# ZAAWANSOWANE UCZENIE MASZYNOWE

# KONSPEKT WSTĘPNY

# TEMAT NR 7: AKTYWNE UCZENIE

**Arkadiusz Koszewski** **Kornel Mrozowski**

28 maja 2021

## Spis treści

**1 Opis problemu** **2**

1.1 Treść Zadania . . . . . 2

1.2 Wstęp . . . . . 2

1.3 Zarys Teoretyczny . . . . . 2

1.3.1 Scenariusze . . . . . 2

1.3.2 Strategie selekcji zapytań . . . . . 4

**2 Propozycja rozwiązania** **5**

2.1 Cel projektu . . . . . 5

2.2 Algorytm . . . . . 5

2.3 Miary jakości i procedury oceny modeli . . . . . 5

2.4 Propozycje strategii selekcji zapytań . . . . . 6

2.5 Propozycje danych wejściowych do problemu klasyfikacji . . . . . 6

**3 Badanie wpływu aktywnego uczenia się na jakość klasyfikacji i regresji** **7**

**4 Wybór technologii** **7**

# 1 Opis problemu

## 1.1 Treść Zadania

Zintegrowane i zautomatyzowane aktywne uczenie się przy tworzeniu modeli klasyfikacji lub regresji na podstawie małych zbiorów trenujących przez zgłaszanie zapytania o prawdziwe wartości atrybutu docelowego dla ograniczonej liczby przykładów z dostarczonego dużego zbioru danych nieetykietowanych wybranych według określonych kryteriów (np. przykłady bliskie granicy decyzyjnej dotychczasowego modelu lub takie, dla których jego predykcje są obarczone największą niepewnością) i iteracyjne doskonalenie modelu na podstawie powiększanego w ten sposób zbioru trenującego. Implementacja w formie opakowania umożliwiającego użycie dowolnego algorytmu klasyfikacji lub regresji dostępnego w R stosującego interfejs formuły (po dostosowaniu jego metody predykcji tak, aby dostarczała informacji pozwalających na wybór przykładów do zapytania). Badanie wpływu użycia aktywnego uczenia się na jakość modeli klasyfikacji i regresji tworzonych za pomocą algorytmów dostępnych w R na podstawie małych zbiorów trenujących.

## 1.2 Wstęp

W dzisiejszych czasach często spotyka się z mnogością nieoznaczonych danych pochgodzących z internetu, ze świata akademickiego lub z biznesu. Dane nieoznaczone są stosunkowo łatwe do pozyskania i kosztowne w etykietowaniu. Firmy będące w posiadaniu takich danych zwykle zatrudniają eksperta lub kilku pracowników, których celem jest ich oznaczanie [1]. Na przykład: firma medyczna ma wiele skanów MRI i musi zatrudnić eksperta, który pomoże zinterpretować te skany. Firma ma ograniczone zasoby i nie może zinterpretować ani oznaczyć wszystkich swoich danych; w tym momencie decydują się na *aktywne uczenie się* ($\mathcal{AL}$). Firma zajmująca się $\mathcal{AL}$ obiecuje, że poprzez iteracyjne zwiększanie rozmiaru starannie wybranych oznaczonych danych możliwe jest osiągnięcie podobnej (lub większej [2]) dokładności uczonego modelu, niż przy użyciu w pełni oetykietowanego zestawu danych. $\mathcal{AL}$ jest uważana za metodę częściowo nadzorowaną, ponieważ używamy próbek oznaczonych w zakresie 0 do 100 procent.

## 1.3 Zarys Teoretyczny

**Aktywne uczenie się** to szczególny przypadek uczenia maszynowego, w którym algorytm uczący się może interaktywnie wysyłać zapytania do użytkownika (lub innego źródła informacji), aby oznaczyć nowe punkty danych żądanymi wynikami. Źródło informacji jest również nazywane nauczycielem lub wyrocznią. Główna hipoteza w aktywnym uczeniu się jest taka, że jeśli algorytm uczący się może wybierać instancje danych, z których chce się uczyć, może działać lepiej niż tradycyjne metody ze znacznie mniejszą ilością danych do szkolenia. Proces selekcji tych instancji na podstawie danych, które do tej pory zostały zebrane, nazywa się *aktywnym uczeniem się*.

### 1.3.1 Scenariusze

W *aktywnym uczeniu się* występują zazwyczaj trzy scenariusze lub też możliwości zgłaszania instancji nieoznaczonego zbioru dancyh do nadania im etykiet:

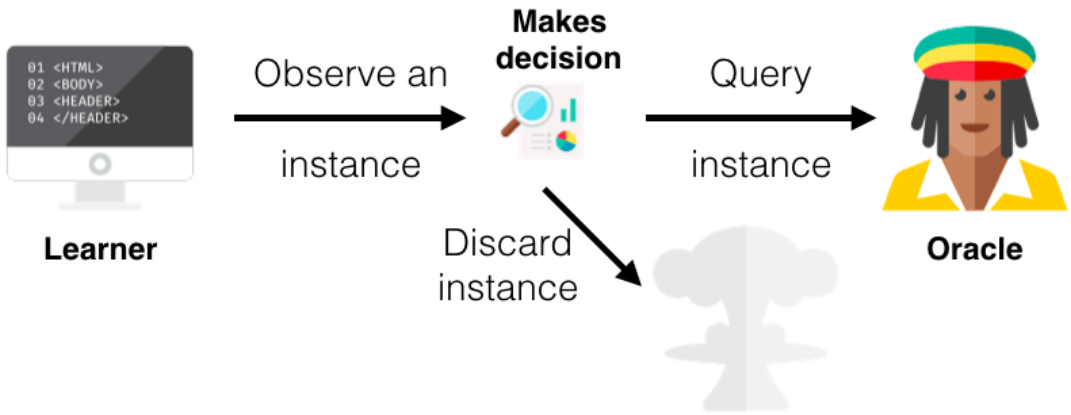
**Synteza zapytań o członkostwo**

Uczeń generuje lub konstruuje instancję przy użyciu rozkładu normalnego. Na przykład, jeśli dane są obrazami cyfr, uczący się utworzy obraz podobny do cyfry, który może zostać obrócony lub część cyfry może zostać wykluczona. Ten utworzony obraz jest wysyłany do wyroczni w celu oznaczenia.


Rysunek 1: Synteza zapytań o członkostwo

**Próbkowanie selektywne oparte na strumieniu**

W tym scenariuszu zakłada się, że uzyskanie kolejnej nieetykietowanej danej jest tanie. Opierając się na tym założeniu, wybiera się pojedynczo każdą nieoznaczoną instancję i pozwala uczniowi określić, czy chce zapytać o etykietę instancji, czy odrzucić ją na podstawie jej informatywności (pouczalności przykładu). Aby określić informatywność instancji, należy użyć strategii zapytań. Zgodnie z powyższym przykładem należy wybrać np. jeden obraz z zestawu obrazów bez etykiety, określić, czy należy go oznaczyć, czy odrzucić, a następnie przejść do kolejnego obrazu.

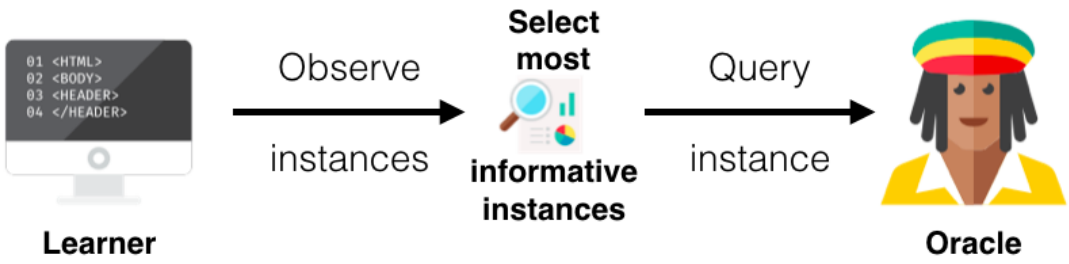



Rysunek 2: Próbkowanie selektywne oparte na strumieniu

**Próbkowanie oparte na puli**

W tym scenariuszu zakłada się, że istnieje duża pula nieoznaczonych danych, podobnie jak w przypadku próbkowania selektywnego opartego na strumieniu. Instancje są następnie pobierane z puli zgodnie z pewną miarą informatywności. Ta miara jest stosowana do wszystkich instancji w puli (lub niektórych podzbiorów, jeśli pula jest bardzo duża), a następnie wybierane są instancje zawierające najwięcej informacji. Jest to najczęstszy scenariusz w aktywnej społeczności

uczącej się. Kontynuując przykład w dwóch powyższych scenariuszach, wszystkie nieoznaczone obrazy cyfr zostaną uszeregowane, a następnie zostaną wybrane najlepsze (najbardziej pouczające) instancje i zażądane ich etykiety.




Rysunek 3: Próbkowanie oparte na puli

### 1.3.2 Strategie selekcji zapytań

Najważniejszym narzędziem w metodzie AL jest funkcja doboru próby, jest to jedyny punkt, w którym wpływamy na proces uczenia się i kluczowy jest wybór właściwej metody. Ten obszar jest gorącym tematem badawczym i istnieje wiele badań, w których proponuje się konkurencyjne funkcje selekcji.
Strategie selekcji zapytań dla metod klasyfikacji:

- **Random Selection** – Instancja jest wybierana ze zbioru walidacyjnego w sposób losowy.
- **Lowest Confidence**(LC) – W tej strategii uczeń wybiera instancję, w przypadku której ma najmniejszą pewność swojej **najbardziej** pewnej etykiety.
- **Margin Selection** – Wadą strategii **LC** jest to, że bierze pod uwagę tylko najbardziej prawdopodobną etykietę i pomija inne prawdopodobieństwa etykiet. Strategia próbkowania marginesu stara się przezwyciężyć tę wadę, wybierając wystąpienie, które ma najmniejszą różnicę między pierwszą a drugą najbardziej prawdopodobną etykietą.
- **Entropy Selection** – Aby wykorzystać wszystkie możliwe prawdopodobieństwa etykiet, używasz popularnej miary zwanej entropią. Formuła entropii jest stosowana do każdego wystąpienia i odpytywane jest wystąpienie o największej wartości.

Strategie selekcji zapytań dla metod regresji:

- **Expected Model Change Sampling** – wybiera instancję która spowodowałaby największą zmianę w modelu gdybyśmy znali jej etykietę. Jedną ze strategii stosujących to podejście jest Epected Gradient Length (EGL). Zmiana w modelu jest równoznaczna z długością gradientu trenującego (wektora służącego do wyznaczenia nowych wartości parametrów).
- **Query By Committee** (QBC) – Podejście QBC obejmuje utrzymywanie komitetu $\mathcal{C} = \{(1), ..., (C)\}$ modeli, z których wszystkie są szkolone w zakresie danych ze zbioru etykietowanego, ale różnią się między sobą w założeniu tym, że reprezentują konkurencyjne hipotezy. Każdy członek komisji może następnie głosować nad etykietami kandydatów do zapytań. Za najbardziej pouczające zapytanie uważa się przypadek, z którym najbardziej się nie zgadzają z sobą.

## 2 Propozycja rozwiązania

### 2.1 Cel projektu

Celem projektu jest stworzenie pakietu w R umożliwiającego proste testowanie algorytmu zautomatyzowanego zintegrowanego aktywnego uczenia. Pakiet ma umożliwiać stosowanie i testowanie różnych algorytmów klasyfikacji na różnych zbiorach danych.

Poniższy algorytm będzie testowany w procesie uczenia modelu klasyfikującego dane na podstawie atrybutów.

### 2.2 Algorytm

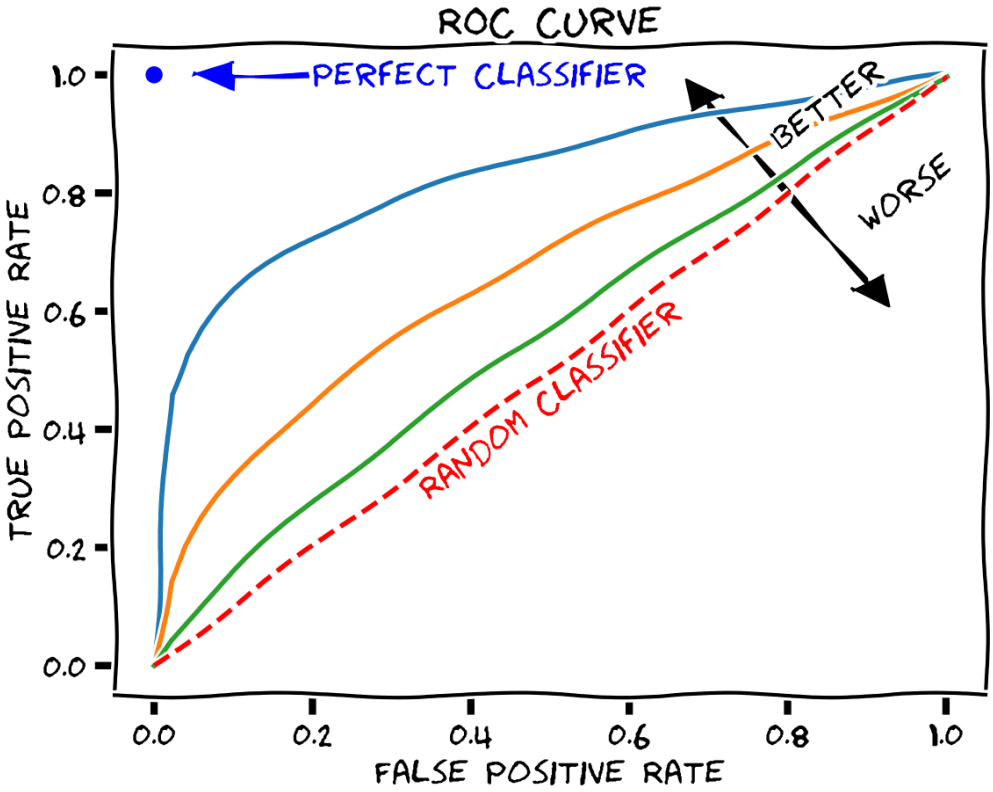
Poniżej przedstawiono listę kroków algorytmu:

```
1. Podzielenie zbioru na pulę oraz zbiór testowy
2. Wybranie k próbek z puli do początkowego zbioru trenującego i podpisanie ich,
pozostałe dane stanowią zbiór walidacyjny
3. Normalizacja wszystkich zbiorów
4. Trenowanie modelu za pomocą zbioru trenującego
5. Użycie modelu na zbiorze walidacyjnym w celu otrzymania prawdopodobieństw dla każdej próbki
6. Użycie modelu na zbiorze testowym i zmierzenie wydajności
7. Wybranie k próbek niosących najwięcej informacji bazując na uzyskanych prawdopodobieństwach
(powinny zostać wybrane te próbki dla których klasyfikator zwrócił najniższe prawdopodobieństwo
tzn. był w ich przypadku najmniej "pewny").
8. Przeniesienie wybranych próbek ze zbioru walidacyjnego do testowego
i sprawdzenie ich etykiet.
9. Odwrotna normalizacja wszystkich zbiorów danych
10. Koniec w przypadku spełnienia warunku kończącego,
w przeciwnym razie powrót do punktu 3
```

**Warunek kończący** – Algorytm kończy iteracje kiedy zostanie osiągnięty limit zapytań wystosowanych do wyroczni.

### 2.3 Miary jakości i procedury oceny modeli

Wykorzystaną procedurą oceny modeli będzie krzywa Receiver Operating Characteristics. ROC oznacza charakterystykę działania odbiornika, a wykres jest wykreślany względem TPR (True-Positive Rate) i FPR (False-Positive Rate) dla różnych wartości progowych. Wraz ze wzrostem TPR wzrasta również FPR. Jak widać na pierwszym rysunku, mamy cztery kategorie i chcemy, aby wartość progowa prowadziła nas bliżej lewego górnego rogu. Porównywanie różnych predyktorów (tutaj 3) na danym zbiorze danych również staje się łatwe, jak widać na rysunku 2, można wybrać próg w zależności od aplikacji. ROC AUC to tylko obszar pod krzywą, im wyższa wartość liczbowa, tym lepiej.




Rysunek 4: Receiver Operating Characteristics

### 2.4 Propozycje strategii selekcji zapytań

W przypadku modelu klasyfikacji proponowanymi strategiami selekcji zapytań są: **Random Selection**, **Lowest Confidence**, **Margin Selection**, **Entropy Selection**. Dla modelu regresji proponowanymi strategiami selekcji zapytań są: **Expected Model Change Sampling** oraz **Query By Committee**.

### 2.5 Propozycje danych wejściowych do problemu klasyfikacji

Do przetestowania algorytmu zamierzamy użyć następujących zbiorów danych:

- **Klasyfikacja:** Job Change of Data Scientists (Kaggle) – predykcja odchodzenia z pracy specjalistów data science
- **Regresja:** Bike Sharing Demand – predykcja zapotrzebowania na bike-sharing na podstawie

## 3 Badanie wpływu aktywnego uczenia się na jakość klasyfikacji i regresji

**Hipoteza badawcza**

Użycie algorytmu aktywnego uczenia się nie wpłynie znacząco na jakość klasyfikacji i regresji za to znacznie ograniczy potrzebną liczbę nadanych etykiet.
Porównujemy algorytm pasywnego uczenia się z aktywnym przy wykorzystaniu dwóch modeli uczących się:

- maszyna wektorów nośnych (SVM)
- losowy las (RF)

Oba algorytmy zostaną wykonane z jedną lub wszystkimi strategiami selekcji przy użyciu wszystkich „k" = [10, 25, 50, 125, 250], łącznie 20 do 60 eksperymentów. Ze względu na losowy charakter niektórych algorytmów i funkcji selekcyjnych wskazane jest przeprowadzenie w kodzie powtórnych eksperymentów w celu obliczenia wyniku istotnego statystycznie.

## 4 Wybór technologii

Projekt zostanie zrealizowany w języku R. Planujemy wykorzystać następujące pakiety:

- *randomForest* - do algorytmu klasyfikacji Random Forest (Losowy Las)
- *e1071* - do algorytmu klasyfikacji Support Vector Machine (Maszyna Wektorów Nośnych)
- *ggplot2* - do rysowania wykresów
- *devtools* - do tworzenia pakietów
- *roxygen2* - do dokumentacji pakietów
- *dplyr, tidyr* - do przekształcania danych
- *stats* - do obliczeń statystycznych, m. in. do regresji
- *lubridate* - do pracy na datach

## Bibliografia


[1] Stefan Hosein *Active Learning: Curious AI Algorithms*.

[2] Shay Yehezkel *High Dimensional Statistical Process Control and Application, M.Sc Thesis.*.

[3] Ilhan, Hamza Osman, and Mehmet Fatih Amasyali. *Active Learning as a Way of Increasing Accuracy.*. International Journal of Computer Theory and Engineering 6, no. 6 (2014): 460.

## Oświadczenia

Potwierdzam samodzielność powyższej pracy oraz niekorzystanie przeze mnie z niedozwolonych źródeł.

Oświadczam, że niniejsza praca stanowiąca podstawę do uznania osiągnięcia efektów uczenia się z przedmiotu Zaawansowane Uczenie Maszynowe została wykonana przeze mnie i osoby z sespołu samodzielnie.

Kornel Mrozowski

Nr albumu: 285776

Potwierdzam samodzielność powyższej pracy oraz niekorzystanie przeze mnie z niedozwolonych źródeł.

Oświadczam, że niniejsza praca stanowiąca podstawę do uznania osiągnięcia efektów uczenia się z przedmiotu Zaawansowane Uczenie Maszynowe została wykonana przeze mnie i osoby z sespołu samodzielnie.

Arkadiusz Koszewski

Nr albumu: 283435