Wrocław University
of Science and Technology

Zastosowania Informatyki w Medycynie

Metody rozpoznawania bez uczenia – analiza skupień (metody grupowania danych)

Spotkanie konsultacyjne

Prowadzący: mgr inż. Paweł Zyblewski
Termin seminarium: Czwartek 15:15-16:55

Skład grupy:
inż. Kamil Zdeb 235871
inż. Paweł Szynal 226026
inż.Tomasz Florczuk 235715

Wydział Elektroniki
Kierunek: Informatyka

Wrocław 11 marca 2021

1 Źródła

1. k-means and hierarchical clustering by hand and in R
2. How DBSCAN works and why should we use it
3. DBSCAN Clustering — Explained
4. Clustering Using OPTICS
5. Selection of K in K-means clustering D T Pham, S S Dimov, C D Nguyen
   sci-hub.se
6. Johnson, S. C. (1967). Hierarchical clustering schemes
   sci-hub.se
7. Cluster analysis for researchers
8. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
9. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

2 Analiza skupień

Analiza skupień inaczej zwana również analizą klasterową (cluster analysis) ma na celu pogrupowanie badanych elementów w podobne do siebie grupy.

Ideą analizy skupień jest takie pogrupowanie badanych osób, aby wedle wyznaczonych kryteriów wyodrębnić podobne do siebie jednostki w oddzielne grupy. Stosowana jest tutaj zasada podobieństwa wewnętrznego i niepodobieństwa zewnętrznego. Innymi słowy, grupowanie polega na takim przyporządkowaniu obiektów do grup, aby wewnątrz każdej z wydzielonych grup jednostki w niej znajdujące się były podobne do siebie, ale różne wyodrębnione grupy były jak najmniej podobne do siebie.

W analizie skupień wybieramy również kryteria, według, których grupujemy obserwacje. Tymi kryteriami są analizowane zmienne. Przykład:

WSTĘP

Podczas prezetnwaoania często będę się posługiwał pojęciami o które są synonimami (grupa - klaster - zbiory- (grupa podobnych rzeczy lub osób umieszczonych lub występujących blisko siebie.) jako osoby zwiazane z IT bardzije przypomina nam to jednostkę pamięci w dyskukładające się lub kilku sektorów;

)

grupa podobnych rzeczy lub osób umieszczonych lub występujących blisko siebie.

Klaster to zbiór obiektów, w których każdy obiekt jest bliżej (bardziej podobny) do prototypu definiującego klaster niż do prototypu dowolnego innego klastra. W przypadku danych z atrybutami ciągłymi prototyp klastra jest często centroidem, tj. Średnią (średnią) wszystkich punktów w klastrze. Kiedy centroid nie ma znaczenia, na przykład gdy dane mają atrybuty kategorialne,

2.1 Zastosowanie

Klasy lub znaczące grupy obiektów, które mają wspólne cechy, odgrywają ważną rolę w tym, jak ludzie analizują i opisują świat.

Istoty ludzkie rzeczywiście potrafią dzielić obiekty na grupy (grupowanie) i przypisywać określone obiekty do tych grup (klasyfikacja).

Na przykład nawet stosunkowo małe dzieci mogą szybko oznaczyć obiekty na zdjęciu jako budynki, pojazdy, ludzi, zwierzęta, rośliny itp.

W kontekście zrozumienia danych, klastry są potencjalnymi klasami , a analiza skupień jest badaniem technik automatyczne wyszukiwanie zajęć. Oto kilka przykładów:

- Biologia.

Biolodzy spędzili wiele lat na tworzeniu taksonomii (hierarchicznej klasyfikacji) wszystkich żywych istot: królestwa, gromady, klasy, porządku, rodziny, rodzaju i gatunku. Dlatego nie jest być może zaskakujące, że znaczna część wczesnych prac nad analizą skupień miała na celu stworzenie dyscypliny taksonomii matematycznej, która mogłaby automatycznie znaleźć takie struktury klasyfikacyjne. Niedawno biolodzy zastosowali grupowanie do analizy dużych ilości informacji genetycznej, które są obecnie dostępne. Na przykład grupowanie zostało użyte do znalezienia grup genów o podobnych funkcjach.

- Klimat

Zrozumienie klimatu Ziemi wymaga znalezienia wzorców w atmosferze i oceanie. W tym celu zastosowano analizę skupień, aby znaleźć wzorce ciśnienia atmosferycznego w regionach polarnych i obszarach oceanu, które mają znaczący wpływ na klimat na lądzie.

- Biznes

Firmy zbierają duże ilości informacji o obecnych i potencjalnych klientach. Klastrowanie można wykorzystać do podzielenia klientów na niewielką liczbę grup w celu przeprowadzenia dodatkowych analiz i działań marketingowych.

2.2 Slajd nr 5

Analiza skupień grupuje obiekty danych tylko na podstawie informacji znalezionych w danych, które opisują obiekty i ich relacje.

Celem jest, aby obiekty w grupie były podobne (lub pokrewne) do siebie i różniły się (lub niezwiązane) z obiektami w innych grupach.

Im większe podobieństwo (lub jednorodność) w grupie i im większa różnica między grupami, tym lepsze lub bardziej wyraźne skupienie. W wielu aplikacjach pojęcie klastra nie jest dobrze zdefiniowane.

Aby lepiej zrozumieć trudność podejmowania decyzji, co stanowi klaster, rozważ rysunek 8.1, który przedstawia dwadzieścia punktów i trzy różne sposoby ich podziału na klastry.

Kształty znaczników wskazują na przynależność do klastra. Rysunki 8.1 (b) i 8.1 (d) dzielą dane odpowiednio na dwie i sześć części. Jednak pozorny podział każdej z dwóch większych gromad na trzy podgrupy może być po prostu artefaktem ludzkiego układu wzrokowego. Nie jest również nierozsądne stwierdzenie, że punkty tworzą cztery skupienia, jak pokazano na rysunku 8.1 (c). Rysunek ten pokazuje, że definicja klastra jest nieprecyzyjna i że najlepsza definicja zależy od charakteru danych i pożądanych wyników. Analiza skupień jest powiązana z innymi technikami używanymi do dzielenia obiektów danych na grupy. Na przykład grupowanie można traktować jako formę klasyfikacji, ponieważ tworzy etykietowanie obiektów etykietami klas (klastrów). Jednak wyprowadza te etykiety tylko z danych. W przeciwieństwie do klasyfikacji

2.3 Wtrącenie

Ponadto, terminy segmentacja i partycjonowanie są czasami używane jako synonimy dla grupowania, terminy te są często używane w podejściach wykraczających poza tradycyjne granice analizy skupień. Na przykład termin partycjonowanie jest często używany w połączeniu z technikami, które dzielą grafy na podgrafy i które nie są silnie związane z grupowaniem. Segmentacja często odnosi się do podziału danych na grupy przy użyciu prostych technik; np. obraz można podzielić na segmenty tylko na podstawie intensywności i koloru pikseli lub ludzi można podzielić na grupy na podstawie ich dochodów. Niemniej jednak niektóre prace związane z partycjonowaniem grafów oraz segmentacją obrazu i rynku są związane z analizą klastrów.

2.4 Różne typy klastrów. Slajd nr 6

Wyróżniamy różne typy klastrów: hierarchiczne (zagnieżdżone) i częściowe (niezagnieżdżone), wyłączne, nakładające się i rozmyte oraz pełne i częściowe.

Grupowanie częściowe to po prostu podział zbioru obiektów danych na nienakładające się podzbiory (klastry) w taki sposób, że każdy obiekt danych znajduje się dokładnie w jednym podzbiorze. Z osobna każdy zbiór klastrów na rycinach 8.1 (b – d) jest grupowaniem częściowym.

W klastrach rozmytych każdy obiekt należy do każdego klastra o wadze przynależności od 0 (absolutnie nie należy) do 1 (absolutnie należy). Innymi słowy, skupienia są traktowane jako zbiory rozmyte. (Matematycznie, zbiór rozmyty to taki, w którym obiekt należy do dowolnego zbioru o wadze od 0 do 1. W przypadku grupowania rozmytego często nakładamy dodatkowe ograniczenie, że suma wag każdego obiektu musi być równa 1 .) Podobnie, probabilistyczne techniki grupowania obliczają prawdopodobieństwo, z jakim każdy punkt należy do każdego klastra, a te prawdopodobieństwa muszą również sumować się do 1. Be-

(well sepratated)

Rysunek 8.2 (a) przedstawia przykład dobrze oddzielonych klastrów, które składają się z dwóch grup punktów w dwuwymiarowej przestrzeni. Odległość między dowolnymi dwoma punktami w różnych grupach jest większa niż odległość między dowolnymi dwoma punktami w grupie. Dobrze oddzielone zbiory nie muszą być kuliste, i mogą mieć dowolny kształt.

(Prototype-Based) prototyp jest często medoidą, tj. Najbardziej reprezentatywnym punktem skupienia. W przypadku wielu typów danych prototyp można uznać za najbardziej centralny punkt iw takich przypadkach zwykle nazywamy klastry oparte na prototypach jako skupione w centrum. Nic dziwnego, że takie gromady są zwykle kuliste. Rysunek 8.2 (b) przedstawia przykład klastrów skupionych w centrum.

- K-oznacza. Jest to oparta na prototypach, partycjonalna technika grupowania, która próbuje znaleźć określoną przez użytkownika liczbę klastrów (K), które są reprezentowane przez ich centroidy.

Aglomeracyjne hierarchiczne grupowanie. To podejście do grupowania odnosi się do zbioru ściśle powiązanych technik tworzenia klastrów, które tworzą hierarchiczne grupowanie, rozpoczynając od każdego punktu jako pojedynczy klaster, a następnie wielokrotnie łącząc dwa najbliższe klastry, aż pozostanie pojedynczy, obejmujący wszystko klaster. Niektóre z tych technik mają naturalną interpretację pod względem grupowania opartego na grafach, podczas gdy inne mają interpretację w kategoriach podejścia opartego na prototypach.

- DBSCAN. Jest to algorytm klastrowania oparty na gęstości, który tworzy grupowanie partycjonowane, w którym liczba klastrów jest automatycznie określana przez algorytm. Punkty w regionach o niskiej gęstości są klasyfikowane jako hałas i pomijane; w związku z tym DBSCAN nie tworzy pełnego klastrowania.

3 Metoda k-means

Metoda wyznacza k środków, a następnie przyporządkowuje dane do jednej z grup. Metoda ta jest przydatna jeżeli znana jest liczba grup, do której przyporządkowujemy obiekty. Do klasyfikacji wykorzystuje odległości między punktami (najczęściej w ujęciu euklidesowym). Jeżeli dane na podstawie których dokonywana jest klasyfikacja są różnego rzędu, należy dokonać ich skalowania. Punkty będące środkami grup wybierane są losowo. W związku z tym można przeprowadzić taką analizę kilkukrotnie dla różnych środków grup w celu uzyskania lepszych rezultatów.

Definicje:

- BSS - Between Sum of Squares
- TSS - Total Sum of Squares
- Cluster

Co poruszyć:

1. Sposoby wyznaczania optymalnej liczby grup (arbitralne, elbow method, average silhouette method, Gap statistic method) - omówienie
2. Wyjaśnić sposób działania (początkowe środki grup, klasyfikacja obiektów)
3. Jakość podziału - sposób obliczania

Zalety:

- Dla złożonych modeli (duża liczba zmiennych) dla niewielkiej liczby grup działa szybciej niż klasteryzacja hierarchiczna
- Stosunkowo prosta implementacja

Wady:

- Znalezienie odpowiedniej liczby grup nie jest trywialne
- Losowy wybór środków grup powoduje niestabilność wyników
- Wartości odstające mogą wpłynąć na klasteryzację

4 Metody klasteryzacji hierarchicznej

W eksploracji danych i statystyce hierarchiczne grupowanie (HCA) jest metodą analizy klastrów, która ma na celu zbudowanie ich hierarchii. Strategie hierarchicznego grupowania ogólnie dzielą się na dwa typy:

1. Aglomeracyjne
2. Podziału

4.1 Aglomeracja a podział hierarchiczny

Hierarchiczne techniki grupowania są drugą ważną kategorią metod klastrowania. Podobnie jak w przypadku K-średnich, podejścia te są stosunkowo stare w porównaniu z wieloma algorytmami grupowania, ale nadal cieszą się szerokim zastosowaniem. Istnieją dwa podstawowe podejścia do generowania hierarchicznego grupowania:

Aglomeracyjny: zacznij od punktów jako oddzielnych skupień i na każdym kroku połącz najbliższą parę skupień. Wymaga to zdefiniowania pojęcia bliskości klastra.

Dzielenie: Rozpocznij od jednego, obejmującego wszystko klastra i na każdym kroku dziel klaster, aż pozostaną tylko pojedyncze skupiska poszczególnych punktów. W takim przypadku musimy zdecydować, który klaster podzielić na każdym kroku i jak to zrobić.

Techniki skupień aglomeracyjnych hierarchicznych są zdecydowanie najpowszechniejsze i w tej sekcji skupimy się wyłącznie na tych metodach. Technika grupowania hierarchicznego, która dzieli się na podziały, została opisana w sekcji 9.4.2. Hierarchiczne grupowanie jest często wyświetlane graficznie za pomocą diagramu przypominającego drzewo zwanego dendrogramem, który wyświetla zarówno podklaster klastra

Powszechnie określane jako AGNES (AGglomerative NESting) działa w sposób bottom-up. Nakażdym etapie algorytmu dwa najbardziej podobne klastry są łączone w nowy większy klaster(węzły). Ta procedura jest powtarzana, aż wszystkie punkty będą członkami tylko jednego dużegoklastra (korzenia). Rezultatem jest drzewo, które można wyświetlić za pomocą dendrogramu

Podziałowe hierarchiczne grupowanie: powszechnie określane jako DIANA (DIvise ANAlysis)działa w sposób odgórny. DIANA jest jak rewers AGNES. Na każdym etapie algorytmu bieżącyklaster jest dzielony na dwa klastry uważane za najbardziej niejednorodne. Proces jest powtarzany,aż wszystkie obserwacje znajdą się we własnym klastrze.

ALGORYTM

Wiele aglomeracyjnych hierarchicznych technik klastrowania to odmiany jednego podejścia: zaczynając od pojedynczych punktów jako klastrów, kolejno łącz dwa najbliższe klastry, aż pozostanie tylko jeden klaster. Podejście to jest bardziej formalnie wyrażone w algorytmie 8.3.

Kluczowym działaniem Algorytmu 8.3 jest obliczenie bliskości pomiędzy dwoma klastrami i to właśnie definicja bliskości klastra, która rozróżnia różne aglomeracyjne techniki hierarchiczne, które będziemy omawiać. Bliskość klastra jest zwykle definiowana z myślą o konkretnym typie klastra - patrz sekcja 8.1.2. Na przykład wiele aglomeracyjnych hierarchicznych technik grupowania, takich jak MIN, MAX i Średnia grupowa, pochodzi z graficznego widoku klastrów. MIN definiuje bliskość klastra jako bliskość między dwoma najbliższymi punktami, które są w różnych klastrach, lub używając terminów grafowych, najkrótszą krawędź między dwoma węzłami w różnych podzbiorach węzłów. Daje to klastry oparte na ciągłości, jak pokazano na rysunku 8.2 (c). Alternatywnie MAX przyjmuje bliskość między dwoma najdalszymi punktami w różnych klastrach jako bliskość klastra lub, używając terminów grafowych, najdłuższą krawędź między dwoma węzłami w różnych podzbiorach węzłów. (Jeśli odległościami są odległości, wówczas nazwy MIN i MAX są krótkie i sugestywne. Jednak w przypadku podobieństw, gdzie większe wartości wskazują bliższe punkty, nazwy wydają się odwrócone. Z tego powodu zwykle wolimy używać nazw alternatywnych, pojedyncze łącze i pełne łącze.) Inne podejście oparte na wykresie, technika średniej grupy, definiuje bliskość klastra jako średnie odległości parami (średnia długość krawędzi) wszystkich par punktów z różnych klastrów. Rysunek 8.14 ilustruje te trzy podejścia.

Wspomnieliśmy wcześniej, że hierarchiczne grupowanie aglomeracyjne nie może być postrzegane jako globalna optymalizacja funkcji celu. Zamiast tego, techniki hierarchicznego grupowania aglomeracyjnego wykorzystują różne kryteria, aby na każdym etapie decydować lokalnie, które klastry należy połączyć (lub podzielić w przypadku podejść dzielących). Takie podejście zapewnia algorytmy grupowania, które pozwalają uniknąć trudności związanych z próbą rozwiązania trudnego problemu optymalizacji kombinatorycznej. (Można wykazać, że ogólny problem klastrowania dla funkcji celu, takiej jak „minimalizacja SSE", jest niewykonalny obliczeniowo). Ponadto takie podejścia nie mają problemów z lokalnymi minimami ani trudnościami w wyborze punktów początkowych. Oczywiście złożoność czasowa O (m log m) i złożoność przestrzenna O (m są w wielu przypadkach przeszkodą).

echniki skupień aglomeracyjnych hierarchicznych są zdecydowanie najpowszechniejsze i w tej sekcji skupimy się wyłącznie na tych metodach. Technika grupowania hierarchicznego, która dzieli się na podziały, została opisana w sekcji 9.4.2. Hierarchiczne grupowanie jest często wyświetlane graficznie za pomocą diagramu przypominającego drzewo zwanego dendrogramem, który wyświetla zarówno podklaster klastra

Dowolne z bliskości klastrów, które omówiliśmy w tej sekcji, można postrzegać jako wybór różnych parametrów (we wzorze Lance-Williamsa pokazanym poniżej w Równaniu 8.7) dla bliskości między klastrami Q i R, gdzie R jest tworzone przez scalanie klastrów A i B. W tym równaniu p (.,.) Jest funkcją bliskości, podczas gdy m, m i m to odpowiednio liczba punktów w klastrach A, B i Q. Innymi słowy, po połączeniu klastrów A i B w celu utworzenia klastra R, bliskość nowej klastra, R, istniejącej klastra Q, jest funkcją liniową odległości Q w stosunku do pierwotnych klastrów A i B. Tabela 8.5 pokazuje wartości tych współczynników dla technik, które omówiliśmy.

Żadna technika hierarchicznego grupowania, którą można wyrazić za pomocą wzoru Lance-Williamsa, nie musi zachowywać oryginalnych punktów danych. Zamiast tego macierz bliskości jest aktualizowana w miarę tworzenia klastrów. Chociaż ogólna formuła jest atrakcyjna, zwłaszcza w przypadku implementacji, łatwiej jest zrozumieć różne metody hierarchiczne, patrząc bezpośrednio na definicję bliskości klastra, której używa każda metoda.

5 Metody DBSCAN i OPTICS

5.1 DBSCAN

1. Metoda analizy skupień bazująca na gęstości, bardziej efektywna w przypadku losowych kształtów oraz wartości odstających. Jest w stanie znaleźć regiony o większym skupieniu i oddzielić je od tych o mniejszej gęstości

2. Główne założenie - punkt należy do zbioru gdy jest w pobliżu wielu innych punktów należących do tego zbioru

3. Główne parametry - eps i minPoints

   (a) eps - wartość określająca sąsiedztwo punktów. Jeśli odległośc między punktami jest mniejsza lub równa tej wartości, punkty uznwane są za sąsiadów. Odległość między punktami obliczana jest tak jak w metodzie k-means. Najpopularniejszą jest klasyczna odległość euklidesowa

   (b) minPoints - minimalna ilość punktów aby utworzyć zbiór Przy użyciu tych dwóch parametrów, można podzielić punkt na trzy jego rodzaje:

   - Punkt źródłowy, jeśli w jego sąsiedztwie określonym przez wartośc eps znajduje się przynajmniej tyle punktów ile wskazuje wartość minPoints
   - Punkt graniczny, jeśli znajduje się w sąsiedztwie punktu źródłowego, ale w jego sąsiedztwie znajduje się mniej punktów niż wskazuje wartość minPoints
   - Punkt odstający, jeśli nie znajduje się w sąsiedztwie żadnego innego punktu

4. Sposób działania

   (a) Wybierane są wartości eps i minPoints

   (b) Losowo wybierany jest punkt początkowy, a następenie sprawdzane jest jego sąsiedztwo. Jeśli punkt jest punktem źródłowym, tworzony jest zbiór. W przeciwnym przypadku punkt określany jest jako szum. Wszystkie punkty w sąsiedztwie punktu źródłowego zaliczane są do tego zbioru. Jeśli któryś z tych punktów również spełnia wymagania by stać się punktem źródłowym, jego sąsiedzi również zawierają się w zbiorze

   (c) Następnie losowo wybierany jest kolejny punkt, który nie został jeszcze sprawdzony. Powtarzany jest poprzedni punkt algorytmu

   (d) Algorytm kończy się gdy wszystkie punkty zostaną sprawdzone

5. Plusy

   - Nie wymaga wcześniejszego podania liczby zbiorów
   - Efektywny ze zbiorami o losowych kształtach
   - Wykrywa oraz dobrze radzi sobie z wartościami odstającymi

6. Minusy

   - Znalezienie odpowiedniej wartości sąsiedztwa(eps) może być problematyczne

5.2 OPTICS

Prawie bezparametryczna metoda analizy skupień bazująca na gęstości mająca wiele wspólnego z algorytmem DBSCAN. Wykorzustuje wykres osiągalności do wyodrębnienia zbiorów oraz jest zdolna do wykrywania wartości odstających

1. Parametry
    - maximum epsilon
    - minPoints
2. Sposób działania
3. Co poruszyć
    - Wykres osiągalności, Core distance(odległość źródłowa) i Reachability distance(odleglość osiągalności) potrzebne do jego utworzenia