

Politechnika Warszawska

W Y D Z I A Ł   E L E K T R Y C Z N Y



Instytut Informatyki  
Zakład Elektrotechniki Teoretycznej  
i Informatyki Stosowanej

# Praca dyplomowa magisterska

na kierunku INFORMATYKA  
w specjalności Inżynieria Systemów Informatycznych

Meta-metody służące do poprawy jakości klasyfikacji

**Konrad Ziaja**  
nr albumu 272170

promotor  
dr inż. Łukasz Skonieczny

Warszawa 2017



Warszawa, XX lutego 2017

POLITECHNIKA WARSZAWSKA  
Wydział Elektroniki i Technik Informacyjnych

### OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa magisterska pt. Meta-metody służące do poprawy jakości klasyfikacji:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Konrad Ziaja.....



# Spis treści

<b>1</b>	<b>Wstęp</b>	<b>1</b>
<b>2</b>	<b>Wstęp teoretyczny</b>	<b>3</b>
2.1	Klasyfikacja danych . . . . .	3
2.2	Wybrane algorytmy klasyfikacji danych . . . . .	4
2.2.1	Drzewo decyzyjne . . . . .	4
2.2.2	Naiwny klasyfikator bayesowski . . . . .	5
2.2.3	Klasyfikator k najbliższych sąsiadów (kNN) . . . . .	5
2.2.4	Las losowy . . . . .	5
2.2.5	Bagging . . . . .	5
2.2.6	Boosting . . . . .	5
2.2.7	Stacking . . . . .	5
2.3	Klasyfikacja danych nie zrównoważonych . . . . .	5
2.4	Dane . . . . .	6
2.4.1	Wstępne przetwarzanie danych . . . . .	7
2.4.2	Brakujące wartości atrybutów . . . . .	7
2.4.3	Usunięcie niekompletnych obserwacji . . . . .	7
2.4.4	Imputacja danych . . . . .	7
2.5	Ocena poprawności klasyfikacji . . . . .	7
2.5.1	Miary jakości klasyfikacji danych . . . . .	7
2.5.2	Metody pomiaru jakości klasyfikacji danych . . . . .	10
	<b>Bibliografia</b>	<b>13</b>



# Rozdział 1

## Wstęp

Trzeba napisać jakiś wstęp :(

# Cel pracy

I jakiś cel pracy :( (:(()))



# Rozdział 2

## Wstęp teoretyczny

### 2.1 Klasyfikacja danych

może coś pisać o uczeniu maszynowym i że klasyfikacja jest nadzorowana?

inline

Klasyfikacja jest to proces przyporządkowania danych do jednej z predefiniowanych klas na podstawie atrybutów tych danych. Algorytm klasyfikacji na podstawie analizy danych trenujących, zawierających atrybuty oraz klasę, tworzy model klasyfikacyjny. Stworzony model klasyfikacyjny wykorzystywany jest do predykcji klasy (kategorii) nowych danych bez określonej klasy. Celem algorytmu budującego model, jest odnalezienie wzorców, w jaki sposób atrybuty obiektu wpływają na przynależność do danej klasy, starając się o to, aby wiedza na temat analizowanych danych była możliwie ogólna oraz niezależna od próby.

Klasyfikacja danych jest procesem dwuetapowym:

- budowa modelu – proces ten polega na analizie obiektów z przyporządkowaną klasą oraz na budowie modelu opisującego predefiniowany zbiór klas danych,
- właściwa klasyfikacja – otrzymany model stosuje się do przydzielania klasy nowym obiektom.

Budowa modelu jest także procesem dwu-etapowym. Dzieli się ona na:

- uczenie – klasyfikator budowany jest w oparciu o dane treningowe,
- ocena jakości klasyfikacji – jakość klasyfikacji badana jest w oparciu o dane testowe.

W zależności od liczebności klas w zbiorze danych, możemy wyróżnić:

- klasyfikację binarną – klasyfikator decyduje o przypisaniu obiektu do jednej z dwóch klas (np. czy człowiek jest zdrowy lub nie)

- klasyfikację wieloklasową – obiektowi przypisuje się jedną z wielu predefiniowanych klas.

Do reprezentacji danych uczących, testowych oraz do klasyfikacji najczęściej stosuje się system informacyjny.

Zachmurzenie	Temp.	Temp. wody	Opady	Wiatr	Pływać	h(x)
słonecznie	32	25	brak	słaby	tak	tak
słonecznie	31	26	brak	umiarkowany	tak	nie
pochmurnie	22	15	brak	b. mocny	nie	nie
pochmurnie	20	18	brak	słaby	tak	tak
całkowite zachmurzenie	12	6	brak	umiarkowany	nie	nie
całkowite zachmurzenie	10	8	duże	słaby	nie	nie
pochmurnie	21	10	brak	mocny	nie	tak
słonecznie	25	17	brak	umiarkowany	tak	nie
pochmurnie	23	17	przelotne	umiarkowany	nie	tak

Tablica 2.1: Przykład danych treningowych składających się z 5 atrybutów oraz klasy decyzyjnej. W ostatniej kolumnie znajduje się wynik klasyfikacji. W pięciu przypadkach, klasyfikator poprawnie wskazał klasę.

## 2.2 Wybrane algorytmy klasyfikacji danych

### 2.2.1 Drzewo decyzyjne

Drzewo decyzyjne jest bardzo często wykorzystywane jako klasyfikator danych. Celem jest stworzenie modelu, który na podstawie danych wejściowych przewidzi poprawnie klasę. Drzewo jest acyklicznym spójnym grafem skierowanym. Korzeń (węzeł na poziomie 0) zawiera w sobie cały zbiór uczący. W każdym węźle przeprowadza się test na wartościach atrybutu, który dzieli zbiór nad podzbiory. Z węzła wychodzi tyle gałęzi ile jest możliwych wyników testu z tego węzła. Pod każdym węzłem znajduje się kryterium podziału dokonywanego w danym węźle, które jest jednakowe dla wszystkich elementów zbioru. Ostatnim elementem drzewa decyzyjnego są liście, które zawierają etykiety, czyli przydział klasowy elementów z tego podzbioru. Drzewo buduje się w sposób rekurencyjny od korzenia do liścia z wykorzystaniem metody "dziel i zwyciężaj".

#### Proces budowy drzewa

1. Stwórz korzeń zawierający cały zbiór uczący.

2. Jeśli wszystkie przykłady należą do tej samej klasy decyzyjnej, to węzeł staje się liściem z etykietą klasy.
3. Jeżeli nie, oblicz kryterium podziału wykorzystując np. entropię, które najlepiej dzieli zbiór treningowy.
4. Dla każdego testu stwórz gałąź i podziel odpowiednio podzbiory do nowych węzłów.
5. Wywołaj rekurencyjnie algorytm dla nowych węzłów.
6. Algorytm kończy się dla kryterium stopu.

Stosuje się różne kryterium stopu:

- wszystkie przykłady należą do tej samej klasy,
- brak możliwości dalszego podziału,
- zbiór pusty,
- osiągnięto zakładany cel, np.: maksymalna głębokość drzewa, maksymalna czystość klas w liściu, minimalny przyrost informacji po podziale.

Jako kryterium podziału można stosować np. wskaźnik Giniego lub entropię. Często zdarza się, że zbudowane drzewa są zbyt duże i tworzy się nadmierne dopasowanie do danych. Wtedy powinno ograniczyć się wysokość drzewa. Istnieje kilka algorytmów drzew decyzyjnych takich jak: ID3, C4.5, CART, CHAID.

wstawić obrazek drzewa decyzyjnego

### 2.2.2 Naiwny klasyfikator bayesowski

### 2.2.3 Klasyfikator k najbliższych sąsiadów (kNN)

### 2.2.4 Las losowy

### 2.2.5 Bagging

### 2.2.6 Boosting

### 2.2.7 Stacking

## 2.3 Klasyfikacja danych niezrównoważonych

Większość istniejących algorytmów klasyfikacji, nastawiona jest na poprawną klasyfikację zbiorów o zrównoważonej liczebności wszystkich klas. Niestety w rzeczywistych problemach, bardzo często zdarza się, że zbiory są mocno niezbilansowane.

## Dane nie zrównoważone

dodac odno-  
sniak do one vs  
all

Dane są nie zrównoważone jeśli klasy decyzyjne nie są przybliżeniu tak samo liczebne. Najmniejsza klasa, nazywana jest klasą mniejszościową (ang. *minority class*), natomiast klasa dominująca, lub pozostałe połączone klasy (można połączyć pozostałe klasy w jedną, doprowadzając do klasyfikacji binarnej, one vs all), nazywana jest klasą większościową (ang. *majority class*). W praktyce klasa mniejszościowa, zazwyczaj liczy około 10-20% wszystkich przykładów. Często zdarzają się jednak takie problemy, gdzie to zróżnicowanie jest większe np.:

- około 2% transakcji kartami kredytowymi w GOCARDLESS to oszustwa [3].

dodac przy-  
kłady danych  
mniejszoscio-  
wych

W przytoczonych przykładach ważniejsza jest klasa mniejszościowa i wykrycie jej stanowi priorytet. Nie zrównoważenie klas w zbiorze danych stanowi problem w fazie uczenia i znacząco obniża jakość klasyfikacji. Ze względu na częstość występowania klasy dominującej, klasyfikator preferuje tę klasę, dążąc do optymalizacji i obniżenia błędu error rate (2.5.1) nie biorąc pod uwagę rozłożenia klas w zbiorze. Klasyfikator może osiągnąć wysoką skuteczność klasyfikacji np. 95% przy niskiej lub zerowej wykrywalności klasy mniejszościowej. Należy oczekiwać od klasyfikatora wysokiej skuteczności wykrywania klasy mniejszościowej, nawet kosztem pogorszenia rozpoznawania klasy większościowej. Poddając analizie sąsiedztwa przykłady z klasy zdominowanej, można wyróżnić przykłady bezpieczne i niebezpieczne w klasyfikacji:

- safe - przykład bezpieczny, w jego sąsiedztwie zdecydowana większość obserwacji jest z tej samej klasy,
- borderline - graniczny, przykład niebezpieczny, w jego sąsiedztwie ilość przykładów z obu klas jest podobna
- outlier - poboczny, przykład niebezpieczny, w jego sąsiedztwie większość obserwacji jest z klasy przeciwnej, dominującej,
- rare - rzadki, przykład niebezpieczny, w jego sąsiedztwie występują tylko przykłady z klasy przeciwnej, większościowej.

dodac obrazek  
danych safe  
border itd.

podac wy-  
kresy przy-  
kłady danych  
niezrównowa-  
żonych

opisac metody  
klasyfikacji  
danych mniej-  
szosciowych,  
preprocessing,  
algorytmy

## 2.4 Dane

Dane nie zrównoważone (ang. *imbalanced data*) są to dane, które zawierają nie zrównoważone liczebnie klasy.

### **2.4.1 Wstępne przetwarzanie danych**

### **2.4.2 Brakujące wartości atrybutów**

Często zdarza się, że bazy danych nie są kompletne, że brakuje kilku wartości różnych atrybutów. Brakujące wartości mogą być wynikiem błędu człowieka, aplikacji, programu pomiarowego, nie podania danych lub z innego powodu. Zazwyczaj brakujące dane oznaczone są pustymi polami,? lub w inny opisany sposób. Istnieje kilka sposobów na rozwiązanie tego problemu.

### **2.4.3 Usunięcie niekompletnych obserwacji**

Najprostszym sposobem jest usunięcie wierszy lub kolumn, w których brakuje wartości. Przed usunięciem, należy przeanalizować dane, sprawdzić, które usunięcie będzie najbardziej korzystne (usunie najmniej danych). Może bowiem zdarzyć się, że zamiast usuwać dużą ilość przykładów (wiersze), bardziej opłaca usunąć się atrybut (kolumnę), który ma dużo pustych komórek. Opisana metoda niesie ze sobą niepożądane konsekwencje. Usuając, niektóre obserwacje lub atrybuty, pozbywa się części informacji. Skutkiem tego zabiegu model predykcyjny może działać słabiej. Następstwem stosowania tego sposobu jest także, w przyszłości brak możliwości predykcji niekompletnych przykładów.

### **2.4.4 Imputacja danych**

Innym pomysłem na rozwiązanie tego problemu jest imputacja danych. Brakujące dane można obliczyć lub wyznaczyć różnymi technikami na podstawie wartości pozostałych obserwacji. Jeżeli atrybut zawiera wartości ciągłe, brakujące elementy można zastąpić wartością średnią lub medianą całej kolumny. W przypadku wartości dyskretnych można uzupełnić je wartością występującą najczęściej. Stosując takie rozwiązanie można wprowadzić szum do danych. Dającym lepsze rezultaty rozwiązaniem, może być zastosowanie klasyfikatora lub regresji w celu imputacji danych.

## **2.5 Ocena poprawności klasyfikacji**

### **2.5.1 Miary jakości klasyfikacji danych**

Jakość klasyfikacji można ocenić na podstawie kilku współczynników. Do ich obliczenia wykorzystuje się macierz pomyłek (tabela 2.2). Tworzona jest ona w oparciu o wynik klasyfikacji. Dla klasyfikacji binarnej macierz składa

się z dwóch kolumn oraz dwóch wierszy. W wierszach znajdują się poprawne klasy decyzyjne, natomiast w kolumnach przewidziane przez klasyfikator. Zaklasyfikowane obiekty, umieszcza się w odpowiedniej grupie. Nazwa grup

		Klasa predykowana	
		pozytywna	negatywna
Klasa rzeczywista	pozytywna	prawdziwie pozytywna (TP)	fałszywie negatywna (FN)
	negatywna	fałszywie pozytywna (FP)	prawdziwie negatywna (TN)

Tablica 2.2: Macierz pomyłek

inspirowana była nazewnictwem medycznym. Dla dwóch wyróżniamy następujące grupy:

- prawdziwie pozytywna (ang. *true positive*), skrót TP: są to obiekty należące klasy pozytywnej oraz zakwalifikowane przez klasyfikator jako pozytywne (trafienie, z ang. *hit*)
- fałszywie negatywna (ang. *false negative*), skrót FN: są to obiekty należące klasy pozytywnej, ale zostały błędnie zakwalifikowane przez klasyfikator jako negatywne (błąd pominięcia, z ang. *miss*)
- fałszywie pozytywna (ang. *false positive*), skrót FP: są to obiekty należące klasy negatywnej, błędnie uznane przez klasyfikator jako pozytywne (fałszywy alarm, ang. *false alarm*)
- prawdziwie negatywna (ang. *true negative*), skrót TN: są to obiekty należące klasy negatywnej, i sklasyfikowane przez klasyfikator jako negatywne (poprawnie odrzucone, ang. *correct rejection*)

Ocenę jakości klasyfikacji przeprowadza się w oparciu o współczynniki wyliczane na podstawie macierzy pomyłek.

Podstawowym kryterium służącym do oceny klasyfikacji jest dokładność (ang. *accuracy*), jest to stosunek wszystkich poprawnie sklasyfikowanych przykładów klasy pozytywnej oraz negatywnej do wszystkich przykładów. Miara ta określa dokładność z jaką klasyfikator podaje poprawny wynik.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Można wyróżnić także błąd klasyfikatora, obliczany na podstawie dokładności.

$$Error\ rate = 1 - accuracy$$

Trzecim wskaźnikiem oceny klasyfikacji jest TPR (ang. *true positive rate*), często określany jako czułość (ang. *sensitivity* lub *recall*). Jest to stosunek

obiektów poprawnie sklasyfikowanych jako pozytywne z wszystkimi pozytywnymi przykładami. Wskaźnik ten pokazuje poprawność klasyfikowania obserwacji pozytywnych. W medycynie, wykorzystując tę miarę można określać skuteczność wykrywania osób chorych.

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

Kolejną miarą oceniającą klasyfikację jest TNR (ang. *true negative rate*), nazywana także specyficznością (ang. *specificity*). Wskazuje ona efektywność klasyfikowania przykładów negatywnych. Jest to stosunek poprawnie przydzielonych przykładów negatywnych do wszystkich negatywnych obserwacji. Z jej pomocą, można ocenić celność klasyfikacji osób zdrowych.

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

Istotnym wskaźnikiem jest także precyzja (ang. *precision*). Określa ona jaką część przykładów uznanych za pozytywne przez klasyfikator została poprawnie oznaczona. Precyzja wyrażana jest jako stosunek prawdziwie pozytywnych przypadków do wszystkich przykładów uznanych za pozytywne. W medycynie, pokazuje procentowo ile osób uznanych za chorych, jest rzeczywiście chora.

$$precision = \frac{TP}{TP + FP}$$

wstawić przykład klasyfikacji dn o wysokiej skuteczności (np 95), ale o niskim wykrywaniu małej klasy Wskaźnik dokładności oraz error rate nie sprawdzają się w przypadku, gdy dane są nie zrównoważone. Klasyfikator może osiągnąć wysoką dokładność np. 90% przy niskiej wykrywalności klasy mniejszościowej. Dlatego oceniając klasyfikator pracujący na nie zrównoważonych danych, należy obliczyć osobno współczynniki precyzji, czułości oraz specyficzności dla każdej kategorii danych. Jak wspomniano wcześniej, bardzo często polepszenie jakości klasyfikacji klasy mniejszościowej połączona jest z pogorszeniem rozpoznawalności klasy większościowej. Mając współczynnik czułości oraz specyficzności ciężko zdecydować, który klasyfikator jest lepszy. Kubat i Matwin zaproponowali połączenie obu tych współczynników, w postaci średniej geometrycznej czułości oraz specyficzności [1].

inline

wycięta o klasyfikacji

$$G - mean = \sqrt{precision * recall}$$

Klasyfikator z wyższym G-mean, zapewnia lepszą rozpoznawalność obu klas, jednocześnie zachowując, aby dokładność w rozpoznawaniu obu klas była

zbilansowana. Współczynnik ten jest niezależny od rozkładu klas w danych [2].

Ocenę klasyfikacji danych nie zrównoważonych możemy dokonać także przy pomocy F-measure. Jest to średnia harmoniczna precyzji oraz czułości. Współczynnik F-measure można obliczyć dla obu klas.  $\beta$  wykorzystywana jest do określenia zależności pomiędzy precyzją oraz czułością.

$$F - measure = \frac{(1 + \beta)^2 * precision * recall}{\beta^2 * precision + recall}$$

Zazwyczaj  $\beta = 1$ , wtedy:

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

- inline napisac o roznych F i o tym jak wyniki wyglada, pokazać test
- inline napisc o roc
- inline napisac gdzieś o nadmiernym dopasowaniu -> walidacja krzyżowa

## 2.5.2 Metody pomiaru jakości klasyfikacji danych

W celu oceny klasyfikatora powinno wykorzystywać się dwa zbiory, treningowy oraz testowy. Najpierw należy zbudować model w oparciu o dane testowe, a następnie wykonać klasyfikację testową w oparciu o zbiór testowy. W celu poprawnej oceny klasyfikacji zbioru testowego konieczna jest znajomość odpowiedniej przynależności jego składników do klas oraz zestawienie jej z przyporządkowaniem składników do klas, które zostały zasugerowane przez klasyfikator. Następnie buduje się macierz pomyłek w oparciu o sklasyfikowane przypadki. Kolejnym krokiem jest obliczenie opisanych wyżej współczynników w oparciu o tę macierz. Istnieją różne schematy postępowania, służące do oceny zbudowanego modelu. może cos napisac jeszcze o celu [https://en.wikipedia.org/wiki/Cross-validation\\_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))

### Metoda z jednym zbiorem

Do budowy klasyfikatora wykorzystywany jest cały zbiór dostępnych danych. W procesie testowania, bierze udział także cały zbiór danych. Metoda ta, nie jest zbyt wartościowa i prowadzi do zawyżenia jakości klasyfikatora. W przypadku nowych danych, taki model osiągnie gorsze wyniki niż wskazywałyby na to obliczone współczynniki.



### **Metoda z wydzielonym zbiorem testowym (ang. *the holdout method*)**

W tej metodzie, zbiór danych dzielony jest w sposób losowy na dwie części. Użytkownik dobiera rozmiar zbioru uczącego (np. 80%) oraz zbioru testowego (np. 20%). Wadą jest, że nie wiadomo ile obiektów danej klasy znajdzie się w zbiorze testowym oraz, że zostaje zmniejszony zbiór uczący. Może to doprowadzić do sytuacji nadmiernego dopasowania (zawyżonych wyników) lub do niedoszacowania klasyfikatora. Ważne jest, aby nie używać ciągle tego samego zbioru testowego do wyboru modeli, ale dokonywać losowania przed każdą oceną.

Ulepszeniem tej metody, może być równy rozkład klas w obu zbiorach, tak aby zostały zachowane proporcje z oryginalnego zbioru.

### **Sprawdzian krzyżowy z p przykładami (ang. *leave-p-out cross-validation*)**

Sprawdzian krzyżowy z p przykładami wykorzystuje p obserwacji jako zbiór testowy, pozostałe elementy tworzą zbiór uczący. Cały proces jest powtarzany do momentu stworzenia i przetestowania wszystkich możliwych kombinacji p przykładów ze zbioru n. Ten rodzaj metody wymaga uczenia i testowania klasyfikatora  $\binom{n}{p}$  razy, gdzie n to liczebność całego zbioru danych. W przypadku dużego zbioru danych oraz  $p > 1$ , obliczenia mogą zająć bardzo dużo czasu, a nawet ze względu na dużą ilość kombinacji, obliczenie ich może być niemożliwe.

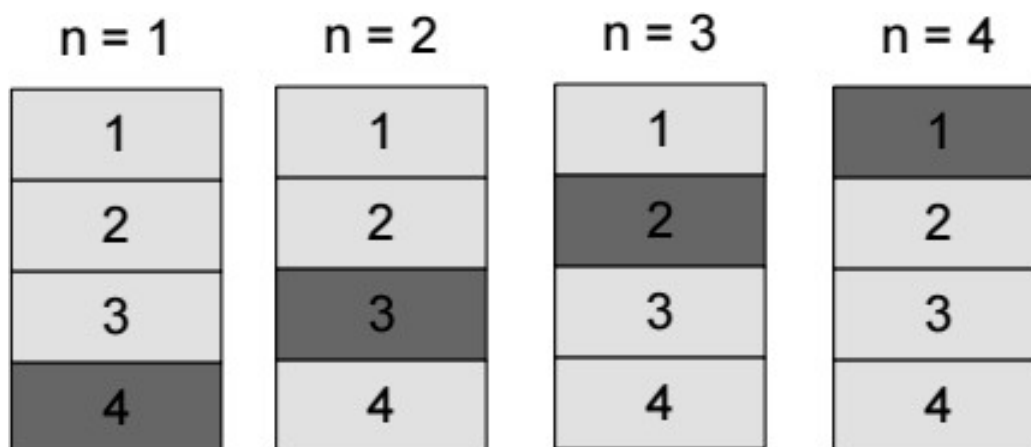
### **Sprawdzian krzyżowy minus jeden element (ang. *leave-one-out cross-validation*)**

Jest to specjalny przypadek sprawdzaniu krzyżowego z p przykładami, dla  $p = 1$ . W tej metodzie zbiór testowy tworzy jeden element, pozostałe tworzą zbiór uczący. Testowania klasyfikatora trwa do momentu użycia wszystkich obserwacji jako zbioru testowego. W przeciwieństwie do poprzedniej metody, ta jest wolna od czasochłonnych obliczeń, gdyż  $\binom{n}{1} = n$ , gdzie n to liczba wszystkich obserwacji. Zazwyczaj ta metoda wykorzystywana jest tylko do małych zbiorów danych.

### **Sprawdzian krzyżowy k-krotny (ang. *k-fold cross-validation*)**

Zbiór danych jest losowo dzielony na k równych podzbiorów. Następnie każdy z podzbiorów w kolejnych k iteracjach staje się kolejno zbiorem testowym, pozostałe zbiory tworzą zbiór uczący, na podstawie, którego buduje się

model. Klasyfikacja i testowanie wykonywane są k-krotnie. Otrzymane wyniki łączy się i uśrednia w celu uzyskania jednego wyniku. Zaletą tej metody jest mały błąd estymacji oraz niższa wariancja błędu niż w przypadku metody minus jednego elementu. Zwykle stosuje się  $k=3..10$ , dla których koszt czasowy jest umiarkowany.



Rysunek 2.1: Przykład sprawdzianu krzyżowego k-krotnego,  $k=4$ .

### Równomierny sprawdzian krzyżowy k-krotny (ang. *Stratified k-fold cross-validation*)

Jest to specjalny przypadek sprawdzianu krzyżowego k-krotnego. Podzbiory tworzone są z zachowaniem proporcji wszystkich klas. Każdy podzbiór powinien zawierać w przybliżeniu podobny procent obserwacji z każdej kategorii.

napisac o tym, że niektóre zbiory są multiklasowe, natomiast dane sprowadzone są do 2 klas

napisac o podejsci one vs all

# Bibliografia

- [1] M. Kubat i S. Matwin. Addressing the curse of imbalanced training sets: one-side selection.
- [2] H. He i E. A. Garcia. Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering, 2009.
- [3] N. Hockham: Machine learning with imbalanced data sets  
<https://www.youtube.com/watch?v=X9MZtvvQDR4>