KLASTERYZACJA

Paweł Buglewicz, Krzysztof Cieśla, Justyna Olczak

4 lutego 2017

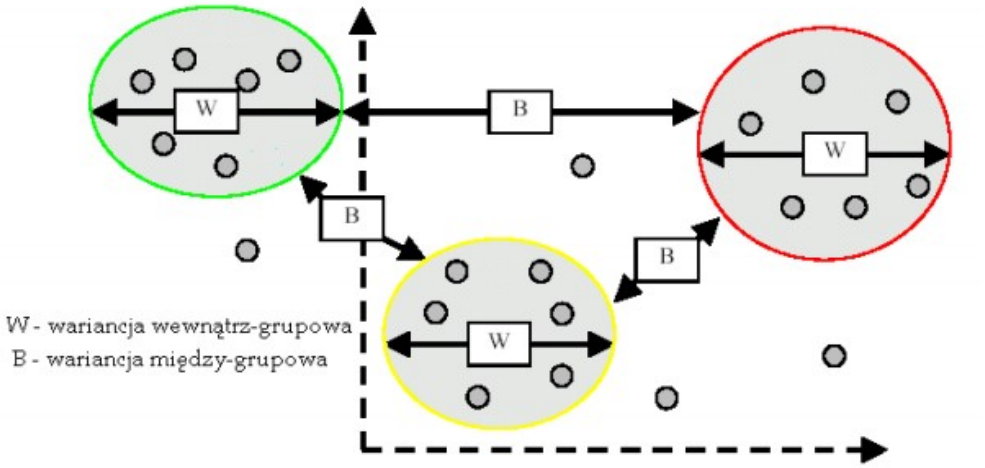
Spis treści

1 Wstęp

Niejednokrotnie może się zdarzyć, że dane pomiarowe wydają się chaotyczne i nieskorelowane. Przedstawienie ich cech na wykresie trójwymiarowym (lub w większej ilości wymiarów) ukazuje jednak, że te dane są powią-

zane. W tej przestrzeni fazowej są one zgrupowane w struktury nazywane klastrami, które czasami można zauważyć gołym okiem.

Gdy złożony zestaw danych rozbije się na mniejsze grupy (co właśnie robi klasteryzacja), opracowanie tych danych, w tym stworzenie modelu opisującego zebrane dane, staje się znacznie prostszym zadaniem. Równocześnie znacznie prostsze staje się zredukowanie ilości wymiarów danego zagadnienia.



Rysunek 1: Idea klasteryzacji. Dążymy do takiego podziału zbioru danych aby wariancja wewnątrz-grupowa (w każdym z klastrów była możliwie mała) a jednocześnie wariancja między-grupowa możliwie duża.

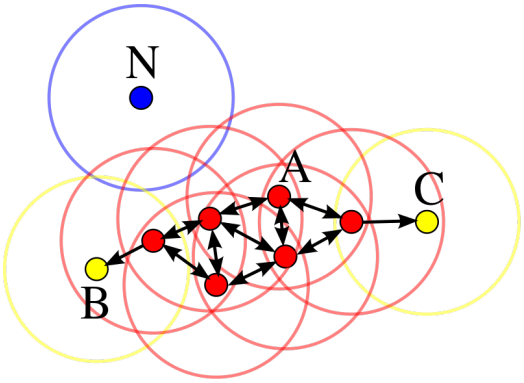
Podział zestawu danych odbywa się na podstawie kryterium podobieństwa, które w przestrzeni wielowymiarowej, można traktować jako odległość (różnie zdefiniowaną dla różnych algorytmów). Funkcja klasteryzująca dąży do zminimalizowania wariancji wewnętrznej danej grupy i maksymalizacji wariancji międzygrupowej, co oznacza, że w danym klastrze znajdują się obiekty jak najbardziej podobne do siebie i jednocześnie klastry różnią się między sobą w maksymalnym stopniu (rys. 1).

2 Algorytmy klasteryzacji

Istnieje wiele powszechnie stosowanych algorytmów klasteryzacji [1]. W naszym projekcie wypróbowaliśmy kilka z nich. Wszystkie z nich to algorytmy, które umożliwiają klasteryzację, gdy nie zna się ilości klastrów.

2.1 DBSCAN

Density-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise - Martin Ester, Hans-Peter Kriegel, Jörg Sander i Xiaowei Xu (1996). DBSCAN jest jedną z najprostszych, najpowszechniejszych oraz najszybszych metod klaseryzacji. Polega ona na szukaniu klastrów jako obszarów o zwiększonej gęstości, oddzielonch obszarami o mniejszej gęstości.



Rysunek 2: Algorytm DBSCAN. Punkty czerwone (rdzeń klastra) należą do klastra oraz spełniają warunek na liczbę sąsiadów. Punkty żółte (krawęź klastra) nie spełniają tego warunku, ale znajdują się w promieniach sąsiedźtwa punktów czerowonych. Punkty niebieskie (szum) nie należą do żadnego z kalstrów.

Zalety:

- Jako parametru początkowego nie wymaga liczby spodziewanych klastrów w danych.
- Wymaga tylko dwóch parametrów początkowych:
  - Minimalna liczba punktów w sąsiedźtwie
  - Maksymalny promień sąsiedźtwa
- Potrafi wyszukać klastry nawet o bardzo skomplikowanym kształcie.

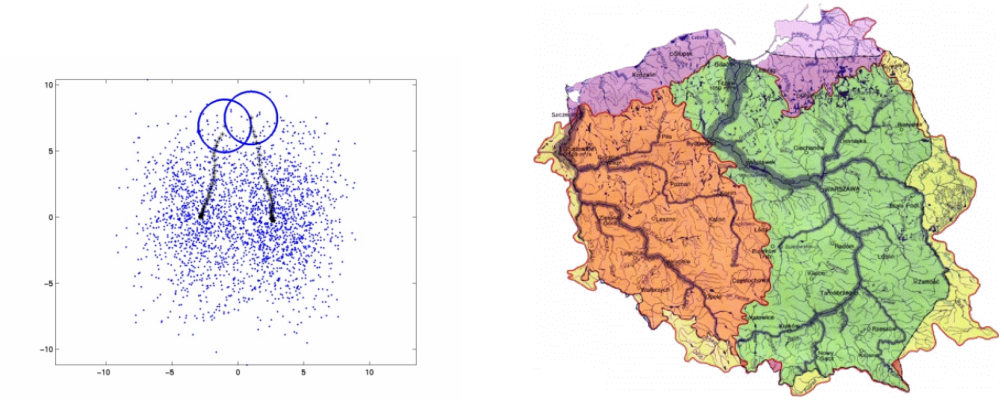
Wady:

- Mało skuteczny dla danych wielowymiarowych.
- Nie działa poprawnie, jeśli różnice w gęstości są zbyt duże.

2.2 Mean Shift

Algorytm poszukuje klastrów, których centra znajdują się w atraktorach danego układu. Atraktorem w tym wypadku nazywamy punkt, do którego docieramy, podążając od każdego punktu w kierunku, w którym gradient jest największy. Ten krok nazywamy *Średnim przesunięciem - mean shift.*Dla danych dyskretnych musi być stosowany odpowiedni estymator gradientu.

Przykład wizualizacji działania tego algorytmu przedstawiony jest na rys. 3a. Bardziej rzeczywistym przykładem może być mapa zlewisk rzek (rys. 3b)



(a) Przykład działania alorytmu Mean Shift. Dwa względnie bliskie sobie punkty zmierzają do dwóch różnych atraktorów.

(b) Mapa zlewisk rzek w Polsce. Atrakторami są tam ich ujścia.

Rysunek 3

Zalety:

- Można zastosować go do dowolnej liczby wymiarów.
- Kształt klastra może być dowolny.

Wady:

- Czuły na wybór parametrów początkowych.
- Przy dużej liczbie próbek algorytm znacznie zwalnia - złożoność $O(n^2)$.

2.3 Affinity propagation

Metoda affinity propagation szuka klastrów poprzez wymianę informacji między punktami. Algorytm określa podobieństwo między parami punktów

(poprzez to, jaką informację wymieniły), jednocześnie je modyfikując (zmienia się odległość tych punktów w przestrzeni wielowymiarowej). Po wielu krokach, jeśli punkty są odpowiednio podobne, czyli odpowiednio bliskie, stają się jednym klastrem. Osobne klastry nie są podobne do siebie i się "odpychają". Cechy punktów są opisywane za pomocą macierzy – wartości liczbowych.

Każdy z klastrów można określić za pomocą kilku parametrów, które opisują jego "jakość". My zwróciliśmy uwagę na następujące parametry:

- Homogeneity - h - każdy klaster zawiera tylko dane z jednego zestawu
- Completness - c - wszystkie dane z jednego zestawu znajdują się w tej samej klasie
- V-measure - v - średnia harmoniczna h oraz c

Warunkowa entropia klas przy chwilowym (w danym kroku algorytmu) przydzieleniu do klastrów:

$$H\left(C|K\right) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n} \cdot log\left(\frac{n_{c,k}}{n_k}\right) \tag{1}$$

Entropia klas:

$$H\left(C\right) = -\sum_{c=1}^{|C|}\frac{n_c}{n} \cdot log\left(\frac{n_c}{n}\right) \tag{2}$$

$$h = 1 - \frac{H\left(C|K\right)}{H\left(C\right)} \qquad c = 1 - \frac{H\left(K|C\right)}{H\left(K\right)} \qquad v = 2 \cdot \frac{h \cdot c}{h + c} \tag{3}$$

Zalety:

- Jako parametr początkowy nie jest potrzebne zadanie liczby klastrów.
- Może być stosowany w dowolnej liczbie wymiarów.
- Stabilny.

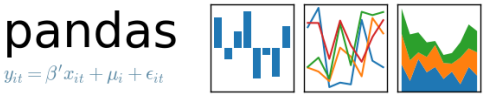
Wady:

- Wrażliwy na szumy.
- Trudny wybór parametrów.
- Powolny - złożoność $O(N^2)$.

3 Wykorzystane oprogramowanie

Używanymi przez nas narzędziami były biblioteki Pythona [2]:

- pandas - umożliwiająca łatwe zarządzanie zebranymi danymi, ich filtrowanie i szeroko pojętą wstępną analizę. [3]



- scikit-learn - wykorzystywany głównie w zagadnieniach machine learnig. Zawiera m.in. wykorzystywane przez nas algorytmy klasteryzacji. [4]

4 Dane

Algorytmy klasteryzacji najlepiej działają dla dosyć dużych zbiorów danych, wielowymiarowych, czyli badających wiele aspektów danego zagadnienia.

W naszym projekcie postanowiliśmy przeprowadzić klasteryzację na zbiorze danych, dostępnych w internecie, opisującym wiele aspektów życia portugalskiej młodzieży[1]. Te dane zostały uzupełnione o odpowiedzi, które uzyskaliśmy przez dodatkową, osobną ankietę, zawierającą analogiczny do oryginalnego badania, zestaw pytań[2].

[1] `https://www.kaggle.com/uciml/student-alcohol-consumption`
[2] `https://goo.gl/forms/aDAEtAKZTxXPf5j32`

Badanych poproszono o udzielenie odpowiedzi na następujące pytania:

- Uczelnia - skrót (drukowane litery np. AGH, PK, UEK itp.)
- płeć
- wiek
- miejsce zamieszkania (miasto, wieś)
- Liczba członków w Twojej rodzinie?
- Czy Twoi rodzice mieszkają razem?
- Wykształcenie matki (podstawowe, średnie zawodowe, średnie ogólnokształcące, policealne, wyższe)
- Wykształcenie ojca (podstawowe, średnie zawodowe, średnie ogólnokształcące, policealne, wyższe)
- W której z podanych branży pracuje Twoja matka? (szkolnictwo, służba zdrowia, służba cywilna, inna)
- W której z podanych branży pracuje Twój ojciec? (szkolnictwo, służba zdrowia, służba cywilna, inna)
- Powód wybrania uczleni (blisko miejsca zamieszkania, reputacja, ciekawy kierunek, inny)
- Czas dojazdu na uczelnię (<15 min, 15-30 min, 30-60 min, >60 min)
- Tygodniowy czas poświęcony na naukę (<2 h, 2-5 h, 5-10 h, >10 h)
- Łączna liczba niezdanych przedmiotów
- Czy uczęszczasz na dodatkowe zajęcia edukacyjne (poza uczelnią)?
- Czy ktoś z Twojej rodziny uczęszcza na dodatkowe zajęcia edukacyjne (poza uczelnią/szkołą)?
- Czy należysz do organizacji uczelnianych (koła naukowe, AZS, samorząd itp.)?
- Czy w domu masz dostęp do internetu?
- Czy jesteś w związku?

- Jak oceniasz swoje relacje z rodziną? (1: b. złe - 5: b. dobre)
- Ile masz wolnego czasu po zajęciach? (1: b. mało - 5: b. dużo)
- Jak często spotykasz się ze znajomymi (poza uczelnią)? (1: b. rzadko - 5: b. często)
- Konsumpcja alkoholu w dzień powszedni. (1: wcale - 5: b. dużo)
- Konsumpcja alkoholu w weekend. (1: wcale - 5: b. dużo)
- Jak oceniasz swoje zdrowie? (1: b. źle - 5: b. dobrze)
- Liczba nieuzasadnionych obecności.
- Średnia ocen z ostatniego semestru.
- Średnia ocen z ostatniego roku.

W danych dostępnych w internecie było około 700 odpowiedzi, na naszą ankietę odpowiedziało kolejnych 600 osób, co razem daje znaczącą próbkę.

Odpowiedzi na poszczególne pytania prezentują się następująco:

- automat dodarłszy na planetę sprawdza, czy nadaje się ona do zamieszkania przez ludzi,
- jeśli nie, używa zasobów planety, aby wytworzyć jak największą liczbę kopii samego siebie,
- następnie kopie te wysyłane są w kierunku najbliższych planet, gdzie procedura jest powtarzana.

5 Analiza

Uzyskane dane są obiektem 28-wymiarowym (na tyle pytań odpowiadali ankietowani). Chociaż wiele algorytmów klasteryzacji pozwala na przeprowadzenie grupowania na tak skomplikowanych obiektach, to przedstawienie wyników, wraz ze wzrostem liczby wymiarów, staje się dużo bardziej skomplikowane.

Z tego też powodu skupiliśmy się na trójwymiarowych klastrach, tworzonych dla każdej kombinacji wymiarów. Innymi słowy, powinniśmy oglądać powiązania między różnymi badanymi aspektami życia ankietowanych.

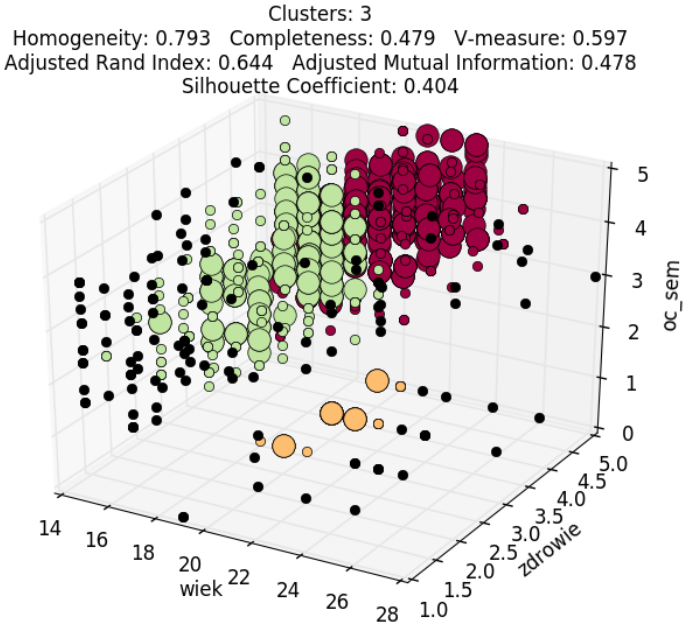
Podczas "analizy" danych napotkaliśmy problemy, które ostudziły nasz zapał badawczy których nie byliśmy w stanie rozwiązać w tak krótkim czasie.

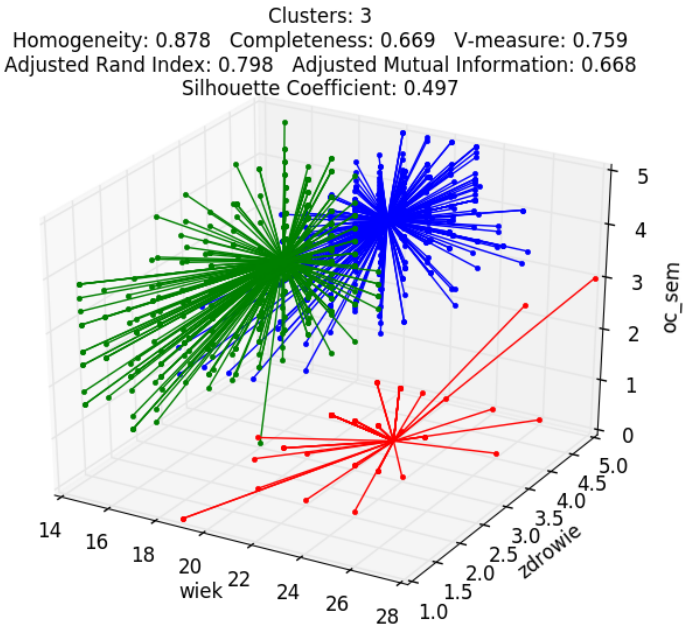
Jednym z największych problemów był dyskretny charakter zebranych przez nas danych. Odpowiedzi na większość pytań są liczbami całkowitymi z zakresu 1-5 lub binarnymi. W przestrzeni trójwymiarowej powoduje to powstawanie równomiernie rozłożonej siatki, o równoodległych punktach. Przeprowadzenie klasteryzacji na takim zbiorze danych jest bardzo trudne, szczególnie gdy używa się tylko podstawowych i szeroko dostępnych narzędzi. Rozwiązaniem tego problemu byłoby zezwolenie na odpowiedzi z szerszego lub "gęstszego" zakresu (np. zmiennego co 0.5 lub 0.25).

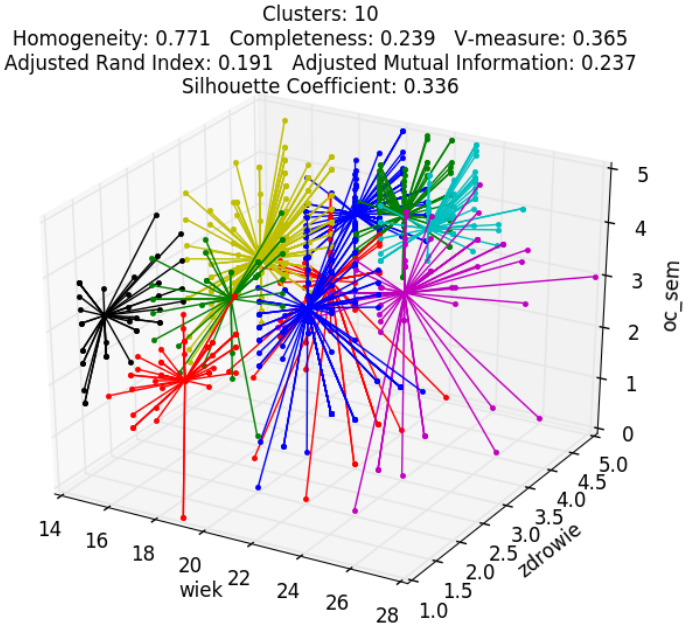
5.1 Zgodność danych

W oryginalnym badaniu ankietowani to młodzież w wieku okołolicealnym, a my przeprowadzaliśmy ankietę wśród studentów. W wielu pytaniach odpowiedzi trzeba by interpretować inaczej. Wśród takich pytań są na przykład: oceny, czas dojazdu do szkoły/na uczelnię, ilość spożywanego alkoholu, czas wolny itd. W Portugalii używany jest 20-stopniowy system ocen. Proste dzielenie, aby sprowadzić go do tego używanego w polsce powoduje powstawanie fałszywych klastrów. Ten podział ze względu na kraj powinien być widoczny w danych, jednakże dało się go zauważyć tylko w niewielu przypadkach. Nie jest również jasne, czy oba zbiory danych można tak bezkrytycznie łączyć i poddawać klasteryzacji.

5.2 Przykładowe klastry







6 Podsumowanie

Chociaż klasteryzacja jest metodą przeznaczoną specjalnie dla danych wielowymiarowych, to jednak korzystanie z niej wymaga pewnej wprawy. Być może problem 28-wymiarowy, dyskretny, gdy nie znamy i nie potrafimy przewidzieć liczby klastrów, to zbyt skomplikowany problem jak na pierwsze ćwiczenie w tym zakresie.

Spośród kilku tysięcy stworzonych przez nas wykresów, bardzo ciężko zauważyć zależności, chociażby przez samą ilość wyników. Jest to główny powód, dla którego nie zaczęliśmy dalszej analizy zebranych danych.

Literatura

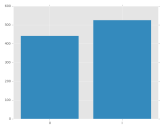
[1] *http://scikit-learn.org/stable/modules/clustering.html*

[2] *http://www.python.org/*

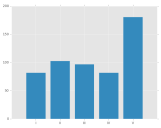
[3] *http://pandas.pydata.org/*

[4] *http://scikit-learn.org/stable/*

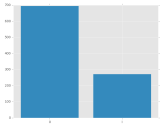
7 Histogramy



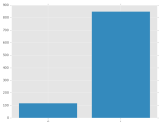
Rysunek 4: Płeć



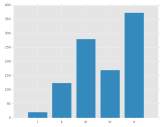
Rysunek 5: Wiek



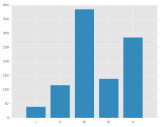
Rysunek 6: Miejsce zamieszkania



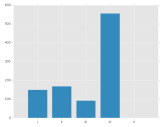
Rysunek 7: rodzice razem



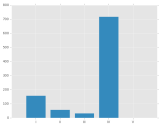
Rysunek 8: Wykształcenie matki



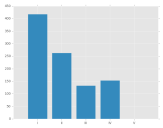
Rysunek 9: Wykształcenie ojca



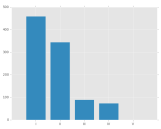
Rysunek 10: Branża m



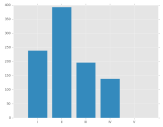
Rysunek 11: Branża o



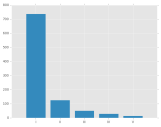
Rysunek 12: Powód wybrania uczelni



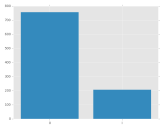
Rysunek 13: Czas dojazdu



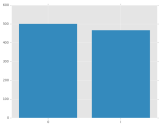
Rysunek 14: Czas poświęcony na naukę



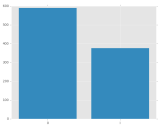
Rysunek 15: Liczba niezdanych przedmiotów



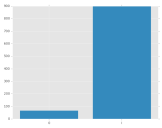
Rysunek 16: Dodatkowe zajęcia



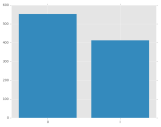
Rysunek 17: Dodatkowe zajęcia (Rodzina)



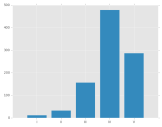
Rysunek 18: Organizacje



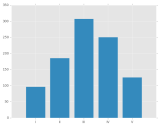
Rysunek 19: Internet



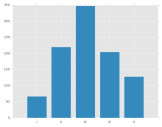
Rysunek 20: Związek



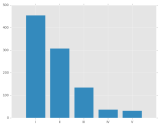
Rysunek 21: Relacje z rodziną



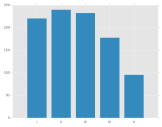
Rysunek 22: Czas wolny



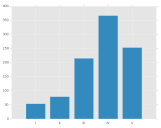
Rysunek 23: Czas - znajomi



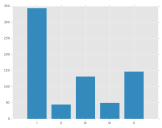
Rysunek 24: Alkohol dzień



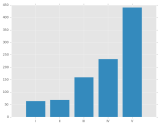
Rysunek 25: Alkohol weekend



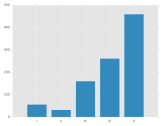
Rysunek 26: Zdrowie



Rysunek 27: Godziny??



Rysunek 28: Oceny semestr



Rysunek 29: Oceny rok