Zastosowanie algorytmów genetycznych do wykrywania wartości odstających

Kelly D. Crawford
Amoco Corporation
kcrawford@amoco.com

Roger L. Wainwright
Uniwersytet w Tulsie
rogerw@penguin.utulsa.edu

Streszczenie

Naukowcy są przyzwyczajeni do niedokładności pomiarów fizycznych i opracowali metody statystyczne, które pomagają radzić sobie z błędami. Wykrywanie odstających obserwacji w danych regresji (wartości odstające) jest ważnym krokiem w analizie tych zestawów danych. Niniejszy artykuł przedstawia algorytm genetyczny zdolny do generowania podzbiorów do diagnostyki wielu przypadków odstających. Algorytm genetyczny wykorzystuje diagnostykę jako funkcje oceny do wyszukiwania dobrych podzbiorów. Testy przeprowadzone na różnych zestawach danych z literatury statystycznej pokazują doskonałą wydajność algorytmu genetycznego w lokalizowaniu najlepszych podzbiorów. Późniejsza analiza tych podzbiorów dokładnie określa, które zbiory są faktycznie odstające. Algorytm genetyczny doskonale radzi sobie z wykorzystaniem diagnostyki wielu przypadków do wyszukiwania tych wartości odstających.

1 WPROWADZENIE

Dane rzadko są doskonałe. Niezależnie od tego, czy problemem jest wadliwy sprzęt, wahania spowodowane wiatrem, czy zwykły błąd ludzki, obserwacje, które rejestrujemy, często zawierają błędy, które mogą prowadzić do mylących wniosków. Błędy te, zwane odstającymi, muszą zostać znalezione i usunięte z danych.

Różne techniki regresji pomagają zlokalizować dane odstające, minimalizując ich wpływ lub lokalizując je bezpośrednio. Klasa technik zwana diagnostyką wielokrotnych przypadków odstających zapewnia sposób pomiaru "odstających" zestawów punktów w stosunku do całego zestawu danych. Kluczowym zadaniem jest wybór odpowiednich podzbiorów do testowania. Kombinatoryczny charakter tego problemu wymaga heurystycznego rozwiązania.

W sekcji 2 omówiono sposoby radzenia sobie z problemem wartości odstających w statystyce. W sekcji 3 przedstawiono algorytm genetyczny, który generuje podzbiory potencjalnie odstających wartości.

w punktach danych regresji. Testy na rzeczywistych danych przedstawiono w sekcji 4, a ich wyniki w sekcji 5.

2 ZEWNĘTRZNE

2.1 NAJMNIEJ KWADRATÓW

Najmniejsza suma kwadratów, częściej nazywana najmniejszymi kwadratami (LS), to technika regresji używana do wyznaczania linii (lub hiperpłaszczyzny) t do zestawu danych. Regresja najmniejszych kwadratów konstruuje linię przechodzącą przez dane w taki sposób, że suma kwadratów reszt (odległości od każdego punktu do linii w kierunku y) jest zminimalizowana. Dostępnych jest wiele dobrych źródeł dotyczących techniki najmniejszych kwadratów, takich jak (Cheney 1980, Hager 1988).

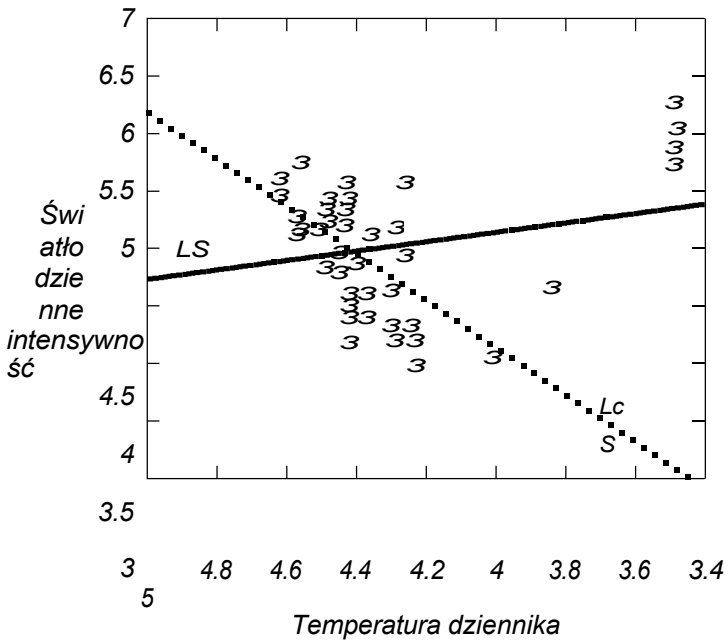
Punkty danych są reprezentowane przez układ nadokreślony y = X + , gdzie n jest liczbą punktów danych, y jest wektorem n zmiennych zależnych, X jest macierzą n p 1 zmiennych niezależnych p 1, jest wektorem p 1 współczynników i jest wektorem n niezależnych błędów. Przyjmując, że y^ = y, rozwiązaniem układu metodą najmniejszych kwadratów jest X^t X = X^t y^. Jest to równoważne minimalizacji błędu resztowego w następujący sposób:

$$LS = \underset{i=1}{\overset{n}{minimalizuj}} \Lambda_{2}$$

Jako przykład rozważmy diagram Hertzsprunga-Russella gromady gwiazd CYG OB1, często analizowany zbiór danych z dziedziny astronomii (Rousseeuw i Leroy 1987). W celach referencyjnych będziemy nazywać go zbiorem danych CYG. Dane, przedstawione na rysunku 1, pochodzą z 47 gwiazd w kierunku Cygnus.

Oś x (pokazana numerycznie w odwrotnej kolejności) reprezentuje logarytm efektywnej temperatury na powierzchni każdej gwiazdy, podczas gdy oś y pokazuje logarytm natężenia światła gwiazdy. Na wykresie znajdują się dwie linie. Linia oznaczona jako LS została utworzona na podstawie testu najmniejszych kwadratów dla całego zbioru danych.

Zwróć uwagę na 4 punkty w prawym górnym rogu ekranu



Rysunek 1: Diagram Hertzsprunga-Russella dla gromady gwiazd CYG OB1

wykres. Odpowiadają one gwiazdom olbrzymom, których pomiary nie korelują liniowo z pomiarami innych gwiazd. Punkty te są uważane za odstające. Przeniesienie tych punktów i ponowne obliczenie linii najmniejszych kwadratów daje linię oznaczoną L^S. Na tym przykładzie można zauważyć, że oryginalna linia najmniejszych kwadratów została dramatycznie zniekształcona przez punkty odstające i nie pomogła w znalezieniu punktów odstających.

2.2 SOLIDNE TECHNIKI

Techniki odporne są tak nazywane, ponieważ próbują zminimalizować wpływ wartości odstających. Jednym z najlepszych przykładów jest najmniejsza mediana kwadratów (LMS) (Rousseeuw 1984). W tej technice minimalizowana jest mediana kwadratów reszt, a nie ich suma. Działa to dobrze, ponieważ dla każdej rozważanej linii używany jest tylko punkt odpowiadający medianie posortowanych reszt kwadratowych. Wynikiem jest

wyrzucając wysokie i niskie wartości rezydualne, które mogą mieć tendencję do wyrzucania najmniejszych kwadratów ze ścieżki.

Dostępnych jest wiele innych odpornych estymatorów, takich jak estymator najmniejszych wartości bezwzględnych, najmniejsze przycięte kwadraty, ważone i ponownie ważone najmniejsze kwadraty, estymatory M i estymatory S. Rousseeuw i Leroy (1987) przedstawiają doskonałą historię rozwoju tych technik, wraz z omówieniem różnych mocnych i słabych stron każdej z nich.

diagnostyka wartości odstających próbuje ocenić wpływ każdego punktu w zbiorze danych. Diagnostyka pojedynczych przypadków koncentruje się na wpływie poszczególnych punktów, podczas gdy diagnostyka wielu przypadków działa z zestawami punktów.

Technikę najmniejszych kwadratów, opisaną w sekcji 2.1, można traktować jako prostą, wielokrotną technikę diagnostyki wartości odstających. Zatem LS(I) to najmniejsze kwadraty zbioru danych z usuniętym podzbiorem punktów I (Crawford, Wainwright i Vasicek 1995). Będziemy odnosić się do tej miary jako LS.

Kwadratowy wzór Cooka na odległość dla diagnostyków w wielu przypadkach (Cook i Weisberg 1982) jest uogólnieniem wersji dla pojedynczego przypadku (Cook 1977). Szczegóły tego wzoru pozostawiono w źródłach, ale w skrócie wzór ten jest następujący

$$CD^2(I) = \frac{(\hat{} \hat{}(I)) X^{tt} X(\hat{} \hat{}(I))}{ps2}$$

gdzie ^ jest równaniem linii utworzonej metodą najmniejszych kwadratów, a ^(I) jest linią najmniejszych kwadratów utworzoną po usunięciu podzbioru punktów I ze zbioru danych. X jest macierzą nadokreśloną skonstruowaną ze zbioru danych, p jest efektywnie wymiarem zbioru danych, a s^2 jest sumą kwadratów reszt podzieloną przez $n-p$

odwrotność stopni swobody ([1]). Będziemy odnosić

2.3 DIAGNOSTYKA WARTOŚCI ODSTAJĄCYCH

Niniejszy artykuł koncentruje się na klasie technik zwanych diagnostyką wartości odstających. Zamiast minimalizować ef-

się do
ten środek jako CD.

Andrews i Pregibon (1978) opracowali diagnostykę wielu przypadków odstających opartą na następującym współczynniku determinacji

$$AP(I) = \frac{\det(Z^t(I)Z(I))}{\det(Z^t Z)}$$

gdzie Z to macierz X z dołączoną zmienną odpowiedzi y, a $Z(I)$ to Z z usuniętym podzbiorem punktów I. Geometrycznie, $AP(I)$ jest miarą "oddalenia" podzbioru punktów I. Miarę tę będziemy określać jako AP.

Odchylenie wybranych podzbiorów punktów ze zbioru danych można oszacować za pomocą LS, CD i AP. Kluczową częścią jest ustalenie, które podzbiory należy ocenić. Wszystkie podzbiory punktów będą musiały zostać sprawdzone, chyba że istnieje mechanizm wyboru odpowiednich podzbiorów. Jest to fundamentalny problem kombinatoryczny wspólny dla wszystkich tego typu systemów diagnozowania wielokrotnych przypadków odstających (Rousseeuw i Leroy 1987). Rozważmy zbiór danych o rozmiarze n. Sprawdzenie wszystkich zestawów $1; 2; ; b^n c$ wymaga obliczeń rzędu

$$\sum_{i=1}^{b^n c} \binom{n}{i} :$$

3 ALGORYTM GENETYCZNY DO GENEROWANIA PODZBIORÓW

Opracowaliśmy algorytm genetyczny (GA) do zarządzania wyborem podzbiorów punktów. Biorąc pod uwagę zbiór danych zawierający n punktów, musimy być w stanie reprezentować podzbiór k punktów jako potencjalny zestaw odstający. Aby to osiągnąć, używamy unikalnej listy indeksów k punktów. Logicznie rzecz biorąc, można to uznać za chromosom oparty na kolejności, w którym najbardziej wysunięte na lewo indeksy k punktów są wartościami odstającymi, a pozostałe $n\ k$ wartości nie są odstające. Jednak w praktyce w chromosomie przechowujemy tylko k odstających indeksów punktów. Skutkuje to znacznie szybszym wykonaniem dla większych zbiorów danych.

Zwrotnica zastosowana w naszym GA jest wariantem dwurodzicielskiej jednolitej zwrotnicy opartej na kolejności (UOX) (Davis 1991). Na każdej pozycji w łańcuchu wybierana jest losowo wartość z jednego z dwóch rodziców w celu utworzenia nowych dzieci. Zduplikowane wartości są usuwane poprzez wybranie nowej unikalnej wartości losowej. W przypadku mutacji losowo wybieramy punkt w chromosomie i zamieniamy go na unikalną wartość losową. Ponadto sortujemy indeksy punktów w chromosomie w celu poprawy wydajności.

Porównaliśmy trzy różne funkcje oceny. Każda z nich opiera się na formułach diagnostyki wielu przypadków odstających LS, CD i AP, opisanych wcześniej w sekcjach 2.1 i 2.3.

testy dla $k = 2$. Następnie przeprowadziliśmy testy dla $k = 3$, osobno i tak dalej, aż do $k = b^n c$.

Będziemy odnosić się do zbiorów danych jako belgium, brain, china, cyg (opisanych w sekcji 2) i siegel. Belgium przedstawia liczbę połączeń międzynarodowych z Belgii w latach 1950-1973. Zbiór danych dotyczących mózgu porównuje masę ciała z masą mózgu 18 różnych gatunków zwierząt. Zbiór danych china jest zapisem rocznych stóp wzrostu średnich cen w głównych wolnych miastach wolnych Chin w latach 1940-1948. Zbiór danych siegel jest dokładnym zbiorem danych. Innymi słowy, po usunięciu wartości odstających wszystkie pozostałe punkty będą dokładnie leżeć na linii prostej. Tabela 1 zawiera szczegółowe informacje na temat rozmiaru i wymiaru każdego zbioru danych, a także liczby znanych wartości odstających.

Tabela 1: Dane testowe GA

ZESTAW DANYCH		LICZBA	LICZBA
NAZWA	PUNKTY		POZOSTAŁE
Belgia	24		6
mózg	28		3
Chiny	9		2
cyg	47		4
siegel	9		3

4 DANE TESTOWE

Przetestowaliśmy 5 różnych zestawów danych z wersjami LS, CD i AP algorytmu genetycznego. Najpierw uruchomiliśmy

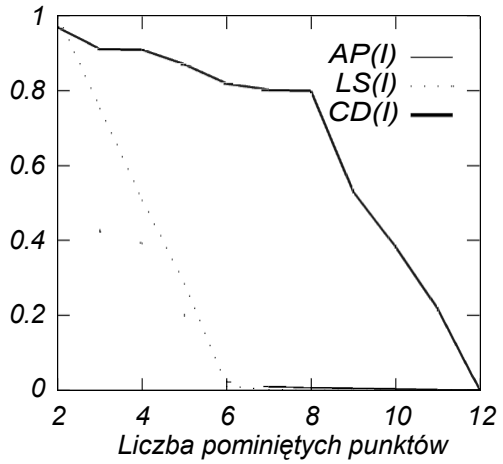
Dla każdego zbioru danych celem jest usunięcie wszelkich odstających punktów danych i znalezienie liniowego t dla danych. Każdy z tych zestawów danych został dokładnie przeanalizowany w literaturze statystycznej i każdy z nich ma znane punkty odstające. Pełny opis każdego zestawu danych można znaleźć w Rousseeuw i Leroy (1987).

Do skonstruowania naszego algorytmu genetycznego wykorzystaliśmy pakiet LibGA (Corcoran i Wainwright 1993). Każdy zestaw danych został przetestowany przy użyciu 100 różnych losowych nasion. Parametry GA to: reprezentacja = liczby całkowite, długość łańcucha = k (oddzielne przebiegi o stałej długości dla k = 2; :::; b^n c), rozmiar populacji = 100, błąd selekcji = 1,8, współczynnik mutacji = 0,05 i maksymalna liczba iteracji = 10000.

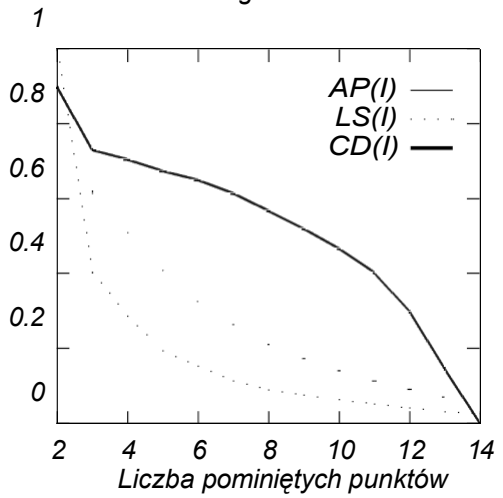
5 WYNIKI

Tabela 2 pokazuje wydajność każdej metody na testowych zbiorach danych w znajdowaniu pojedynczego podzbioru zawierającego rzeczywiste wartości odstające. "% znalezionych wartości odstających" odnosi się do tego, ile ze 100 przebiegów testowych faktycznie znalazło najlepszy podzbiór, podczas gdy "średnia liczba wartości odstających" wskazuje, ile podzbiorów zostało ocenionych przed znalezieniem najlepszego.

Ogólnie rzecz biorąc, CD osiągnął najlepsze wyniki, z wyjątkiem zestawu danych Siegel. W tym zbiorze danych CD uszeregował podzbiór z dwoma odstającymi punktami i jednym nieodstającym punktem wyżej niż tness poprawnych trzech punktów. Tak więc, podczas gdy GA działał poprawnie, CD doprowadził nas do nieprawidłowej odpowiedzi (w rzeczywistości GA znalazł ten zestaw danych).



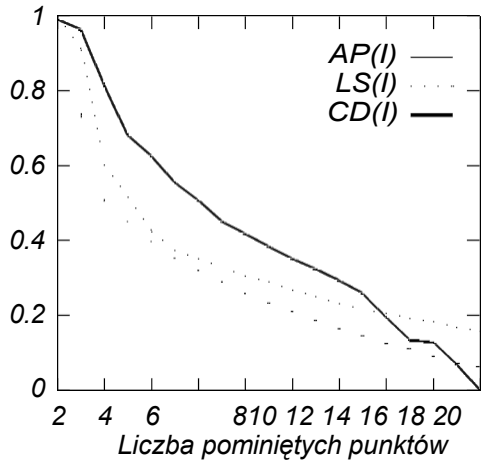
Rysunek 2: Zestaw danych z Belgii



Rysunek 3: Zestaw danych mózgu

dla wszystkich 100 przebiegów). Najbardziej prawdopodobnym wyjaśnieniem jest to, że CD nie obsługuje poprawnie dokładnych zestawów danych.

Znormalizowane wykresy wyników GA są używane do zlokalizowania najbardziej prawdopodobnego zbioru odstającego. GA lokalizuje najlepszy podzbiór o rozmiarze k (nazwij go I_k), dla k = 2; :::; b^n c. Znormalizowane wartości LS(I_k), CD(I_k) i AP (I_k) są wykreślone jako trzy linie na Rysunku 2, Rysunku 3 i Rysunku 4 odpowiednio dla zbiorów danych belgium, brain i cyg. Stosunkowo małe wartości na osi y wskazują na lepszą wartość t dla danych. Idealny t, w którym wszystkie punkty w zestawie znajdują się dokładnie na jednej linii (lub hiperpłaszczyźnie), miałby wartość y równą 0.



Rysunek 4: Zbiór danych CYG

reagują bezpośrednio z rzeczywistymi wartościami odstającymi. CD okazuje się rozczarowujące dla celów analizy końcowej.

6 WNIOSKI

Nasze wyniki wskazują, że algorytmy genetyczne zapewniają użyteczne środki do generowania podzbiorów do diagnostyki wielu przypadków odstających. Zdolność algorytmów genetycznych do pokonywania kombinatorycznych przestrzeni poszukiwań czyni je głównym kandydatem do tego rodzaju badań. Autorzy uważają, że diagnostyka wielokrotnych przypadków odstających nie jest szeroko stosowana na dużych zbiorach danych z powodu tego ograniczenia. Nasz algorytm genetyczny sprawia, że taka analiza jest wykonalna.

Każda diagnostyka wielu przypadków ma mocne i słabe strony. Różne modele danych i zbiory danych generują różne "krajobrazy". Podobnie jak w przypadku solidnych regresji
Stosunkowo duża zmiana nachylenia krzywych przedstawionych na Rysunku 2, Rysunku 3 i Rysunku 4 wskazuje na prawdopodobny zestaw wartości odstających. Zwróć uwagę na względną zmianę nachylenia linii LS i AP po pominięciu 6 punktów dla Belgii (rysunek 2), 3 punktów dla mózgu (rysunek 3) i 4 punktów dla cyg (rysunek 4). Te korelacje

i diagnostyki wartości odstających, nie ma uniwersalnego rozwiązania.
mula do wykrywania wartości odstających. Jednak GA doskonale sprawdził się w wykorzystaniu diagnostyki wielu przypadków do wyszukiwania wartości odstających. Dalsze badania nad innymi technikami diagnostycznymi wielokrotnych przypadków odstających z pewnością p r z y n i o s ą dodatkowy wgląd w ten intrygujący problem.

Podziękowania

Badania te były częściowo wspierane przez OCAST Grant AR2-004. Autorzy pragną również podziękować za wsparcie Sun Microsystems, Inc.

Tabela 2: Wyniki GA

DATASET NAZWA	OCENA FUNKCJA	% CZASU OUTLIERS FOUND	SREDNIA # ITERS (100 RUNS)
Belgia	LS	96	769
	CD	91	1526
	AP	81	4846
mózg	LS	100	245
	CD	100	217
	AP	100	391
Chiny	LS	100	217
	CD	100	110
	AP	100	100
cyg	LS	87	2925
	CD	100	323
	AP	100	759
siegel	LS	100	261
	CD	0	195
	AP	96	2438

Referencje

D. F. Andrews i D. Pregibon (1978). Znajdowanie wartości odstających, które mają znaczenie. Journal of the Royal Statistical Society B, 40:85{93.

W. Cheney i D. Kincaid (1980). Matematyka numeryczna i obliczenia. Brooks/Cole, Monterey, Kalifornia.

R. D. Cook (1977). Wykrywanie obserwacji in uential w regresji liniowej. Technometrics, 19:15{18.

R. D. Cook i S. Weisberg (1982). Residuals and In uence in Regression. Chapman and Hall, London.

A. L. Corcoran i R. L. Wainwright (1993). LibGA: Przyjazne dla użytkownika środowisko pracy do badań nad algorytmami genetycznymi opartymi na kolejności. In Proceedings of the 1993 ACM/SIGAPP Symposium on Applied Computing, strony 111{117.

K. D. Crawford, R. L. Wainwright i D. J. Vasicek (1995). Wykrywanie wielokrotnych wartości odstających w danych regresji przy użyciu algorytmów genetycznych. In Proceedings of the 1995 ACM/SIGAPP Symposium on Applied Computing.

Lawrence Davis (1991). Podręcznik algorytmów genetycznych. Van Nostrand Reinhold, New York, NY.

W. W. Hager (1988). Applied Numerical Linear Algebra. Prentice Hall, Englewood Cli s, New Jersey.

P. J. Rousseeuw (1984). Least median of squares re- gression. Journal of the American Statistical Associa- tion, 79:871-880.

P. J. Rousseeuw i A. M. Leroy (1987). Robust Regression & Outlier Detection. John Wiley & Sons, New York, NY.