Uniwersytet Zielonogórski
Instytut Sterowania i Systemów Informatycznych

Eksploracja danych

2023

Laboratorium nr 10: Klasteryzacja danych

I. Zagadnienia teoretyczne

Wstęp

Analiza skupień zwana również klasteryzacją to technika wielowymiarowa grupująca obserwacje, które mają podobne wartości dla wielu zmiennych. Zazwyczaj obserwacje nie są rozproszone równomiernie w n-wymiarowej przestrzeni, ale raczej tworzą skupiska lub klastry. Identyfikacja tych klastrów zapewnia głębsze zrozumienie danych.

Klasteryzacja hierarchiczna polega na łączeniu kolejnych klastrów. Metoda rozpoczyna się od utworzenia przez każdą obserwacją własnego klastra. Na każdym etapie proces grupowania oblicza odległość między wszystkimi parami klastrów i łączy dwa skupienia, które są najbliżej siebie. Proces ten trwa do momentu, gdy wszystkie punkty znajdą się w jednym klastrze. Klastrowanie hierarchiczne jest również nazywane klastrowaniem aglomeracyjnym ze względu na stosowane w nim podejście łączenia. Proces aglomeracji jest przedstawiony jako drzewo, zwane dendrogramem. Aby pomóc Ci zdecydować o liczbie klastrów, JMP udostępnia wykres odległości. Możesz wybrać liczbę klastrów, określając, kiedy odległości między klastrami nie mają już praktycznego znaczenia. Klastrowanie hierarchiczne obsługuje również kolumny znakowe.

Metoda centoridów (**K Means**) konstruuje określoną liczbę klastrów przy użyciu algorytmu iteracyjnego, który dopasowuje obserwacje ze zbioru. Metoda opiera się na iteracyjnym procesie dopasowywania. Algorytm centroidów najpierw wybiera zestaw n punktów zwanych centroidami klastrów jako wstępne oszacowanie centralnego położenia wewnątrz klastrów. Każda obserwacja jest przypisywana do najbliższego centroidu klastra w celu utworzenia zestawu tymczasowych klastrów. Centroidy są następnie zastępowane przez nowo wyznaczone środki klastrów, następnie punkty są ponownie przydzielane, a proces jest kontynuowany, dopóki w klastrach nie wystąpią żadne dalsze zmiany. Algorytm centroidów jest szczególnym przypadkiem algorytmu EM, gdzie E oznacza oczekiwanie, a M maksymalizację. W przypadku algorytmu centroidów obliczenie tymczasowych średnich skupień reprezentuje krok Oczekiwanie, a przypisanie punktów do najbliższych skupień reprezentuje krok Maksymalizacja. Klastrowanie centroidów obsługuje tylko kolumny numeryczne. Klastrowanie centroidów ignoruje typy modelowania (nominalne i porządkowe) i traktuje wszystkie kolumny liczbowe jako ciągłe. Musisz wcześniej określić liczbę klastrów, k, lub zakres wartości dla k. Możesz jednak porównać wyniki różnych wartości k, aby wybrać optymalną liczbę skupień dla swoich danych.

Hierarchical Cluster jest przydatny w przypadku mniejszych tabel zawierających do kilkudziesięciu tysięcy wierszy i umożliwia pracę z danymi znakowymi. Hierarchiczne grupowanie łączy wiersze w hierarchiczną sekwencję, która jest przedstawiana jako drzewo. Możesz wybrać liczbę klastrów, która jest najbardziej odpowiednia dla twoich danych po zbudowaniu drzewa.

K Means Cluster jest odpowiedni dla większych tabel zawierających do milionów wierszy i pozwala tylko na zastosowanie danych liczbowych. Musisz wcześniej określić liczbę klastrów, k. Algorytm losuje punkty początkowe klastrów. Następnie przeprowadza iteracyjny proces naprzemiennego przypisywania punktów do klastrów i ponownego obliczania centrów klastrów.

II. Instrukcja do wykonania ćwiczeń

Klasteryzacja hierarchiczna

1) W tym ćwiczeniu pogrupujemy kraje według ich wskaźników urodzeń i zgonów na 100 000 osób w 1976 roku.

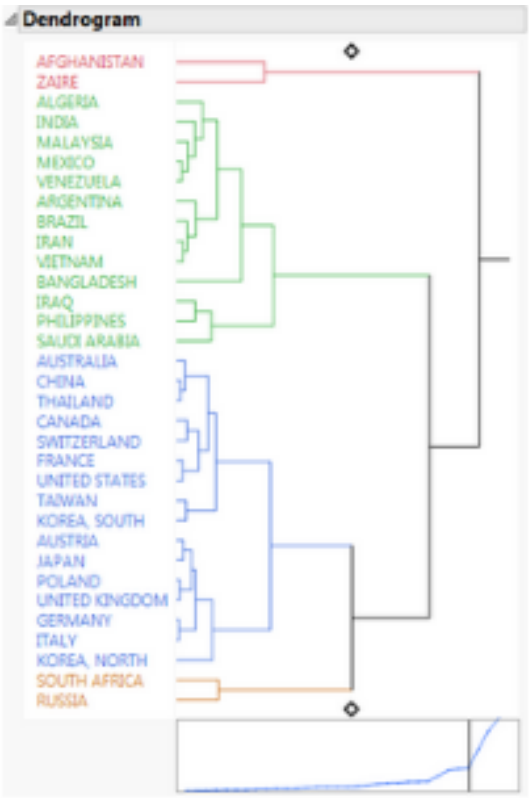
2) Wybierz **Help** > **Sample Data Library** i otwórz plik **Birth Death Subset.jmp**

3) Wybierz **Analyze** > **Clustering** > **Hierarchical Cluster**.

4) Wybierz **birth** oraz **death** i kliknij **Y, Columns**. Ten wybór zapewnia, że kolumna kraju, a nie numer wiersza, zostanie użyta do oznaczenia dendrogramu, który pojawi się po kliknięciu OK. Kliknij **OK**.

5) Kliknij czerwony trójkąt obok **Hierarchical Clustering** i wybierz **Color Clusters**.

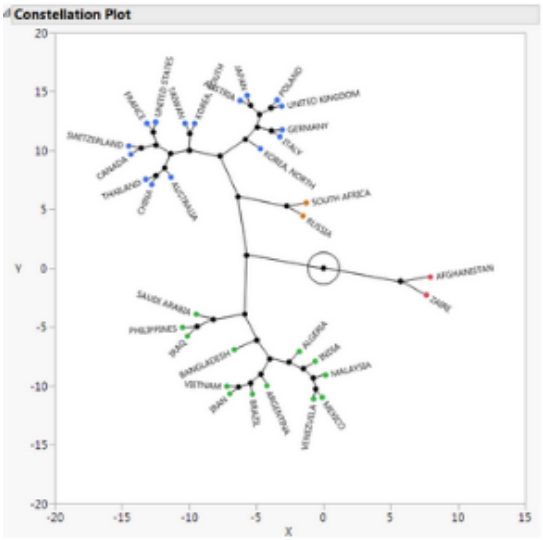
6) Dendrogram pokazuje, w jaki sposób przeprowadzana jest klasteryzacja. Proces klasteryzacji można obejrzeć, czytając dendrogram od lewej do prawej. Każdy krok polega na połączeniu dwóch najbliższych klastrów w jeden klaster. Na dendrogramie względne odległości między skupieniami są określone przez poziome odległości między pionowymi liniami łączącymi skupienia. Na przykład Afganistan i Zair różnią się bardziej niż Malezja od klastra składającego się z Meksyku i Wenezueli.



7) Wykres, który pojawia się pod dendrogramem, posiada punkty dla każdego kroku, w którym dwa klastry są połączone w jeden klaster. Współrzędne poziome reprezentują liczbę klastrów i maleją od lewej do prawej. Pionowa współrzędna punktu to odległość między dwoma skupieniami, które są połączone w celu utworzenia określonej liczby skupień. Możesz kliknąć dowolny romb na dendrogramie i przeciągnąć linię, aby wybrać liczbę klastrów, które najlepiej reprezentują dane. Możesz także użyć opcji Liczba klastrów w menu czerwonego trójkąta, aby wybrać liczbę klastrów. Wykres odległości ma zauważalną zmianę nachylenia w czterech klastrach. Zmiana nachylenia wskazuje, że różnice w klastrach, które są połączone do punktu, w którym pozostają cztery klastry, są stosunkowo małe. Sugeruje to, że cztery to dobry wybór dla liczby klastrów. Zwróć uwagę, że jest to liczba klastrów pokazywana domyślnie.

8) Kliknij czerwony trójkąt obok **Hierarchical Clustering** i wybierz **Constellation Plot**

9) Ten wykres konstelacji organizuje kraje jako punkty końcowe, a każdy klaster łączy się jako nowy punkt. Linie reprezentują członkostwo w klastrze. Długość linii między łączeniami klastrów jest zbliżona do odległości między klastrami, które zostały połączone. Wykres konstelacji wskazuje, że klaster zawierający Afganistan i Zair jest mniej więcej tak odległy od klastra pozostałych krajów, jak dwa skupienia składające się z pozostałych krajów w górnej połowie wykresu i tych w dolnej połowie wykresu.



Klasteryzacja algorytmem centroidów

1) W tym ćwiczeniu wykorzystasz tabelę przykładowych danych *Cytometry.jmp* do grupowania obserwacji za pomocą **K Means Cluster**. Cytometria służy do wykrywania markerów na powierzchni komórek, a odczyty z tych markerów pomagają w diagnozowaniu niektórych chorób. W tym przykładzie obserwacje są pogrupowane na podstawie odczytów czterech markerów w analizie cytometrycznej.

2) Wybierz **Help** > **Sample Data Library** i *otwórz Cytometry.jmp*.

3) Wybierz **Analyze** > **Clustering** > **K Means Cluster**.

4) Select CD3, CD8, CD4, oraz MCB następnie kliknij **Y, Columns**. Click **OK**.

5) Wpisz 3 obok pola **Number of Clusters**

6) Wpisz 15 obok **Range of Clusters**. Ponieważ **Range of Clusters** jest ustawiony na 15, platforma zapewnia dopasowanie od 3 do 15 klastrów. Następnie możesz określić preferowaną liczbę klastrów. Kliknij **Go**.

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	23.1784	
K-Means Clustering	4	8.80709	
K-Means Clustering	5	29.5123	
K-Means Clustering	6	52.5517	
K-Means Clustering	7	49.5876	
K-Means Clustering	8	56.5308	
K-Means Clustering	9	54.053	
K-Means Clustering	10	69.8707	
K-Means Clustering	11	70.5239	Optimal CCC
K-Means Clustering	12	61.5326	
K-Means Clustering	13	68.1277	
K-Means Clustering	14	66.4044	
K-Means Clustering	15	69.9928	

7) Raport **Cluster Comparison** jest wyświetlany u góry okna raportu. Najlepsze dopasowanie określa najwyższa wartość **CCC**. W tym przypadku najlepsze dopasowanie występuje, gdy dopasujesz 11 klastrów.

8) Przewiń do raportu **K Means NCluster**=11. Raport **Cluster Summary** pokazuje liczbę obserwacji w każdym z jedenastu skupień. Raport **Cluster Means** pokazuje średnie z czterech odczytów znaczników dla każdego skupienia.

K Means NCluster=11

Columns Scaled Individually

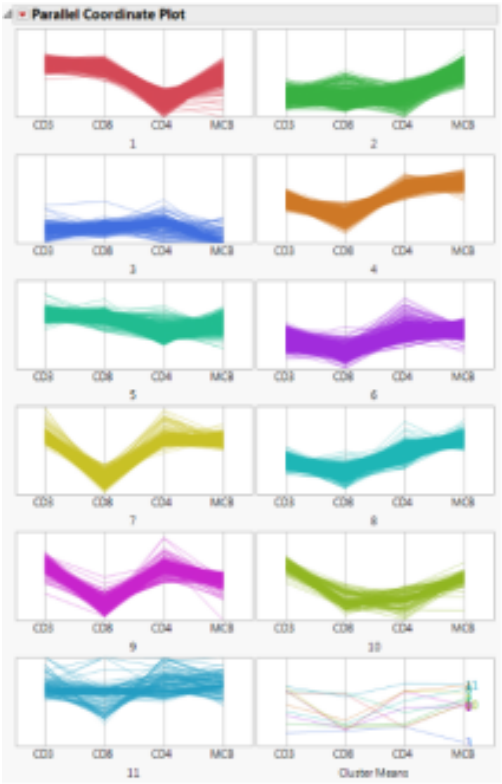
Cluster Summary

Cluster	Count	Step	Criterion
1	816	24	0
2	482		
3	157		
4	577		
5	856		
6	498		
7	447		
8	549		
9	377		
10	123		
11	118		

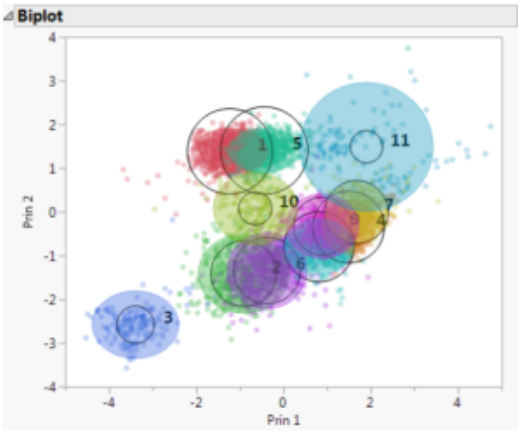
Cluster Means

Cluster	CD3	CD8	CD4	MCB
1	314.073529	300.270833	106.615196	183.4375
2	140.091286	116.365145	127.024896	205.014523
3	90.7898089	87.5477707	109.356688	15.0191083
4	247.426343	148.064125	314.074523	261.831889
5	320.287383	307.372664	193.117991	193.174065
6	188.686747	99.0863454	220.062249	171.682731
7	347.496644	89.9574944	320.782998	231.557047
8	210.258652	126.125683	260.327869	245.588342
9	328.206897	89.7824934	312.909814	175.965517
10	322.105691	113.715447	116.666667	181.731707
11	349.864407	284.813559	360.932203	267.474576

9) Kliknij czerwony trójkąt przy **K Means NCluster=11** i wybierz **Parallel Coord Plots**. Równoległe wykresy współrzędnych przedstawiają strukturę obserwacji w każdym skupieniu. Użyj tych wykresów, aby zobaczyć, jak różnią się klastry. Klastry 4, 6, 7, 8 i 9 mają zwykle stosunkowo niskie wartości CD8 i wysokie wartości CD4. Z drugiej strony klaster 1 ma wyższe wartości CD8 i niższe wartości CD4.



10) Kliknij czerwony trójkąt **K Means NCluster=11** i wybierz **Biplot**. Klastry, które wydają się najbardziej oddzielone od innych w oparciu o ich pierwsze dwie główne składowe, to klastry 3, 10 i 11. Potwierdzają to ich równoległe wykresy współrzędnych, które różnią się od wykresów dla innych klastrów.



III. Pytania podsumowujące

i. Czy do klasteryzacji metodą hierarchiczną można wykorzystać dane znakowe?

ii. Dlaczego klasteryzacja metodą centroidów wymaga danych numerycznych?

iii. Czy do badania odległości pomiędzy punktami w metodzie centroidów można użyć metryki euklidesowej?

iv. Czym jest standaryzacja i w jaki sposób pomaga w klasteryzacji?

v. Jakie jest zadanie kryterium CCC?

vi. Czym jest dendrogram?

vii. Do czego stosujemy wykres konstelacji?

viii. Czym się różni klasteryzacja od klasyfikacji?