Autorzy

Mateusz Ostaszewski 325203
Michał Sadowski 325221

Zadanie

Połączenie lasu losowego z SVM w zadaniu klasyfikacji. Postępujemy tak jak przy tworzeniu lasu losowego, tylko co drugi klasyfikator w lesie to SVM. Jeden z klasyfikatorów (SVM lub drzewo ID3) może pochodzić z istniejącej implementacji.

Interpretacja i doprecyzowanie treści zadania

Celem zadania jest stworzenie hybrydowego modelu klasyfikatora, który łączy drzewa ID3 i maszyny wektorów nośnych (SVM).

Algorytmy

1. Drzewo decyzyjne (ID3): Zaimplementujemy algorytm ID3 do budowy drzew decyzyjnych, który wybiera podział w węźle na podstawie maksymalizacji zysku informacyjnego (information gain).
2. SVM (Support Vector Machine): Będziemy korzystać z dostepnej implemenacji SVM z biblioteki scikit-learn.

Integracja w modelu hybrydowym:

- Dla każdego klasyfikatora generujemy losowy podzbiór danych treningowych oraz losowy podzbiór cech.
- Co drugi klasyfikator jest zastępowany SVM. Wynik końcowy jest określany na podstawie głosowania większościowego.

Opis algorytmów

ID3

1. Na wejściu mamy zbiór danych treningowych D o n przykładach, każdy z m cechami i przypisaną klasą.
2. W każdym węźle obliczamy zysk informacyjny dla każdej cechy:

$$IG(A) = H(D) - \sum_{v \in \text{Wartości}(A)} \frac{|D_v|}{|D|} H(D_v)$$

- gdzie:
  H(D) - entropia zbioru danych D obliczana jako

$$H(D) = -\sum_k p_k log_2(p_k)$$

- Dv - podzbiór danych dla wartości v cechy A

3. Wybieramy cechę A o maksymalnym IG(A) jako kryterium podziału.
4. Tworzymy nowe węzły dla każdej wartości cechy A.
5. Powtarzamy proces, aż wszystkie przykłady w węźle należą do tej samej klasy lub osiągniemy maksymalną głębokość drzewa.

SVM

1. Na wejściu mamy zbiór danych:
   - wektor cech

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

   - etykieta klasy

$$y_i \in \{-1, 1\}$$

2. Optymalizujemy funkcję celu:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i$$

   z ograniczeniami:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, n$$

   gdzie:
   - w - wektor wag,
   - b - bias,
   - x_i -: zmienne slack pozwalające na błędną klasyfikację,
   - C - parametr regularyzacji.
   - ξ_i - zmienne tolerancji

3. W przypadku nieliniowych danych stosujemy funkcję jądra RBF:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

Integracja (las hybrydowy)

1. Wygenerujemy N klasyfikatorów
   - Dla nieparzystych i: stwórzymy drzewo **ID3** na losowym podzbiorze danych i cech.
   - Dla parzystych i: wytrenujemy **SVM** na losowym podzbiorze danych i cech.
2. Podczas predykcji przeprowadzimy głosowanie większościowe wśród klasyfikatorów.
   - Dla każdej próbki wybierzemy klasę y z największą liczbą głosów.

Planowane eksperymenty

Zbiory danych

1. Iris

- Liczba przykładów: 150
- Liczba klas: 3
- Liczba cech: 4 (długość/szerokość kielicha i płatka)

2. Wine Quality

- Liczba przykładów: 4898 (wino białe), 1599 (wino czerwone)
- Liczba klas: od 0 do 10 – oceny jakości wina
- Liczba cech: 11 (chemiczne właściwości)

3. Telecom Churn

- Liczba przykładów: 3150
- Liczba klas: 2 (churn lub nie churn)
- Liczba cech: 21 (informacje o klientach, użycie usług).

Walidacja modeli zostanie przeprowadzona za pomocą walidacji krzyżowej z podziałem na 5 podzbiorów.

Eksperymenty numeryczne

Wpływ parametrów modelu na skuteczność

Cel: Analiza wpływu parametrów modeli SVM i ID3 na skuteczność hybrydowego lasu.

- **Opis:**
  W tym eksperymencie skupimy się na testowaniu różnych wartości hiperparametrów dla SVM i ID3. Skuteczność będzie oceniana pod kątem dokładności i F1 score.
- **Hiperparametry do testowania:**
  - **SVM:**
    - C: wartość regularyzacji (np. 0.01, 0.1, 1, 10, 100).
    - gamma: współczynnik w funkcji RBF (np. 0.001, 0.01, 0.1, 1).
  - **ID3:**
    - Maksymalna głębokość drzewa: np. 3, 5, 10, brak limitu.
    - Minimalna liczba próbek w liściu: np. 1, 5, 10.
    - Minimalna liczba próbek do podziału węzła: np. 2, 10, 20.

Skuteczność hybrydowego modelu

Cel: Porównanie skuteczności hybrydowego lasu (ID3 + SVM) z klasycznym lasem losowym oraz samym SVM.

- **Opis:**
  Modele będą testowane na trzech wybranych zbiorach danych (Iris, Wine Quality, Telecom Churn). Dla każdego zbioru porównamy dokładność, F1, precyzję i czułość trzech wariantów klasyfikatorów:
  1. Hybrydowy model (ID3 + SVM) - najlepszy(wyłoniony podczas poprzednich eksperymentów).
  2. Klasyczny las losowy z drzewami ID3.
  3. Sam SVM jako oddzielny klasyfikator.

Podane wartości hiperparametrów są orientacyjne po przeprowadzeniu eksprymentów dostosujemy przestrzeń przeszukiwań parametrów w celu znalezianie jak najlepszego modelu

Każdy z eksprymentów zostanie przeprowadzony kilka razy a wyniki zostaną uśrednione.

Metryki

- $$\text{Dokładność} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- $$\text{F1} = 2 \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{\text{Precyzja} + \text{Czułość}}$$
- $$\text{Precyzja} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- $$\text{Czułość} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
- Macierz pomyłek