Projekt: Nienadzorowana detekcja anomalii z wykorzystaniem globalnych i lokalnych wskaźników niepodobieństwa

Autorzy

Mateusz Krakowski

Bartosz Latosek

Temat projektu

Nienadzorowana detekcja anomalii na podstawie globalnych i lokalnych wskaźników niepodobieństwa do sąsiadów z możliwością użycia dowolnej miary niepodobieństwa. Porównanie z nienadzorowaną detekcją anomalii za pomocą algorytmów klasyfikacji jednoklasowej dostępnych w środowisku R lub Python.

Cel projektu

Implementacja i analiza efektywności nienadzorowanych metod detekcji anomalii z wykorzystaniem:

- **Globalnych wskaźników niepodobieństwa** (analiza w kontekście całego zbioru danych)
- **Lokalnych wskaźników niepodobieństwa** (analiza w kontekście najbliższego sąsiedztwa)
- **Połączenia obu podejść**
- **Porównanie** z klasycznymi metodami klasyfikacji jednoklasowej

Zakres tematyczny

- Implementacja własnego algorytmu detekcji anomalii
- Walidacja na danych ze zbiorów "donor" i "fraud"
- Analiza porównawcza z istniejącymi rozwiązaniami

Istniejące rozwiązania które będą porównywane do naszego rozwiązania

- Isolation Forest
- One-Class SVM

oba te rozwiązania są dostępne w bibliotece scikit-learn

Opis implementacji

```
class NeighborAnomalyDetector:
    def __init__(self, method='combined', metric = 'euclidean', contamination =
0.172, local_n_neighbors=10, global_n_neighbors=10): ...
    def fit(self, X): ...
    def predict(self, X): ...
```

Nasz algorytm zamknie się w jednej klasie, podczas inicjalizacji instancji klasy zostaną wybrane zmienne takie jak:

- method - metoda wykorzystująca tylko lokalny wskaźnik, tylko globalne wskaźnik lub oba
- metric - metryka oceny odległości
- contamination - parametr algorytmu określający jaką cześć danych zaklasyfikować jako anomalię, jest to liczba arbitralnie wybrana przez nas, na potrzebę eksperymentów ustalona ona zostanie jako procent anomalii w ogóle danych na których testujemy algorytm

Opis wskaźników odległości

- lokalny - będzie to odległość od k-tego sąsiada, k ustalane na podstawie local_n_neighbors
- globalny, będzie to średnia odległość od k sąsiadów, k ustalane na podstawie global_n_neighbors

Pytania do badań:

- Która metoda (globalna/lokalna/połączona) zapewnia najlepsze wyniki?
- Jak nasz algorytm radzi sobie w porównaniu z innymi metodami?
- Jak zmiana hiperparametrów wpływa na jakość detekcji anomalii?

Miary jakości użyte do porównania modeli

1. Recall (Czułość, Sensitivity)
2. Precision (Precyzja)
3. F1-Score
4. AUC-ROC (Area Under ROC Curve)
5. AUC-PR (Area Under Precision-Recall Curve)
6. Accuracy Najważniejszą metryką w tym przypadku problemu zdaje się Recall, gdyż zależy nam na tym aby wykrywane były anomalie, koszt zaklasyfikowania nie anomalii jako anomalię jest znacznie mniejszy niż pominięcie wykrycia faktycznej anomalii. Accuracy w tym przypadku jest pomijalną metryką oceny, gdyż anomalie występują rzadko, a co za tym idzie zbiory danych będą niezbalansowane.

Opis zbioru danych dotyczących darowizn

Charakterystyka zbioru danych

Zbiór danych dotyczy zbiórek pieniędzy i zawiera informacje o darowiznach oraz cechy charakterystyczne darczyńców. Dane służą do przewidywania sukcesu zbiórek lub analizy zachowań darczyńców.

Struktura danych

Zbiór zawiera **10 cech** wejściowych i **1 zmienną docelową** (`class`). Każdy wiersz reprezentuje jeden przypadek darowizny. Zmienna docelowa oznacza, czy zbiórka jest zaklasyfikowana jako wyjątkowo udana.

Opis cech

Cechy binarne (tak/nie)

Nazwa cechy	Opis	Wartości
at_least_1_teacher_referred_donor	Czy do zbiórki donacje dała osoba która dostała linka od nauczyciela	1 / 0
fully_funded	Czy projekt został w pełni sfinansowany	1 / 0
at_least_1_green_donation	Czy była przynajmniej jedna "zielona" darowizna, czyli taka która została zrealizowana przez kartę kredytową, PayPal, Amazon lub czek	1 / 0
great_chat	Projekt ma stronę z komentarzami	1 / 0
three_or_more_non_teacher_referred_donors	non-teacher to osoba która trafiła na stronę nie poprzez link od nauczyciela	1 / 0
one_non_teacher_referred_donor_giving_100_plus	czy osoba która trafiła na stronę nie poprzez link od nauczyciela wpłaciła więcej niż 100$	1 / 0
donation_from_thoughtful_donor	Czy pojawiła się darowizna od specjalnego darczyńcy, czyli takiego ze specjalnej listy	1 / 0

Cechy ciągłe

Nazwa cechy	Opis	Zakres
great_messages_proportion	Proporcja pozytywnych wiadomości (do niepozytywnych)	0-1
teacher_referred_count	Liczba darczyńców poleconych przez nauczycieli (znormalizowana)	0-1
non_teacher_referred_count	Liczba darczyńców poleconych (ale nie przez nauczycieli) (znormalizowana)	0-1

Zmienna docelowa

Nazwa	Opis	Wartości
class	Czy zbiórka jest wyjątkowa	0 (nie) / 1 (tak)

Liczebność anomalii

Zbiór danych składa się z 36710 przypadków zbiórek wyjątkowych i 582616 niewyjątkowych. Oznacza to że tylko 5,927% zbiórek jest wyjątkowa (zbiórka przeszła najśmielsze oczekiwania).

Przykładowa interpretacja wiersza

Dla wiersza: 0,1,1,0,1,1,0,0.5,0.02649,0.029605,0

Interpretacja:

- **0**: Brak darczyńców poleconych przez nauczyciela
- **1**: Projekt w pełni sfinansowany
- **1**: Przynajmniej jedna "zielona" darowizna
- **0**: Brak znaczącej komunikacji
- **1**: Trzech lub więcej darczyńców, niepoleconych przez nauczyciela
- **1**: Przynajmniej jedna darowizna ≥$100 od darczyńcy niepoleconego przez nauczyciela
- **0**: Brak darowizny od specjalnego darczyńcy
- **0.5**: 50% wiadomości wysokiej jakości
- **0.02649**: Znormalizowana liczba darczyńców poleconych przez nauczyciela
- **0.029605**: Znormalizowana liczba darczyńców niepoleconych przez nauczyciela
- **0**: Negatywny wynik, nie jest to zbiórka wyjątkowa

Kluczowe notatki

1. **Duża nierównowaga klas**: Tylko 5,927% zbiórek jest wyjątkowa
2. ***Dane binarne i ciągłe**: Oznacza to, że aby użyć wszystkich cech należy użyć metryki odległości do danych mieszanych, taką metryką jest odległość Gowera. Innym rozwiązaniem byłoby użycie tylko danych ciągłych, co bardzo zmniejsza liczbę analizowanych cech.

Zbiór danych "Wykrywanie oszustw kart kredytowych"

Ogólne informacje

Zbiór zawiera transakcje kartami kredytowymi europejskich klientów z września 2013 roku. Dane są silnie niezbalansowane - większość to transakcje prawidłowe (Class=0), a tylko niewielki procent stanowią oszustwa (Class=1).

Charakterystyka zbioru

- **Liczba transakcji**: 284 807
- **Transakcje zaklasyfikowane jako oszustwa**: 492 (0.172%)
- **Cechy**: 30 (Time + 28 komponentów PCA + Amount + Class)
- **Typ danych**: Wszystkie numeryczne (ciągłe)

Opis cech

Nazwa cechy	Opis	Wartości
Time	Liczba sekund od pierwszej transakcji w zbiorze	0-172792
V1-V28	Główne składowe uzyskane metodą PCA (oryginalne cechy zanonimizowane)	0-1

Nazwa cechy	Opis	Wartości
Amount	Kwota transakcji (nieznormalizowana)	0.0 - 25691.16
Class	Zmienna docelowa (0 = prawidłowa, 1 = oszustwo)	1 lub 0

Kluczowe notatki

1. **Duża nierównowaga klas**: Tylko 492 oszustwa na 284 807 transakcji
2. **Zanonimizowane cechy**: Oryginalne dane przekształcone metodą PCA, są już znormalizowane więc nie wymagają obróbki
3. **Bezużyteczność czasu**: Dane pochodzą z 2 dni, kolumna "czas transakcji" wydaje się bezużyteczna w kontekście przewidywania anomalii, zostanie wyrzucona.
4. **Wymóg znormalizowania kwoty transakcji**: Trzeba znormalizować kwotę transakcji do przedziału od 0 do 1