Problemem przynależności do dwóch klas (np. przyporządkowanie pacjenta do jednej z dwóch grup: **zdrowy – chory**). W szczególnym przypadku **decyzja** może być podejmowana na podstawie **jednej zmiennej** – wskaźnika diagnostycznego. Badamy wówczas zależność pomiędzy tą zmienną, najczęściej mierzoną na skali ilościowej, a **wybraną zmienną dwustanową** (np. wiek pacjenta a występowanie powikłań po zabiegu). Często zakłada się, że badana zależność jest **monotoniczna** (wraz ze wzrostem wartości zmiennej diagnostycznej rosną szanse na wystąpienie badanego zjawiska lub na odwrót)

reguła decyzyjna: wartość wskaźnika diagnostycznego np. x = 50

wiek > 50 oznacza powikłania po zabiegu

wiek < 50 oznacza brak powikłań po zabiegu

Błędne decyzje

- Nieodłącznym elementem podejmowania decyzji jest popełnianie błędów
- Błędne decyzje modelu klasyfikującego są nieuniknione, ponieważ często klasy nie są całkowicie separowalne

Błędne decyzje

Łatwo wyobrazić sobie taką sytuację, kiedy dwa obiekty charakteryzowane są za pomocą takich samych wartości zmiennych niezależnych, ale należą do dwóch różnych klas. Często wynika to z niepełnej wiedzy o badanym zjawisku lub po prostu niekompletnych danych.

Celem jest przewidywanie klasy wyróżnionej i w związku z tym poprawne decyzje to: prawidłowe wskazanie wyróżnionej klasy (TP – true positive) oraz prawidłowe nie wskazanie drugiej z klas (TN – true negative). Błędy popełniamy w sytuacji, gdy niepoprawnie wskazujemy wyróżniona klasę (FP – false positve) lub niewskazujemy klasy wyróżnionej w sytuacji, gdy powinniśmy ją wskazać (FN – false negative).

Tab. 1. Macierz klasyfikacji – stan faktyczny i wskazanie modelu.

	Zaobserwowano stan wyróżniony	Nie zaobserwowano stanu wyróżnionego
Przewidywano stan wyróżniony	TP	FP
Nie przewidywano stanu wyróżnionego	FN	TN

- TP, FP, FN oraz TN oznaczają liczbę obserwacji, które trafiły do danej komórki tabeli

- Dobry model to taki, który minimalizuje liczbę błędów, czyli FN oraz FP.

Najlepsza reguła decyzyjna ma nam zapewnić najlepsze wyniki – jak najmniejszą liczbę błędów. Aby móc precyzyjnie zdefiniować odpowiednie kryterium wprowadza się **miary jakości** reguł decyzyjnych:

- specyficzność
- czułość

- czułość to prawidłowe wskazanie klasy pozytywnej / suma obserwacji stanu wyróżnionego

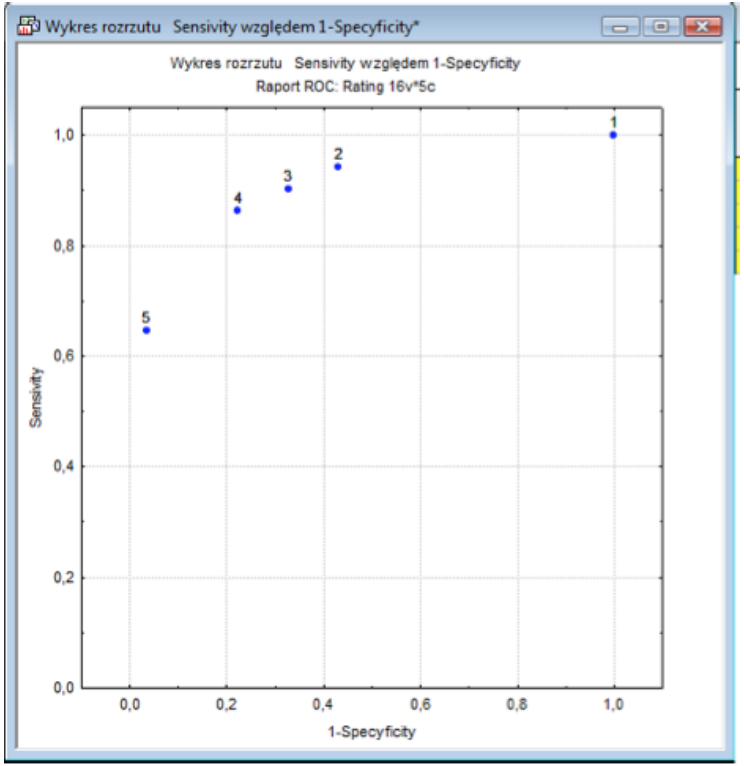
$$\text{Czułość} = \frac{TP}{TP + FN},$$

- specyficzność to prawidłowe wskazanie klasy negatywnej / suma obserwacji stanu niewyróżnionego

$$\text{Specyficzność} = \frac{TN}{TN + FP}.$$

Reasumując, dla każdego z możliwych punktów odcięcia obliczamy czułość i specyficzność, a następnie zaznaczamy otrzymane wyniki na wykresie.

Uzyskane punkty ze sobą łączymy. Im więcej różnych wartości badanego wskaźnika, tym gładsza uzyskana krzywa.



Rys. 2. Przykład konstrukcji krzywej ROC.

Podsumowanie

Oprócz wspomagania wyboru optymalnego punktu odcięcia krzywa ROC używana jest do porównywania różnych modeli, czy to zbudowanych na podstawie różnych zmiennych niezależnych czy też różnymi metodami. Zaletą tej metody jest to, że pokazuje siłę wpływu predyktora na występowanie wybranej klasy dla wszystkich możliwych punktów odcięcia. Zatem krzywe ROC są narzędziem do wyboru progu decyzyjnego, ale też narzędziem do wizualizacji całej sytuacji decyzyjnej.