Leszek **Albrzykowski**

UCZENIE MASZYNOWE

Elementy matematyki w analizie danych

Helion

Redaktor prowadzący: Szymon Szwajger
Recenzja naukowa: dr Zbigniew Leśniak, Instytut Matematyki Uniwersytetu Pedagogicznego w Krakowie
Projekt okładki: Studio Gravite / Olsztyn Obarek, Pokoński, Pazdrijowski, Zaprucki

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/uczmae
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

ISBN: 978-83-283-9139-0

Printed in Poland.

Spis treści

Algorytm Isolation Forest

Do detekcji anomalii możemy wykorzystać algorytm zaproponowany w 2007 roku[3], bazujący na drzewach decyzyjnych. Podstawą działania drzew decyzyjnych jest sekwencyjny podział zbioru, tak aby minimalizować nieuporządkowanie obserwacji.

Przykład

Załóżmy, że dysponujemy historią spłat kredytu przez klientów banku:

KLIENT	WIEK	STAŁE DOCHODY?	SPŁACONY KREDYT?
C_1	23	Nie	Nie
C_2	35	Nie	Tak
C_3	21	Tak	Tak
C_4	42	Tak	Tak

Na podstawie powyższych danych drzewo decyzyjne mogłoby skonstruować reguły oceny zdolności kredytowej:

REGUŁA	WARUNEK
R_1	$WHEN\ (Client\ Age < 35\ \wedge\ Regular\ Income = False)\ THEN\ Credit\ Score = Low$
R_2	$WHEN\ (Client\ Age \geq 35\ \wedge\ Regular\ Income = False)\ THEN\ Credit\ Score = High$
R_3	$WHEN\ (Client\ Age < 35\ \wedge\ Regular\ Income = True)\ THEN\ Credit\ Score = High$
R_4	$WHEN\ (Client\ Age \geq 35\ \wedge\ Regular\ Income = True)\ THEN\ Credit\ Score = High$

[3] Liu F. T. i in., *Isolation Forest, https://www.researchgate.net/profile/Fei-Tony-Liu/publication/224384174_Isolation_Forest/links/5bbd5c0ca6fdcc9552dd04f0/Isolation-Forest.pdf* [dostęp: 26.07.2022].

Zauważmy, że w powyższym przykładzie na podstawie żadnej z cech klienta (wiek, stały dochód) nie można by prawidłowo sklasyfikować klienta, gdyby cechy były brane pod uwagę indywidualnie. Dopiero połączenie informacji na podstawie dwóch cech pozwoliło osiągnąć cel.

W przypadku kiedy model (klasyfikator) nie jest złożony z pojedynczego drzewa decyzyjnego, lecz z wielu drzew, mówimy o lasach (ang. *forest*).

Podstawą działania algorytmu Isolation Forest jest założenie, że anomalie to takie obserwacje, które w czasie podziału zbioru D obserwacji są najłatwiejsze do odizolowania od pozostałych. Innymi słowy, im prostsza reguła pozwalająca na odizolowanie danej obserwacji od pozostałych, tym bardziej prawdopodobne, że ta obserwacja jest anomalią.

W algorytmie Isolation Forest węzły T drzew nazywamy zewnętrznymi (liśćmi), jeżeli nie posiadają dzieci (rozgałęzień), lub wewnętrznymi, jeżeli posiadają dzieci. Zadaniem węzłów wewnętrznych jest rozdzielenie zbioru na podstawie reguł, przy czym w opisywanym algorytmie reguły te mają postać $q < p$, gdzie:

- q jest cechą obserwacji,
- p jest wartością cechy q spełniającą warunek:

$$min(q) < p < max(q).$$

Wybór cechy q i wartości p jest losowy. W przypadku stosowania algorytmu drzew losowych wybór cech i punktu podziału określany jest często na podstawie miar takich jak Entropia Shannona lub gini impurity których szczegóły przedstawimy w dalszej części rozdziału.

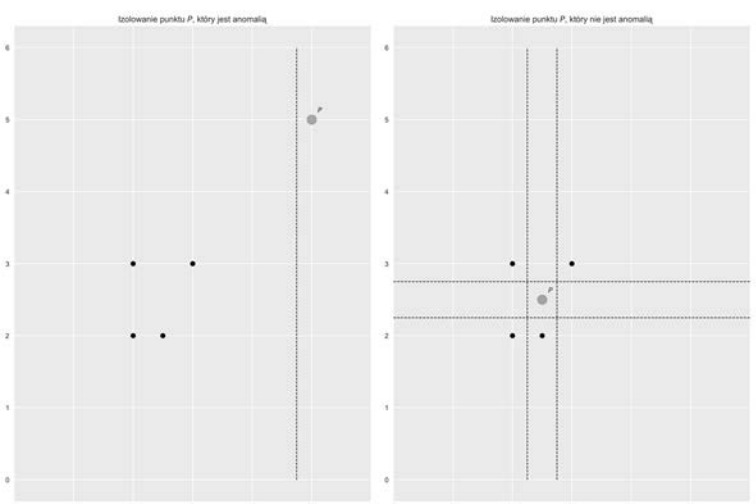
Warunek $q < p$ prowadzi do binarnego podziału węzła T na dwa kolejne węzły:

- $T_{\text{left}} = \{d \in D : q < p\}$,
- $T_{\text{right}} = \{d \in D : q \geq p\}$.

Węzły T_{left} i T_{right} mogą być dalej rozdzielane.

Algorytm kończy działanie po odizolowaniu wszystkich punktów lub jeśli głębokość drzewa (liczba podziałów) osiągnie wskazany *explicite* próg.

Oceniając, czy dana obserwacja jest anomalią, bierzemy pod uwagę średnią liczbę podziałów (dokonanych osobno przez każde z drzew klasyfikatora) konieczną do odizolowania obserwacji od pozostałych.



RYSUNEK 4.2. Punkty będące anomaliami jest łatwiej odizolować od pozostałych, gdy dzieli się obserwacje na podstawie ich cech

KOMENTARZ

Dwiema powszechnie używanymi miarami służącymi do rozdziału zbiorów w drzewach decyzyjnych są entropia i *gini impurity*.

Do najpowszechniejszych miar nieuporządkowania zbioru w teorii informacji i analizie danych należy entropia Shannona[4].

Za pomysłem Shannona stoi problem określenia ilości informacji w przesyłanym komunikacie. Gdyby ktoś poinformował nas, że wydarzy się coś, co jest mało prawdopodobne, zapewne uznalibyśmy taki komunikat za znacznie bardziej wartościowy, niż gdyby ktoś poinformował nas o czymś bardzo prawdopodobnym (oczywistym).

Entropię Shannona definiujemy jako:

$$H(T) = -\sum_{i=1}^{n} P(C_i) log_2\big(P(C_i)\big).$$

Drugą z powszechnie używanych miar jest *gini*[5] *impurity*, która wskazuje nam na prawdopodobieństwo, z jakim losowo wybrana obserwacja zostanie błędnie sklasyfikowana:

$$gini\ impurity\ (I_G) = 1 - \sum_{i=1}^{n} P^2(c_i),$$

gdzie $P(c_i)$ jest prawdopodobieństwem przynależności elementów do danej klasy.

Porównajmy wartości obydwu miar. Wartość entropii dla przykładowego zbioru:

[4] Claude Elwood Shannon (ur. 1916 w Gaylord w Michigan, zm. 2001 w Medford w Massachusetts) — amerykański matematyk i inżynier.

[5] Corrado Gini (ur. 1884 w Motta di Livenza, zm. 1965 w Rzymie) — włoski statystyk i demograf.

ZBIÓR	ELEMENTY	ENTROPIA
S_1	O_{C1}, O_{C1}, O_{C1}, O_{C1}	$H(T) = -[P(C_1) \cdot log_2(P(C_1))] = -[1 \cdot log_2(1)] = 0$
S_2	O_{C1}, O_{C1}, O_{C1}, O_{C2}	$H(T) = -[P(C_1) \cdot log_2(P(C_1)) + p(C_2) \cdot log_2(P(C_2))] = -[0{,}75 \cdot log_2(0{,}75) + 0{,}25 \cdot log_2(0.25)] \approx 0{,}81$
S_3	O_{C1}, O_{C1}, O_{C2}, O_{C2}	$H(T) = -[P(C_1) \cdot log_2(P(C_1)) + P(C_2) \cdot log_2(P(C_2))] = -[0{,}5 \cdot log_2(0{,}5) + 0{,}5 \cdot log_2(0{,}5)] = 1$

oraz *gini inpurity*:

ZBIÓR	ELEMENTY	*GINI IMPURITY*
S_1	O_{C1}, O_{C1}, O_{C1}, O_{C1}	$I_G = 1 - P^2(C_1) = 1 - 1^2 = 0$
S_2	O_{C1}, O_{C1}, O_{C1}, O_{C2}	$I_G = 1 - (P^2(C_1) + P^2(C_2)) = 1 - (0{,}75^2 + 0{,}25^2) = 0{,}375$
S_3	O_{C1}, O_{C1}, O_{C2}, O_{C2}	$I_G = 1 - (P^2(C_1) + P^2(C_2)) = 1 - (0{,}5^2 + 0{,}5^2) = 0{,}5$

Zauważmy, że dla klasyfikacji binarnej obydwie miary przyjmują wartości:

- 0, jeżeli zbiór zawiera elementy należące do jednej klasy (S_1), przy czym 0 jest równocześnie wartością minimalną obydwu miar;
- maksymalne (odpowiednio, 1 i 0,5, dla entropii i *gini inpurity*), jeżeli zbiór zawiera po tyle samo elementów należących do jednej i drugiej klasy.

Istotne różnice pomiędzy miarami to:

- zakres wartości przyjmowanych przez miary;
- łatwiejsze (szybsze) obliczenie wartości dla miary *gini inpurity* w porównaniu z entropią wynikające z braku konieczności obliczania wartości logarytmu w przypadku tej pierwszej.


PROGRAM
PARTNERSKI
GRUPY HELION
1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ
Zmień swoją stronę WWW w działający bankomat!
Dowiedz się więcej i dołącz już dzisiaj!
http://program-partnerski.helion.pl
GRUPA
Helion

Prosto o sprawach skomplikowanych

Większość książek poświęconych uczeniu maszynowemu, w szczególności analizie danych, skupia się na zastosowaniu modeli matematycznych, a pomija omówienie podstaw. Albo – dla odmiany – zgłębiające tę dziedzinę prace mają postać podręczników akademickich, opasłych tomów wymagających od odbiorców znajomości wyższej matematyki, a także sporego nakładu pracy i czasu.

Ta książka została pomyślana tak, by wypełnić lukę między opracowaniami opartymi na podejściu stricte praktycznym a tymi mocno teoretyzującymi. Autor ma nadzieję, że czytelnicy odnajdą radość w zagłębianiu się w kolejne zagadnienia matematyczne i dostrzegą, w jak znacznym stopniu ich znajomość ułatwia rozwiązywanie problemów dotyczących analizowania danych.

Najważniejsze zagadnienia:

- **Liniowe zależności danych**
- **Wnioskowanie bayesowskie**
- **Czynniki wpływające na wyniki modelu**
- **Detekcja anomalii**
- **Ewaluacja modeli**

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej! ▶

ISBN 978-83-283-9139-0

9 788328 391390

Cena: 39,90 zł