MOW Projekt 13

Porównanie algorytmów grupowania i klasyfikacji do detekcji anomalii.

Katarzyna Piórkowska, 259078
Tomasz Korzeniowski, 265753

6 kwietnia 2018

1 Zadanie

1.1 Treść

Nienadzorowana detekcja anomalii za pomocą odpowiednio opakowanych wybranych algorytmów grupowania dostępnych w R. Porównanie z nadzorowaną detekcją anomalii za pomocą dostępnych w R algorytmów klasyfikacji.

1.2 Interpretacja

W ramach projektu należy zbadać czy możliwe jest wykrycie anomalii w danych korzystając ze standardowych algorytmów grupowania dostępnych w R. W tym celu należy zapewnić mechanizm generowania modelu grupowania, który zwróci podział danych trenujących na grupy. Następnie, dla każdego nowego przykładu testowego, trzeba ocenić w jakim stopniu jest on podobny do którejś z wyznaczonych grup. Ocena polegać będzie na wyznaczeniu wskaźnika nietypowości, który jest miarą liczbową wskazującą na ile badany przykład jest niepodobny do rozważanych grup.

Do symulowania anomalii w danych przyjmiemy, że jedna z klas w nich występująca (np. najmniej liczna) będzie stanowiła anomalie. Przykłady należące do tej klasy nie zostaną wykorzystane do budowy modelu grupowania.

Częścią implementacji projektu będzie zapewnienie opakowania wybranych algorytmów grupowania w funkcję, która zwraca model pogrupowanych danych trenujących. Użytkownik będzie miał możliwość podania metody grupowania z jakiej chce skorzystać (wraz z jej niezbędnymi parametrami). Inna funkcja będzie miała za zadanie skorzystać z wyznaczonego modelu oraz dopasować dane testowe przez wyznaczenie zadanego wskaźnika nietypowości.

W części analitycznej zostanie przeprowadzona symulacja wykrywania anomalii na kilku zbiorach danych. Wyniki otrzymane przy użyciu zaimplementowanych funkcji zostaną porównane z wynikami uzyskanymi przez zastosowanie znanych algorytmów klasyfikacji dostępnych w R. W tym drugim przypadku algorytmy będą znały klasy do jakich należą obserwacje by wskazać anomalie jako jedną z klas. Porównanie wyników obu podejść będzie polegało na wyznaczeniu wskaźników jakości (dokładności) powyższych rozwiązań.

2 Algorytmy

Do realizacji zadania zostaną wykorzystane trzy algorytmy grupowania (k-średnich, k-medoidów, grupowania hierarchicznego) oraz dwie metody klasyfikacji (drzewa decyzyjne, k najbliższych sąsiadów). Każdy z nich zostanie pokrótce opisany wraz ze wskazaniem kluczowych parametrów.

2.1 Algorytm k-średnich

Algorytm k-średnich jest jednym z najprostszych algorytmów rozwiązujących zadanie grupowania. Ideą algorytmu jest przyporządkowanie pewnego zbioru N przykładów do przyjętej a priori liczby grup K. Każda grupa posiada dokładnie jeden centroid, czyli punkt reprezentujący wartość średnią grupy. Pojedynczą obserwację $x_i = (x_1, x_2, \ldots, x_N)$ można przyporządkować tylko do jednego z centroidów $c_j = (c_1, c_2, \ldots, c_K)$ Oznacza to minimalizację funkcji:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \|x_i - c_j\|^2 \tag{1}$$

Po zakończeniu pojedynczej iteracji grupowania należy uaktualnić położenie centroidów i przyporządkować obserwacje ponownie. Przebieg grupowania przedstawia algorytm 1.

Algorytm 1 k-średnich

1. Wyznacz początkowe położenie centroidów
2. Przyporządkuj każdej obserwacji najbliższy jej centroid.
3. Gdy wszystkie obserwacje zostaną przyporządkowane, wyznacz ponownie położenie centroidów, znajdując wartość średnią obserwacji przypisanych do centroidu:

$$c_{ji} = \frac{1}{M} \sum_{m=1}^{M} x_m \quad , gdzie\ M - liczba\ obserwacji\ w\ c_j$$

4. Powtarzaj kroki 2. i 3., dopóki centroidy zmieniają swoje położenie lub nie zostanie osiągnięta maksymalna liczba iteracji.

Do zalet algorytmu należy proste znajdowanie podziału grup dobrze odseparowanych od siebie. Największą wadą algorytmu jest konieczność podania liczby grup na jakie chcemy podzielić dane. Ponadto początkowe położenie centroidów determinuje wynik grupowania. Algorytm nie radzi sobie również z danymi silnie zaszumionymi i/lub zawierającymi obserwacje odstające (zaburzenie średniej).

W środowisku R istnieje funkcja *kmeans* w standardowym pakiecie *stats*, która realizuje algorytm k-średnich.

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"), trace=FALSE)
```

Jej główne parametry wejściowe to:

x	–	zbiór danych (numerycznych) do pogrupowania
centers	–	wstępne położenie centroidów lub ich liczba oznaczająca na jak wiele grup należy podzielić dane
iter.max	–	maksymalna liczba iteracji algorytmu, warunek stopu

Pozostałe parametry związane są z wewnętrzną implementacją algorytmu w pakiecie R i nie będą rozpatrywane w ramach projektu. Wynikiem działania algorytmu jest model zawierający wektor przyporządkowania obserwacji do grup, środki wyznaczonych grup oraz pomocnicze miary (suma kwadratów odległości między przykładami w grupie, liczność grup, liczba wykonanych iteracji)

2.2 Algorytm k-medoidów

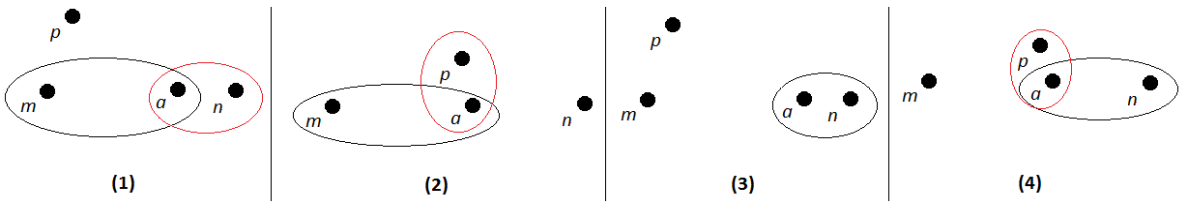
Struktura algorytmu k-medoidów jest niemal taka sama jak algorytmu k-średnich. Jedną różnicą jest przyjęcie, że środkiem grupy jest medoid, a nie wartość średnia. Medoid jest

najbardziej centralnym przykładem w grupie (średnia grupy może się nie pokrywać z żadnym przykładem należącym do wyznaczonej grupy). Powoduje to uodpornienie algorytmu na wartości odstające.

Aby wybrać nowy medoid, w kolejnych iteracjach algorytmu, rozważane są wszystkie przykłady p. Istnieją 4 możliwe przypadki, które należy sprawdzić by stwierdzić czy przykład p (niebędący medoidem) może zastąpić medoid m:

1. Przykład a należy do grupy medoidu m – jeśli zastąpimy m przykładem p oraz a znajduje się bliżej medoidu n to przydziel a do grupy n.
2. Przykład a należy do grupy medoidu m – jeśli zastąpimy m przykładem p oraz a znajduje się bliżej przykładu p to przydziel a do nowego medoidu p.
3. Przykład a należy do grupy medoidu n – jeśli zastąpimy medoid m przykładem p oraz a znajduje się bliżej medoidu n to nie zmieniaj przydziału przykładu a.
4. Przykład a należy do grupy medoidu n – jeśli zastąpimy medoid m przykładem p oraz a znajduje się bliżej p to przydziel a do nowego medoidu p.

Powyższe możliwości ilustruje wykres 1.



Wykres 1: Przypadki zamiany medoidu m przykładem p.

Przebieg grupowania przedstawia algorytm 2.

Algorytm 2 k-medoidów

1. Wybierz K przykładów jako medoidy.
2. Przyporządkuj każdej obserwacji najbliższy jej medoid.
3. Dla każdego medoidu m wybierz inny przykład p, który zmniejszy odległość między przykładami w obrębie nowej grupy.
4. Powtarzaj kroki 2. i 3., dopóki medoidy zmieniają się lub nie zostanie osiągnięta maksymalna liczba iteracji.

W języku R istnieje algorytm *pam* (ang. partitioning around medoids) w pakiecie cluster, który implementuje algorytm k-medoidów.

```
pam(x, k, diss = inherits(x, "dist"), metric = "euclidean",
    medoids = NULL, stand = FALSE, cluster.only = FALSE, do.swap = TRUE,
    keep.diss = !diss && !cluster.only && n < 100,
    keep.data = !diss && !cluster.only, pamonce = FALSE, trace.lev = 0)
```

Najważniejsze parametry wejściowe to:

x	–	zbiór danych (numerycznych) do pogrupowania
k	–	docelowa liczba grup
metric	–	metryka, według której obliczana jest odległość
stand	–	flaga binarna określająca czy dane mają być standaryzowane

W wyniku działania algorytmu otrzymujemy klasę, która zawiera informacje o wyznaczonych medoidach oraz przydział grupy dla każdego przykładu. Ponadto można znaleźć informację o miarach odległości między obserwacjami.

2.3 Algorytm hierarchiczny

Algorytmy hierarchiczne pozwalają na graficzną reprezentację struktury klasteryzacji w postaci drzewa. Można wyróżnić tu dwa podejścia: aglomeracyjne, gdzie każda z obserwacji stanowi na początku oddzielną grupę oraz partycjonujące - rozpoczynające od jednej grupy zawierającej wszystkie próbki. Przebieg aglomeracji przedstawia algorytm 3.

Algorytm 3 hierarchiczny

1. Utwórz jednoelementowe grupy.
2. Zbuduj macierz odległości pomiędzy rozpatrywanymi elementami.
3. Znajdź parę elementów, między którymi odległość jest najmniejsza.
4. Połącz znalezione grupy w jedną i wyznacz ich nowy środek ciężkości jako średnią środków cieżkości grup składowych.
5. Powtarzaj kroki 2-4 aż do uzyskania jednego skupiska zawierającego wszystkie próbki.

Do zalet algorytmów hierarchicznych należy brak konieczności początkowego określenia liczby klas w przeciwieństwie do opisanych wyżej algorytmów k-średnich i k-medoidów. Do wad można zaliczyć dość duże zróżnicowanie wyników w zależności od wybranych metod łączenia grup, wśród których znajdują się m.in. metody najbliższych i najdalszych sąsiadów, metoda średnich (centroidów) oraz minimalnej wariancji Warda.

Algorytmem aglomeracyjnym dostępnym w języku R jest agnes (ang. agglomerative nesting) w pakiecie cluster.

```
agnes(x, diss = inherits(x, "dist"), metric = "euclidean",
      stand = FALSE, method = "average", par.method,
      keep.diss = n < 100, keep.data = !diss, trace.lev = 0)
```

Jego najważniejsze parametry wejściowe to:

x	–	zbiór danych (numerycznych) do pogrupowania
metric	–	metryka, według której obliczana jest odległość
method	–	metoda łączenia grup

Wynikiem działania algorytmu jest graficzna reprezentacja struktury grupowania - dendrogram. Ponadto obliczany jest współczynnik aglomeracji, który charakteryzuje wygląd dendrogramu: niski współczynnik oznacza węższe struktury.

2.4 Drzewa decyzyjne

Pierwszą z metod klasyfikacji stosowaną w celu porównania z algorytmami grupowania są drzewa decyzyjne. W języku R istnieje wiele jej implementacji, z których wykorzystany zostanie algorytm C4.5 przedstawiony poniżej.

Algorytm 4 Algorytm C4.5

1. Utwórz zbiór treningowy T.
2. Wybierz taki atrybut, który najlepiej różnicowałby przykłady ze zbioru T.
3. Utwórz węzeł drzewa odpowiadający wybranemu atrybutowi.
4. Do węzła dodaj podwęzły, z których każdy reprezentuje pewną wartość badanego atrybutu.
5. Powtarzaj podziały dla kolejnych podwęzłów.

Do niewątpliwych zalet drzew decyzyjnych należy czytelna forma reprezentacji, efektywność pamięciowa oraz wszechstronność metody. Konieczny jest w niej jednak kompromis pomiędzy wielkością drzewa a jakością klasyfikacji. Drzewa mogą być również podatne na zjawisko nadmiernego dopasowania.

Wykorzystaną w projekcie funkcją będzie J48 z pakietu RWeka.

```
J48(formula, data, subset, na.action,
    control = Weka_control(), options = NULL)
```

Jego najważniejsze parametry wejściowe to:

formula – symboliczny opis modelu
data – dane treningowe

Wynikiem działania funkcji jest schemat drzewa decyzyjnego wraz z opisem jego węzłów oraz dane o jego wielkości.

2.5 Metoda k najbliższych sąsiadów

Metoda k najbliższych sąsiadów zaliczana jest do grupy algorytmów tzw. leniwego uczenia. Nie tworzy ona żadnej reprezentacji danych w postaci modelu, a szuka rozwiązania dopiero w momencie pojawienia się przykładu do klasyfikacji. Aby zastosować metodę najbliższego sąsiada konieczne jest przedstawienie obiektów w n-wymiarowej przestrzeni po czym umieszczenie w niej obiektu testowanego. Klasyfikacja sprowadza się do sprawdzenia, do jakiej klasy należy obiekt najbliższy obiektowi testowanemu. Jeżeli wybrany został wariant metody z k sąsiadów to najpierw konieczne jest rozstrzygnięcie, która klasa dominuje wśród nich.

Algorytm 5 Algorytm k najbliższych sąsiadów

1. Oblicz odległości klasyfikowanego przykładu od przykładów ze zbioru treningowego.
2. Znajdź k najbliższych sąsiadów.
3. Sklasyfikuj przykład na podstawie klas sąsiadów.

Podstawową zaletą algorytmu kNN jest jego prostota. Ma on jednak szereg wad, do których należy długi czas obliczeń w przypadku licznych zbiorów treningowych, duże wymagi pamięciowe oraz konieczność wstępnej normalizacji danych. Język R oferuje tutaj funkcję kNN.

```
knn(train, test, cl, k = 1, prob = FALSE,
algorithm=c("kd_tree", "cover_tree", "brute"))
```

Jego najważniejsze parametry wejściowe to:

train – zbiór trenujący
test – zbiór danych testowych
cl – prawdziwe klasy, do których należą przykłady zbiory trenującego
k – liczba sąsiadów

3 Opis badań

3.1 Planowane eksperymenty

W projekcie przeprowadzone zostaną następujące eksperymenty:

1. detekcja anomalii przy wykorzystaniu różnych wskaźników nietypowości dla różnych metod grupowania,

2. porównanie grupowania z klasyfikacją za pomocą drzew decyzyjnych,

3. porównanie grupowania z klasyfikacja za pomocą metody k najbliższych sąsiadów.

W celu stwierdzenia czy testowany przykład należy zaliczyć do jednej z wyznaczonych grup czy też oznaczyć go jako anomalię wykorzystamy niestandardowe wskaźniki omówione w [2].

Pierwszym z proponowanych wskaźników jest CBLOF liczony jako iloczyn odległości d badanej próbki p od najbliższej dużej grupy ($C \in LC$) i liczby elementów w grupie, do której obiekt został zaklasyfikowany. Koncepcja małych (SC) i dużych (LC) grup nie jest precyzyjnie określona - możliwy jest wybór algorytmu podziału.

$$CBLOF(p) = \begin{cases} |C_i| \cdot min(d(p, C_j)), & \text{jeśli } C_i \in SC, \text{ gdzie } p \in C_i \text{ oraz } C_j \in LC \\ |C_i| \cdot d(p, C_j), & \text{jeśli } C_i \in LC, \text{ gdzie } p \in C_i \end{cases} \tag{2}$$

Wskaźnik ten powinien rosnąć wraz z odległością próbki od dużej grupy, a zatem wskazywać na stopień anomali - im wyższa jego wartość, tym obiekt bardziej oddalony od grup. Jednak ze względu na fakt, że uwzględniana jest w nim również liczność grupy algorytm ten może dawać nieprawidłowe wyniki. Jako anomalie mogą zostać zaklasyfikowane próbki znajdujące się blisko bardzo licznych zbiorów.

Lepszym rozwiązaniem może być zatem nieważony wskaźnik CBLOF oparty jedynie na odległości od grup, z pominięciem ich liczności. W projekcie zostaną zastosowane obie wersje ocen anomalii i wykonane zostanie porównanie między nimi.

$$u - CBLOF(p) = \begin{cases} min(d(p, C_j)), & \text{jeśli } p \in SC, \text{ gdzie } C_j \in LC \\ d(p, C_i), & \text{jeśli } p \in C_i \in LC \end{cases} \tag{3}$$

Inną miarą oceny anomalii jest LDCOF charakteryzująca się normalizacją wyników dla próbki względem jej sąsiedztwa. Definiowana jest jako iloraz odległości próbki od najbliższej dużej grupy i średniego dystansu między elementami tej dużej grupy i jej środkiem.

$$distance_{avg}(C) = \frac{\sum_{i \in C} d(i, C)}{|C|} \tag{4}$$

$$LDCOF(p) = \begin{cases} \frac{min(d(p,C_j))}{distance_{avg}(C_j}, & \text{jeśli } p \in C_i \in SC, \text{ gdzie } C_j \in LC \\ \frac{d(p,C_j)}{distance_{avg}(C_j}, & \text{jeśli } p \in C_i \in LC \end{cases} \tag{5}$$

3.2 Zbiory danych

Badania przeprowadzone zostaną na kilku zbiorach danych o numerycznych typach atrybutów. Wynika to z parametrów przyjmowanych przez wybrane algorytmy grupowania. Nie oznacza to jednak, że nie można grupować atrybutów dyskretnych. Wymagałoby to ich przewartościowania na wartości numeryczne.

Dla wszystkich zbiorów danych konieczne będzie określenie sposobu postępowania z brakującymi wartościami. W przypadku pojedynczych braków pewne próbki zostaną najprawdopodobniej pominięte lub też zastąpione średnią wartością atrybutu. Jeżeli brakujących wartości byłoby więcej, lepsze wyniki dałaby predykcja danej wartości.

Wstępne przetwarzanie atrybutów będzie odbywać się przed dostarczeniem danych do metody wyznaczającej grupowanie. Możliwe będzie również ograniczenie liczby wykorzystanych atrybutów. Klasyfikator k najbliższych sąsiadów będzie z kolei prawdopodobnie wymagał normalizacji danych.

Wykorzystane zostaną przykładowe zbiory danych pochodzące z UCI Machine Learning Repository:

- Letter Recognition Data Set – zbiór danych dotyczących zdjęć liter alfabetu, zawiera 16 atrybutów oraz 26 klas
- Mushroom Data Set – zbiór danych opisujących grzyby. Możliwe są dwie klasy - grzyby jadalne lub trujące, zawiera 22 atrybuty
- Dataset for Sensorless Drive Diagnosis Data Set – zbiór danych diagnostycznych dla napędów komputerowych, możliwość klasyfikacji na podstawie 49 atrybutów do 11 klas

3.3 Ocena jakości

Aby ocenić jakość rozwiązania trzeba sprawdzić skuteczność algorytmów. Do tego budowana będzie macierz pomyłek. W jej wierszach znajdują się klasy oryginalne do których należały obiekty, a w kolumnach klasy przewidziane. Na przecięciu wstawiana jest liczba obiektów poprawnie lub niepoprawnie sklasyfikowanych.

W przypadku binarnym rozważmy dwie klasy: pozytywną i negatywną. Do klasy pozytywnej będziemy zaliczali wszystkie przykłady zaliczone do swoich prawdziwych grup, a do klasy negatywnej obserwacje stanowiące anomalie. W sytuacji, gdy pewna pozytywna obserwacja zostanie zaklasyfikowana jako negatywna lub odwrotnie, korzystamy z tabeli 1. Przynależność do klasy rzeczywistej oznacza faktyczną klasyfikację obserwacji do jednej z klas, a wynik klasyfikacji to decyzja o przynależności podjęta przez algorytm. Możliwe wyniki to:

– TP (ang. *true positive*) – poprawne zaklasyfikowanie do rzeczywistej grupy

– FN (ang. *false negative*) – błędne zaklasyfikowanie przykładu (jako anomalii) do klasy negatywnej, gdy tak naprawdę należy do jednej ze znalezionych grup

– FP (ang. *false positive*) – błędne zaklasyfikowanie obserwacji (do jednej ze znalezionych grup) do klasy pozytywnej, gdy tak naprawdę są to anomalie

– TN (ang. *true negative*) – poprawne wykrycie anomalii

Tabela 1: Macierz pomyłek

		wynik klasyfikacji	
		pozytywna	negatywna
klasa rzeczywista	pozytywna	TP	FN
	negatywna	FP	TN

Na tej podstawie można wyznaczyć wskaźniki jakości rozwiązania:

– dokładność – liczba poprawnie sklasyfikowanych obserwacji wśród wszystkich wyników klasyfikacji

$$dokładność = \frac{TP + TN}{TP + TN + FP + FN}$$

– precyzja – liczba poprawnie sklasyfikowanych przykładów wśród wszystkich obserwacji zaklasyfikowanych do znalezionych grup

$$precyzja = \frac{TP}{TP + FP}$$

4 Kwestie otwarte

W chwili pisania niniejszego sprawozdania nie przewidujemy zmiany domyślnych parametrów algorytmów, które nie są bezpośrednio związane z zadaniem (np. wybór implementacji algorytmu k-średnich w ramach funkcji *kmeans*). Jeśli podczas testów okaże się, że pewne parametry nieomówione powyżej będą zmieniane, zostanie podana stosowna adnotacja.

Do oceny jakości klasyfikacji zostały wybrane wskaźniki dokładności i precyzji. W sytuacji gdy wskaźniki te nie będą w stanie wystarczająco jednoznacznie wyznaczyć, które z rozwiązań jest lepsze, zostaną rozważone i opisane inne miary jakości.

Literatura

[1] http://wazniak.mimuw.edu.pl/images/8/86/ED-4.2-m11-1.01.pdf

[2] https://www.goldiges.de/publications/Anomaly_Detection_Algorithms_for_RapidMiner.pdf