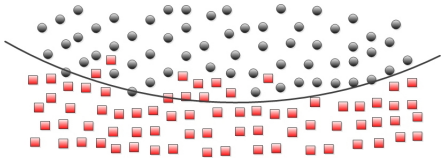
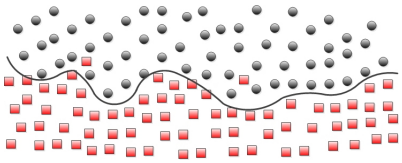
Ćwiczenie 3

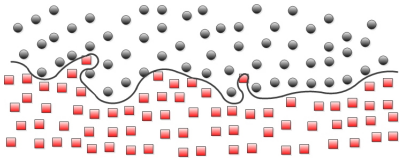
Metody Szacowania Efektywności Klasyfikatorów



Rysunek 1: Przykład modelu niedotrenowanego



Rysunek 2: Przykład modelu dobrze wytrenowanego



Rysunek 3: Przykład modelu przetrenowanego (przedopasowanego)

Zadanie do wykonania

1) Utwórz na pulpicie katalog w formacie Imię_nazwisko, w którym umieść wszystkie pliki związane z ćwiczeniem.

2) Przeczytaj teorię związaną z metodami oceny jakości klasyfikatorów.

3) Zaimplementuj w dowolnym języku programowania (preferowany C++) następujące metody szacowania efektywności klasyfikatorów:

a) Metoda Trenuj i Testuj (Train and Test) ($T\&T$) - ratio podziału 0.5,

b) Metodę Walidacji krzyżowej Monte Carlo (Monte Carlo Cross Validation) - 5 ($MCCV-5$) (ratio podziału 0.5),

c) Metodę Walidacji Krzyżowej (Cross Validation) - 5 ($CV-5$),

d) Metodę Leave One Out,

e) Metodę Bagging - 5.

4) Do klasyfikacji użyj Naiwny Klasyfikator Bayesa z ćwiczenia 2. Jako dane wejściowe użyj systemu australian.txt (15 atrybutów / 690 obiektów). Raport z klasyfikacji przedstaw za pomocą tablicy predykcji (patrz Tab. 1). Na podstawie tablicy predykcji określ Dokładność globalną i zbalansowaną klasyfikacji.

5) W razie problemów z programowaniem, zapoznaj się z programem demonstracyjnym C++, na stronie http://wmii.uwm.edu.pl/∼artem w zakładce Dydaktyka/Systemy Sztucznej Inteligencji. Program umieść w swoim katalogu na pulpicie.

0.1 Teoria do ćwiczeń - metody oceny jakości klasyfikatorów

Dla uproszczenia opisu metod szacowania efektywności modeli klasyfikacji, naszym wstępnym założeniem jest użycie jednego klasyfikatora, nawet gdy model składa się z kilku mniejszych pod modeli.

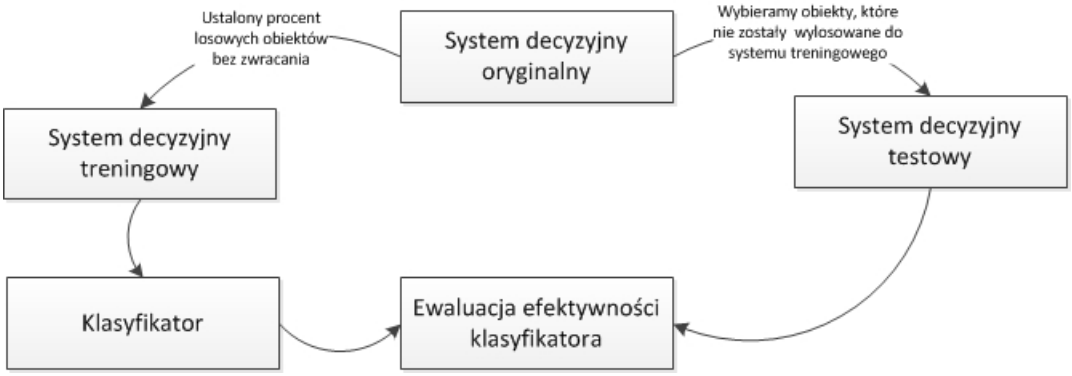
Przejdźmy do opisania pierwszej metody.

0.1.1 Trenuj i testuj (Train and Test) ($T\&T$)

Metoda Trenuj i Testuj jest klasycznym sposobem oceny jakości modeli klasyfikujących, bazuje na podziale oryginalnego systemu decyzyjnego posiadającego rozwiązania eksperta, na dwa podzbiory, system treningowy, na podstawie którego uczymy wybrany klasyfikator i system testowy, na którym sprawdzamy jego efektywność. Podział na system treningowy i testowy jest realizowany w określonych proporcjach, gdzie obiekty są losowane bez zwracania do poszczególnych systemów. Rozważany klasyfikator podejmuje decyzje na obiektach systemu testowego, oszacowanie efektywności klasyfikacji polega na porównywaniu tych decyzji z decyzjami eksperta, które są przypisane do wszystkich obiektów testowych i podczas klasyfikacji najczęściej są całkowicie ukryte. Użycie metody $T\&T$ w wielu przypadkach jest niewystarczające, wynik oszacowania może odbiegać od realnej efektywności z powodu losowego ułożenia obiektów w poszczególnych podzbiorach. Klasyfikacja może być zbyt optymistyczna, w przypadku gdy w systemie testowym i treningowym pojawią się równomiernie rozłożone podobne obiekty lub zbyt pesymistyczna, gdy do systemu treningowego i testowego wylosujemy obiekty bardzo różne. Między innymi z tego powodu metoda $T\&T$ jest stosowana tylko na dużych danych, dla uzyskania szybkiego wstępnego oszacowania efektywności klasyfikatorów, lub jako część składowa innych metod. Wizualizację metody Trenuj i Testuj możemy zobaczyć na Rys. 4

Ewaluacja efektywności klasyfikatora bazuje na określaniu parametrów jakości przeprowadzonej klasyfikacji. Jednymi z bardziej popularnych są parametry wyliczane na podstawie macierzy predykcji takie jak dokładność, pokrycie oraz trafność w klasę, dokładny opis tych parametrów mamy w sekcji 0.1.7.

Polityka budowania modelu w dużym stopniu zależy od kontekstu danych, które klasyfikujemy, w przypadku oceny jakości klasyfikacji opartej na wyliczaniu dokładności, pokrycia i trafności w klasy, w wielu przypadkach pożądanym efektem klasyfikacji jest uzyskanie jak największej wartości dokładności przy jak największym pokryciu i procencie trafności w poszczególne klasy. Warto wspomnieć, że



Rysunek 4: Wizualizacja sposobu oceny klasyfikatora metodą Trenuj i Testuj

jeżeli zależy nam na bardzo dokładnej klasyfikacji, w której każda błędna decyzja może spowodować poważne konsekwencje, ocena klasyfikacji może być ukierunkowana na uzyskanie jak największej dokładności klasyfikacji, gdzie stopień pokrycia systemu decyzyjnego testowego schodzi na drugi plan i może nie być duży. Idea, która przyświeca tego typu klasyfikacji może brzmieć, 'Lepiej nie podejmować żadnej decyzji niż podjąć decyzję złą'. Typ stosowanych parametrów i oczekiwania wobec ich wartości mogą również wynikać z rodzaju klas decyzyjnych, ich rangi oraz mocy w sensie liczby obiektów. W ćwiczeniu zakładamy, że wszystkie obiekty testowe są sklasyfikowane, czyli pokrycie systemu testowego zawsze jest równe 1.0.

Przejdźmy do krótkiego opisania metody wielokrotnego wykonywania testu $T\&T$, która może dawać realniejsze oszacowania efektywności modeli klasyfikujących.

0.1.2 Walidacja krzyżowa Monte Carlo

Walidacja krzyżowa Monte Carlo (Monte-Carlo Cross Validation) ($MCCV$) bazuje na wielokrotnym losowym podziale systemu decyzyjnego oryginalnego na część treningową i testową w ustalonych proporcjach, przy czym, obiekty są losowane bez zwracania, a liczba testów (foldów) jest liczbą przeprowadzonych podziałów. W każdym foldzie przeprowadzany jest test Trenuj i Testuj (patrz Rys. 4), efektywność klasyfikacji jest określana jako średnia efektywność ze wszystkich przeprowadzonych testów. Metoda daje lepsze oszacowanie jakości klasyfikatorów, ponieważ rozważa różnorodne podziały systemu oryginalnego, których uśrednienie zbliża nasze oszacowanie do wartości realnej.

Zważywszy na fakt, że poszczególne systemy testowe i treningowe są losowane z tego samego systemu oryginalnego, cechą szczególną metody $MCCV$ jest możliwość wylosowania tych samych obiektów do wielu systemów treningowych i testowych.

Inną konsekwencją stosowanej polityki ewaluacji, jest możliwość nie wystąpienia pewnych obiektów w systemie treningowym lub testowym. Wskazane cechy symulują sytuację rzeczywistą, w której pewne dane, obiekty mogą pojawiać się wielokrotnie, pewnej wiedzy, potrzebnej do rozwiązywania problemów możemy nie posiadać, oraz pewnych problemów do rozwiązania, możemy nie zaobserwować.

Przejdźmy do opisania metody analogicznej do metody $T\&T$, której cechą szczególną jest losowanie obiektów do systemu treningowego ze zwracaniem.

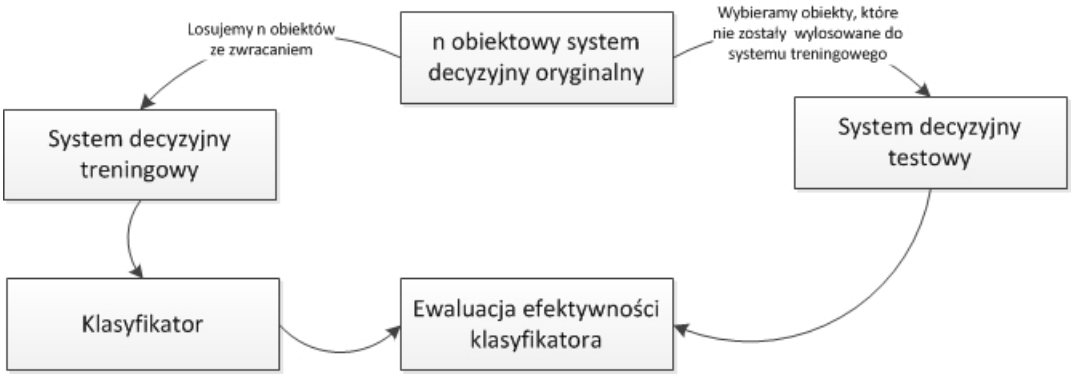
0.1.3 Bootstrap

Metodą analogiczną do metody $T\&T$ jest metoda *Bootstrap*, w której z danego zbioru n obiektów systemu oryginalnego, losuje się ze zwracaniem n obiektów tworząc zbiór treningowy, a pozostałe niewylosowane obiekty tworzą zbiór testowy.

Po zastosowaniu takiej procedury w systemie treningowym te same obiekty mogą pojawić się wielokrotnie, ze statystycznego punktu widzenia, wartość oczekiwana liczby obiektów treningowych nie mających kopii wynosi 0.632. Czyli oczekiwany podział systemu oryginalnego na system treningowy i testowy wynosi 63.2%-36.8%.

Metoda Bootstrap prowadzi do zwielokrotnienia danych, może być stosowana w mało zróżnicowanych danych w celu ich większego zróżnicowania. Przykładowym zastosowaniem jest ewaluacja efektywności klasyfikatorów, które uwzględniają te same obiekty wielokrotnie, zwiększając rangę obiektów, które występują najczęściej.

Wizualizację metody Bootstrap możemy zobaczyć na Rys. 5.



Rysunek 5: Wizualizacja sposobu oceny klasyfikatora metodą Bootstrap

Analogicznie do metody $T\&T$, metoda Bootstrap daje mało dokładne, pobieżne oszacowanie skuteczności klasyfikatorów, zależne od losowego pojedynczego podziału na system treningowy i testowy. Lepszym rozwiązaniem wydaje się wykonywanie metody Bootstrap wielokrotnie, co opiszemy w kolejnym podrozdziale.

0.1.4 Bagging

Metoda Bagging, bazuje na wielokrotnym wykonywaniu testu Bootstrap (patrz Rys. 5). Podobnie jak w metodzie $MCCV$, wyniki z każdego testu składowego są uśredniane, co prowadzi do lepszego, bliższego rzeczywistemu oszacowania efektywności rozważanego klasyfikatora.

Kolejna metoda, którą opiszemy bazuje na jak najlepszym oszacowaniu możliwości klasyfikacyjnych wewnątrz systemu decyzyjnego oryginalnego, poprzez wykonywanie wielokrotnego próbkowania w jego obrębie.

0.1.5 Walidacja Krzyżowa

Często spotykaną w literaturze alternatywną nazwą Walidacji Krzyżowej (Cross Validation) (CV) jest Próbkowanie Wielokrotne (Resampling). Metoda wielokrotnego próbkowania jest popularnym sposobem badania efektywności klasyfikatorów i może dawać dużo realniejsze wyniki ewaluacji klasyfikacji od prostego testu Train and

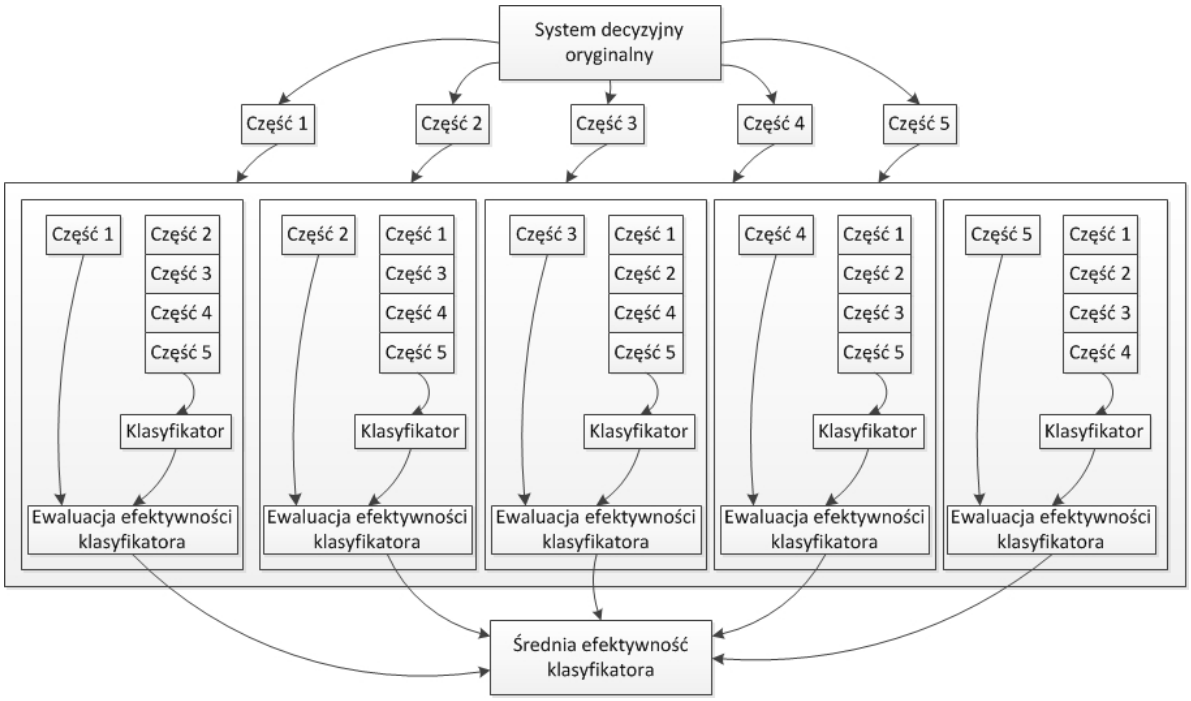
Test. Metoda polega na podziale systemu oryginalnego na ustalone k w miarę możliwości równych części, do których obiekty dobierane są w sposób losowy bez zwracania. Następnie wykonywanych jest k testów Trenuj i Testuj, gdzie za każdym razem inna z k części jest traktowana jako system testowy, a pozostałe jako system treningowy. Po przeprowadzeniu k testów, ostateczne wyniki ewaluacji są uśrednieniem ich efektywności. Metoda CV daje obraz przekrojowej klasyfikacji w całym systemie oryginalnym, jednak jej jednokrotne przeprowadzenie nie zawsze daje dobre oszacowanie skuteczności klasyfikatorów, ze względu na obciążenie związane z losowym podziałem. W celu zniwelowania tego problemu, metoda CV może być stosowana wielokrotnie do momentu uzyskania odpowiednio małego odchylenia standardowego średnich wyników.

Podczas stosowania metody CV napotykamy pewne ograniczenia. Naturalnie, liczba części na jakie możemy podzielić system decyzyjny oryginalny nie może być większa od liczby jego obiektów.

Gdy stosujemy zbyt dużą liczbę podziałów w stosunku do wielkości rozważanego systemu decyzyjnego oryginalnego, w niektórych podsystemach mogą nie występować pewne klasy systemu decyzyjnego oryginalnego. Co podczas wyliczania zbalansowanych parametrów jakości klasyfikacji (patrz sekcja 0.1.8), zmusza, do wyliczania średniej z poszczególnych klas decyzyjnych przy uwzględnianiu systemów testowych zawierających daną klasę.

Warto również wspomnieć, że gdy liczba obiektów systemu decyzyjnego oryginalnego jest podzielna przez k z resztą, pozostają nieprzydzielone obiekty, które można traktować na różne sposoby, wybrane strategie wymienimy w punktach:
• obiekty są usuwane,
• dokładamy po jednym obiekcie do wybranych części, hierarchicznie lub losowo, aż do ich wyczerpania,
• dokładamy wszystkie pozostałe obiekty do każdej części, strategia jest stosowana między innymi w systemie RSES [**?**].

Metoda Walidacji krzyżowej, dla ustalonego $k = 5$ została przedstawiona na Rys. 6.



Rysunek 6: Zobrazowanie metody CV-5; System decyzyjny oryginalny jest dzielony na 5 w miarę możliwości równych części; Wykonywanych jest pięć testów $T\&T$, ostatecznie, wyliczana jest średnia ze wszystkich pięciu oszacowań skuteczności klasyfikacji

Przejdźmy do opisania szczególnego przypadku metody Walidacji Krzyżowej.

0.1.6 Leave one out

Skrajna wersja Metody Walidacji Krzyżowej, w której tworzymy tyle testów składowych ile mamy obiektów w systemie decyzyjnym oryginalnym jest zwana (*Leave-one-out*)(Wyrzuć jeden na zewnątrz). W poszczególnych foldach tej metody pojedynczy obiekt jest traktowany jak system testowy, a zbiorem treningowym jest $(n-1)$ pozostałych obiektów. Dla ułatwienia wyliczania parametrów zbalansowanych, obiekty testowe wszystkich n foldów są traktowane jak jeden system decyzyjny i parametry są wyliczane na podstawie jego klasyfikacji.

0.1.7 Parametry oceny jakości klasyfikacji

Zacznijmy od przedstawienia parametrów, bazujących na macierzy predykcji (confusion matrix), czyli raportu z przeprowadzonej klasyfikacji, w sensie liczby obiektów przydzielanych do poszczególnych klas. Przykładowa macierz predykcji jest pokazana w Tab. 1.

0.1.8 Parametry wyliczane na bazie macierzy predykcji

Zakładając, że klasyfikujemy system decyzyjny (U, A, d), gdzie U jest uniwersum obiektów, A zbiorem atrybutów, a $d \notin A$ atrybutem decyzyjnym. W systemie decyzyjnym mamy klasy decyzyjne postaci $\{c_1, c_2, \ldots, c_k\}$. Po zakończeniu klasyfikacji

Tabela 1: Tablica predykcji - zawiera informacje o klasyfikacji poszczególnych klas systemu decyzyjnego. Wartości d_{ij}, to liczba obiektów klasy c_i, które zostały zaklasyfikowane do klasy decyzyjnej c_j

	c_1	c_2	...	c_k	
c_1	d_{11}	d_{12}	...	d_{1n}	$\|c_1\|$
c_2	d_{21}	d_{22}	...	d_{2n}	$\|c_2\|$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k	d_{n1}	d_{n2}	...	d_{nn}	$\|c_n\|$

tworzymy tablicę, w której zawieramy informacje o zgodności wyniku naszej klasyfikacji z decyzjami podjętymi przez zewnętrznego eksperta. Na podstawie tych informacji możemy określić parametry mówiące o różnych aspektach jakości klasyfikacji. Pierwszym parametrem jest ogólna jakość klasyfikacji, dokładność globalna (global accuracy) mówiąca o procencie poprawnej klasyfikacji w całym systemie decyzyjnym.

Dokładność globalna ($acc_{globalne}$)

$$acc_{globalne} = \frac{liczba\ obiektów\ poprawnie\ sklasyfikowanych\ w\ całym\ systemie\ TST}{liczba\ obiektów\ chwyconych\ w\ systemie\ TST}$$

Liczba obiektów poprawnie sklasyfikowanych w całym systemie TST jest liczbą obiektów testowych, na których decyzja podjęta przez nasz klasyfikator zgadza się z ukrytą decyzją eksperta. Liczba obiektów chwyconych w systemie TST to liczba obiektów systemu TST, które dostały jakąkolwiek decyzję, zostały sklasyfikowane.

Parametr jest często stosowany w systemach decyzyjnych o zbalansowanych, w sensie liczby obiektów, klasach decyzyjnych. Stosowanie parametru w klasach niezbalansowanych może skutkować niedostatecznym uwzględnianiem mniejszych klas decyzyjnych.

Kolejny parametr, dokładność zbalansowana (balanced accuracy), jest stosowany, gdy klasy systemu decyzyjnego są niezbalansowane pod względem mocy, wszystkie klasy są równoważnie uwzględniane, niezależnie od ich wielkości.

Dokładność zbalansowana ($acc_{zbalansowane}$)

$$acc_{zbalansowane} = \frac{acc_{c_1} + acc_{c_2} + ... + acc_{c_k}}{k}$$

Następny parametr, pokrycie globalne (global coverage) jest stosowany gdy klasy decyzyjne są podobnej wielkości, określa procent obiektów sklasyfikowanych w całym systemie decyzyjnym testowym.

Pokrycie globalne ($cov_{globalne}$)

$$cov_{globalne} = \frac{liczba\ obiektów\ chwyconych\ w\ całym\ systemie\ TST}{liczba\ obiektów\ systemu\ TST}$$

Parametr, pokrycie zbalansowane (balanced coverage) jest stosowany gdy klasy decyzyjne są niezbalansowane, uwzględnia procentowe pokrycie poszczególnych klas, jest wyliczany jako średnia ze wszystkich pokryć.

Pokrycie zbalansowane ($cov_{zbalansowane}$)

$$cov_{zbalansowane} = \frac{cov_{c_1} + cov_{c_2} + ... + cov_{c_k}}{k}$$

Jednym z istotniejszych parametrów dającym procentową trafność w klasę decyzyjną jest stopień trafności w klasę (true positive rate), definiowany w następujący sposób.

Stopień trafności w klasę decyzyjną (TPR_c)

$$TPR_c = \frac{x}{x + liczba\ obiektów\ z\ pozostałych\ klas\ błędnie\ trafiających\ do\ klasy\ c}$$

Innym ciekawym parametrem używanym do oceny jakości klasyfikacji niezbalansowanych systemów decyzyjnych binarnych jest indeks Youdena (Youden index) [**?**], [**?**] oraz współczynnik korelacji Matthew (Matthews correlation coefficient [**?**], [**?**]).

Indeks Youdena oraz współczynnik korelacji Matthew

Pierwszy parametr jest definiowany pośrednio za pomocą wartości $Sensitivity$ i $Specificity$,

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$YoudenIndex = Sensitivity + Specificity - 1$$

Kolejny jest definiowany bezpośrednio na podstawie macierzy predykcji z klasyfikacji binarnej (patrz Tab. 2).

$$MatthewsCorrelationCoefficient = \frac{TP * TN - FP * FN}{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}$$

Tabela 2: Tablica predykcji - w wierszach mamy wartości aktualne (eksperta), w kolumnach przewidziane

	TAK	NIE
TAK	TP	FP
NIE	FN	TN

TP - liczba obiektów dobrze sklasyfikowanych z decyzją TAK,
TN - liczba obiektów dobrze sklasyfikowanych z decyzją NIE,
FP - liczba obiektów źle sklasyfikowanych z decyzją TAK,
FN - liczba obiektów źle sklasyfikowanych z decyzją NIE,

Zauważmy, że dla wartości $Sensitivity = 0.8, Specificity = 0.8$, $YoudenIndex = 0.6$, ale $YoudenIndex$ przyjmuje wartość 0.6 również dla wartości $Sensitivity = 0.6, Specificity = 1.0$, jednak opcja, w której $Sensitivity$ i $Specificity$ są zbalansowane jest preferowana. Wartość $|Sensitivity - Specificity|$ jest rozsądną miarą niezbalansowania. Nawiązując do tej idei $YoudenIndex$ po uwzględnieniu zbalansowania parametrów $Sensitivity$ i $Specificity$, wygląda następująco,

$$BalancedYoudenIndex = Sensitivity + Specificity - |Sensitivity - Specificity|$$

czyli

$$BalancedYoudenIndex = min(Sensitivity, Specificity)$$

Parametry zbalansowane są rozsądnymi miarami liczenia jakości klasyfikacji binarnej, zapobiegają sytuacji, w której przydzielenie obiektom dużej liczbie etykiet w sposób przypadkowy powoduje zwiększenie indeksu Youdena.