

Projekt 7: Anomalia i uczenie maszynowe

RAPORT 1

Teoria i plan

Zaawansowane zagadnienia algorytmiki i programowania

Rok akademicki 2025/2026

Spis treści

1. Wstęp
2. Przegląd metod detekcji anomalii
 - 2.1. Local Outlier Factor (LOF)
 - 2.2. Isolation Forest
 - 2.3. Principal Component Analysis (PCA)
3. Plan eksperymentów
4. Wybór zbiorów danych
5. Harmonogram realizacji
6. Bibliografia

1. Wstęp

Detekcja anomalii stanowi kluczowy element wielu systemów analitycznych, od wykrywania oszustw finansowych po diagnostykę medyczną i monitorowanie bezpieczeństwa systemów informatycznych. Anomalie, określane także jako wartości odstające (outliers), to obserwacje znacząco odbiegające od normy ustalonej przez większość danych w zbiorze. Ich identyfikacja wymaga zastosowania zaawansowanych technik uczenia maszynowego, które potrafią modelować złożone wzorce w danych wielowymiarowych.

Celem projektu jest przeprowadzenie kompleksowej analizy porównawczej trzech wiodących metod detekcji anomalii: Local Outlier Factor (LOF), Isolation Forest oraz Principal Component Analysis (PCA). Każda z tych metod reprezentuje odmienne podejście do problemu identyfikacji wartości odstających, co pozwoli na wszechstronną ocenę ich skuteczności w różnych scenariuszach aplikacyjnych.

Raport składa się z trzech części:

- Pierwsza przedstawia teoretyczne podstawy wybranych algorytmów, omawiając ich założenia matematyczne i mechanizmy działania.
- Druga część zawiera szczegółowy plan eksperymentów, definiując metryki oceny oraz metodologię testowania.
- Trzecia część opisuje wybrane zbiory danych, które posłużą do empirycznej weryfikacji skuteczności analizowanych metod.

2. Przegląd metod detekcji anomalii

2.1. Local Outlier Factor (LOF)

Local Outlier Factor to algorytm detekcji anomalii oparty na lokalnej gęstości, zaproponowany przez Breunig i współautorów (2000). LOF identyfikuje wartości odstające poprzez porównanie lokalnej gęstości danego punktu z gęstością jego sąsiadów. Kluczową zaletą tej metody jest zdolność wykrywania anomalii lokalnych, które mogą być niewidoczne dla metod globalnych.

Podstawowe koncepcje

Algorytm LOF opiera się na kilku kluczowych koncepcjach. Pierwszą z nich jest k-distance, czyli odległość punktu do jego k-tego najbliższego sąsiada. Następnie definiuje się reachability-distance między dwoma punktami jako maksimum z rzeczywistej odległości między nimi a k-distance drugiego punktu. Ta miara zapewnia stabilność algorytmu w przypadku punktów znajdujących się bardzo blisko siebie.

Lokalna gęstość osiągalności (local reachability density - LRD) punktu jest odwrotnością średniej odległości osiągalności do jego k najbliższych sąsiadów. Wartość ta informuje o tym, jak gęsto rozmieszczone są punkty w lokalnym otoczeniu analizowanego obiektu.

Obliczanie współczynnika LOF

Współczynnik LOF dla punktu p obliczany jest jako średni stosunek lokalnej gęstości osiągalności jego sąsiadów do jego własnej lokalnej gęstości. Wartość LOF bliska 1

oznacza, że punkt znajduje się w regionie o podobnej gęstości jak jego otoczenie. Wartości znacznie większe od 1 wskazują na anomalie - punkt jest położony w regionie o niższej gęstości niż jego sąsiedzi.

Zalety i wady

- **Zalety:** Wykrywa anomalie lokalne, jest odporny na różne gęstości w zbiorze danych
- **Wady:** Wysoka złożoność obliczeniowa $O(n^2)$, wrażliwość na wybór parametru k

2.2. Isolation Forest

Isolation Forest, wprowadzony przez Liu i współautorów (2008), wykorzystuje fundamentalnie odmienne podejście do detekcji anomalii. Zamiast próbować modelować normalne zachowanie, algorytm bezpośrednio izoluje anomalie. Podstawowym założeniem jest to, że anomalie są rzadkie i różnią się od normalnych obserwacji, przez co są łatwiejsze do izolowania.

Mechanizm izolacji

Algorytm buduje las drzew izolacji (isolation trees), gdzie każde drzewo jest konstruowane poprzez rekurencyjne, losowe podziały przestrzeni cech. W każdym węźle losowo wybiera się cechę i wartość podziału między minimum a maksimum tej cechy w aktualnym podzbiorze danych. Proces ten kontynuowany jest do momentu izolacji każdego punktu lub osiągnięcia maksymalnej głębokości drzewa.

Kluczową obserwacją jest to, że anomalie wymagają średnio mniejszej liczby podziałów do izolacji niż punkty normalne. Anomalie znajdują się zwykle na peryferiach rozkładu danych i mogą być oddzielone od reszty stosunkowo szybko. Punkty normalne, będące częścią gęstych regionów, wymagają znacznie więcej podziałów.

Wynik anomalii

Wynik anomalii (anomaly score) dla każdego punktu obliczany jest na podstawie średniej długości ścieżki w wszystkich drzewach lasu. Krótsze ścieżki wskazują na większe prawdopodobieństwo bycia anomalią. Wynik jest normalizowany względem oczekiwanej długości ścieżki w drzewie BST, co pozwala na porównywanie wyników między różnymi zbiorami danych.

Zalety i wady

- **Zalety:** Niska złożoność $O(n \log n)$, skuteczny dla danych wielowymiarowych, nie wymaga założeń o rozkładzie
- **Wady:** Może mieć trudności z anomaliami lokalnymi, wrażliwy na bardzo wysokie wymiary

2.3. Principal Component Analysis (PCA)

Principal Component Analysis, szczegółowo omówiona przez Jolliffe i Cadima (2016), to technika redukcji wymiarowości, która może być skutecznie wykorzystana do detekcji anomalii. PCA transformuje dane do nowej przestrzeni zdefiniowanej przez składowe główne, które są kierunkami maksymalnej wariancji w danych.

Podstawy matematyczne

PCA rozpoczyna od standaryzacji danych i obliczenia macierzy kowariancji. Następnie wyznacza wektory własne i wartości własne tej macierzy. Wektory własne, uporządkowane według malejących wartości własnych, definiują składowe główne. Pierwsza składowa główna odpowiada kierunkowi największej wariancji w danych, druga - największej wariancji prostopadłej do pierwszej, i tak dalej.

Transformacja danych do przestrzeni składowych głównych polega na projekcji oryginalnych punktów na nowe osie zdefiniowane przez wektory własne. Ta transformacja zachowuje strukturę danych, jednocześnie redukując wymiarowość poprzez odrzucenie składowych odpowiadających najmniejszym wartościom własnym.

Detekcja anomalii przy użyciu PCA

W kontekście detekcji anomalii, PCA może być wykorzystana na dwa główne sposoby. Pierwszy polega na analizie błędu rekonstrukcji - anomalie często mają wysoki błąd rekonstrukcji po projekcji na przestrzeń o zredukowanej wymiarowości i odwrotnej transformacji. Drugi sposób wykorzystuje odległości w przestrzeni składowych głównych - anomalie mogą mieć ekstremalne wartości w niektórych składowych.

Szczególnie skuteczna jest metoda oparta na wykresie Q-Q (quantile-quantile) odległości Mahalanobisa w przestrzeni składowych głównych. Punkty znacząco odbiegające od prostej na wykresie Q-Q są potencjalnymi anomaliami. Dodatkowo, można analizować wkład poszczególnych obserwacji do każdej składowej głównej, identyfikując punkty o nietypowo wysokich wkładach.

Zalety i wady

- **Zalety:** Redukcja wymiarowości, interpretowalna transformacja, skuteczna dla liniowych zależności
- **Wady:** Zakłada liniowość, wrażliwa na skalę danych, może przeoczyć nieliniowe anomalie

3. Plan eksperymentów

Plan eksperymentów został zaprojektowany w celu kompleksowej oceny skuteczności trzech analizowanych metod detekcji anomalii. Eksperymenty będą przeprowadzone systematycznie, zgodnie z przedstawioną poniżej metodologią, zapewniając rzetelność i powtarzalność wyników.

3.1. Etapy eksperymentów

1. **Przygotowanie danych:** Standaryzacja, obsługa brakujących wartości, podział na zbiory treningowe i testowe
2. **Implementacja algorytmów:** Kodowanie LOF i PCA od podstaw, wykorzystanie bibliotek dla walidacji
3. **Dostrajanie hiperparametrów:** Grid search dla optymalnych wartości k (LOF), liczby drzew (IF), liczby składowych (PCA)
4. **Ewaluacja:** Obliczenie metryk dla każdej metody na wszystkich zbiorach danych

5. **Analiza wyników:** Porównanie skuteczności, analiza czasu wykonania, wizualizacje

3.2. Metryki oceny

Do oceny skuteczności algorytmów wykorzystane zostaną następujące metryki:

Metryka	Opis
Precision	Stosunek prawdziwych anomalii do wszystkich wykrytych anomalii
Recall	Stosunek wykrytych anomalii do wszystkich rzeczywistych anomalii
F1-score	Średnia harmoniczna precision i recall
AUC-ROC	Pole pod krzywą ROC, miara ogólnej jakości klasyfikatora
Czas wykonania	Czas trenowania i predykcji dla różnych rozmiarów zbiorów

4. Wybór zbiorów danych

Do przeprowadzenia eksperymentów wybrano zróżnicowane zbiory danych, które pozwolą na wszechstronną ocenę analizowanych metod. Każdy zbiór charakteryzuje się odmiennymi właściwościami, co umożliwi zbadanie zachowania algorytmów w różnych kontekstach.

4.1. KDD Cup 99

Klasyczny zbiór danych wykorzystywany w detekcji włamań sieciowych. Zawiera 41 cech opisujących połączenia sieciowe, z których około 20% stanowią anomalie (ataki). Zbiór ten pozwoli ocenić skuteczność metod w kontekście cyberbezpieczeństwa, gdzie szybka i dokładna detekcja zagrożeń jest kluczowa.

4.2. Credit Card Fraud Detection

Zbiór transakcji kartami kredytowymi zawierający 284807 transakcji, z których tylko 0.172% to oszustwa. Ekstremalne niezbilansowanie klas stanowi wyzwanie typowe dla rzeczywistych problemów detekcji anomalii. Dane zostały już przetransformowane przy użyciu PCA ze względów prywatności, co umożliwi interesującą analizę skuteczności PCA na danych już przetransformowanych tą metodą.

4.3. Breast Cancer Wisconsin

Zbiór medyczny zawierający 569 obserwacji z 30 cechami opisującymi charakterystyki komórek nowotworowych. Około 37% przypadków to nowotwory złośliwe, które można traktować jako anomalie. Ten zbiór pozwoli ocenić

przydatność metod w diagnostyce medycznej, gdzie błędna klasyfikacja może mieć poważne konsekwencje.

4.4. Synthetic Dataset

Dodatkowo wygenerowany zostanie syntetyczny zbiór danych z kontrolowanymi anomaliami. Pozwoli to na precyzyjną ocenę zdolności każdej metody do wykrywania różnych typów anomalii: punktowych, kontekstowych i zbiorowych. Zbiór będzie zawierał 10000 obserwacji w przestrzeni 10-wymiarowej z 5% anomalii o znanej lokalizacji.

5. Harmonogram realizacji

Etap	Zakres prac	Termin
Raport 1	Przegląd literatury, teoria algorytmów, plan projektu	Tydzień 1-2
Raport 2	Implementacja LOF i PCA, testy podstawowe	Tydzień 3-5
Raport 3	Implementacja Isolation Forest, optymalizacje, autoenkoder	Tydzień 6-9
Raport 4	Analiza porównawcza, wizualizacje, wnioski końcowe	Tydzień 10-12

6. Bibliografia

1. Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). Springer.
2. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104.
<https://doi.org/10.1145/342009.335388>
3. Choraś, M., & Kozik, R. (2020). Artificial Intelligence Tools for Cyber Threat Detection and Mitigation. Springer.
4. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Wprowadzenie do algorytmów. WNT.
5. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065).
<https://doi.org/10.1098/rsta.2015.0202>
6. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining (ICDM), 413–422.
<https://doi.org/10.1109/ICDM.2008.17>