Rok akademicki 2012/2013

Politechnika Warszawska

Wydział Elektroniki i Technik Informacyjnych

Instytut Informatyki



praca dyplomowa magisterska

inż. Bartłomiej Jańczak

Tytuł pracy dyplomowej

Opiekun pracy

prof. nzw. dr hab. Marzena Kryszkiewicz

Ocena:

Podpis Przewodniczącego

Komisji Egzaminu Dyplomowego

Kierunek: Informatyka

Specjalność: Inżynieria Systemów Informatycznych

Data urodzenia: 1988.03.06

Data rozpoczęcia studiów: 2007.10.01

Życiorys

Podpis studenta

EGZAMIN DYPLOMOWY

Złożył egzamin dyplomowy w dniu 20\_\_ r

z wynikiem

Ogólny wynik studiów:

Dodatkowe wnioski i uwagi Komisji:

STRESZCZENIE

Streszczenie pracy w języku polskim.

Słowa kluczowe:

THESIS TITLE IN ENGLISH

Summary in English.

Keywords:

Spis treści

[1. Wprowadzenie 1](#_Toc351415218)

[1.1. Przegląd literatury 2](#_Toc351415219)

[1.2. Motywacja i cel pracy 3](#_Toc351415220)

[1.3. Układ pracy 3](#_Toc351415221)

[2. Miary odległości i podobieństwa 4](#_Toc351415222)

[2.1. Metryki odległości 4](#_Toc351415223)

[2.2. Miara odległości kosinusowej (1) 5](#_Toc351415224)

[2.3. Wyznaczanie kosinusowego sąsiedztwa za pomocą sąsiedztwa opartego na odległości euklidesowej 6](#_Toc351415225)

[3. Użyte algorytmy 10](#_Toc351415226)

[3.1. Grupowanie gęstościowe na przykładzie DBSCAN 11](#_Toc351415227)

[3.2. Wyszukiwanie k-najbliższych sąsiadów 15](#_Toc351415228)

[4. Szacowanie odległości 17](#_Toc351415229)

[4.1. Wykorzystanie nierówności trójkąta 17](#_Toc351415230)

[4.2. Wykorzystaniem indeksu metrycznego 20](#_Toc351415231)

# 

# 1. Wprowadzenie

Współczesne systemy komputerowe agregują i generują ogromną ilość danych zawierających cenną dla biznesu trudno odkrywalną wiedzę. Jej znajdowaniem zajmuje się dziedzina informatyki zwana odkrywaniem wiedzy. Mimo, że jest ona stosunkowo młoda, to stworzyła wiele technik eksploracji danych, które dzięki swojej skuteczności oraz wydajności znalazły szerokie praktyczne zastosowanie w rozwiązywaniu problemów związanych z analizą danych.

Zasadniczą przyczyną rychłego rozwoju odkrywania wiedzy jest spopularyzowanie wydajnych metod pozyskiwania i gromadzenia informacji. Zjawisko to nie byłoby możliwe bez postępu technologicznego w dziedzinie urządzeń agregujących dane i systemów bazodanowych oraz dzięki upowszechnieniu urządzeń umożliwiających automatyczną rejestrację sposobu ich wykorzystania. Z punktu widzenia konsumenta można tu wyszczególnić kody kreskowe, karty płatnicze oraz szeroko pojęte urządzenia mobilne. Kolejnym wartym uwagi źródłem danych jest sieć Internet, w której możliwa jest rejestracja wielu czynności korzystających z niej użytkowników, którzy ponadto dobrowolnie umieszczają w niej wiele informacji o sobie, przykładowo na portalach społecznościowych.

Wyżej wymienione zjawiska mają wpływ na osiąganie ogromnych rozmiarów przez współczesne zbiory danych. Tempo ich wzrostu jest szybsze niż przewidywano jeszcze kilka lat temu. Ich gromadzenie i przechowywanie na nośnikach pamięci masowej nie stanowi problemów dla współczesnych systemów, natomiast działanie na takiej ilości danych, pomimo stale wzrastającej mocy obliczeniowej komputerów, wciąż jest wyzwaniem dla dzisiejszej informatyki. Zbiory danych same w sobie nie stanowią wielkiej wartości, jednakże rozsądnie wykorzystane mogą stać się cennym źródłem szczególnej wiedzy. Nierzadko użyteczna wiedza ukryta jest między pewnymi składowymi danych, więc do jej odkrycia konieczne są właściwe algorytmy. Zagadnieniom tym poświęcona jest dziedzina informatyki zwana eksploracją danych, której ideą jest wykorzystanie komputera do znajdowania ukrytych dla człowieka wartościowych prawidłowości w danych zgromadzonych w dużych repozytoriach.

Odkrywanie wiedzy jest procesem złożonym, na który najczęściej składają się następujące etapy:

* analiza danych – poznanie charakteru danych i określenie celu eksploracji,
* selekcja danych – czyszczenie, weryfikacja poprawności i wybór danych, które zostaną poddane dalszej analizie,
* transformacja danych – przekształcenie danych do odpowiedniej postaci, określenie strategii wobec danych niepełnych,
* eksploracja danych – ekstrakcja wiedzy z danych,
* interpretacja wyników – logiczna i graficzna wizualizacja wyników, wybór najbardziej interesującej wiedzy, wnioskowanie.

Kluczową fazą procesu odkrywania wiedzy jest eksploracja danych. Do zasadniczych metod eksploracji danych należą:

* grupowanie,
* klasyfikacja,
* odkrywanie asocjacji,
* regresja.

Każda z metod ujawnia różnego rodzaju korelacje pomiędzy danymi, z czego wynika ich odmienne zastosowanie.

W tej pracy skoncentrowałem się na zagadnieniu grupowania danych, które określane jest jako wyznaczanie zbiorów obiektów podobnych przy zachowaniu właściwości maksymalizacji podobieństwa obiektów należących do tych samych grup i minimalizacji podobieństwa obiektów należących do innych grup.

## 1.1. Przegląd literatury

Grupowanie danych jest popularną metodą o wielu zastosowaniach, dlatego nie trudno o jej opis w literaturze. W przypadku algorytmów, na których skupiłem się w niniejszej pracy wyjątkowo przydatne okazały się artykuły naukowe.

Jednym z najpopularniejszych algorytmów gęstościowego grupowania danych jest DBSCAN [[1](#MEs66)] stanowiący często punkt odniesienia dla porównań z innymi algorytmami gęstościowych grupowań. [TODO]

Nową koncepcją zwiększenia wydajności wyżej wymienionych algorytmów jest wykorzystanie nierówności trójkąta do redukcji liczby kosztownych operacji wyznaczania podobieństwa obiektów. Na przykładzie algorytmu k-środków [[2](#CEl03)] przedstawiane już były próby wykorzystania nierówności trójkąta w algorytmach grupowania danych. Natomiast po raz pierwszy została ona użyta w celu porządkowania dostępu do danych w algorytmach gęstościowego grupowania TI-DBSCAN [[3](#Kry10)] i TI-NBC [[4](#MKr11_2)]. Dokonano również badania wpływu liczby punktów referencyjnych i strategii ich wyboru na efektywność tych algorytmów [Wawer??].

## 1.2. Motywacja i cel pracy

Grupowanie danych to proces powszechnie stosowany w porządkowaniu produktów, segmentacji klientów, organizacji obiektów czy rozpoznawaniu i analizie obrazów. Procesy te wymieniane są pośród kluczowych elementów, na których bazuje szeroko rozumiana sztuczna inteligencja. We współczesnym świecie algorytmy grupowania danych znajdują coraz szersze zastosowanie. Ich popularność rozpala zainteresowanie naukowców, którzy opracowują coraz sprawniejsze algorytmy lub modyfikują istniejące, które dotychczas wydawały się optymalne. Nierzadko zdarza się, że usprawnienia po wielokroć zwiększają wydajność dotychczasowych rozwiązań, co z kolei umożliwia przetwarzanie zbiorów danych z większą liczbą obiektów bądź atrybutów. Niekiedy może to oznaczać sposobność użycia tych algorytmów w nieosiągalnych dotychczas obszarach.

Jednym z najnowszych pomysłów na zwiększenie wydajności algorytmów grupowania danych jest zastosowanie nierówności trójkąta. [TODO]

Celem pracy jest … [TODO]

## 1.3. Układ pracy

[TODO]

# 2. Miary odległości i podobieństwa

Podobieństwo jest pojęciem fundamentalnym w niemal każdej dziedzinie naukowej. Przykładowo, w matematyce, geometryczne metody oceny podobieństwa wykorzystywane są do określania przystawania jak również w dziedzinach pokrewnych takich jak trygonometria. W biologii molekularnej ważnym problemem jest mierzenie podobieństwa par białek. Zbiory rozmyte wykształciły własne miary podobieństwa znajdujące zastosowanie na polach zarządzania, medycyny czy meteorologii. Przegląd wszystkich zastosowań podobieństwa jest samym w sobie wdzięcznym tematem na pracę dyplomową. W niniejszym rozdziale skupię się na wybranych miarach podobieństwa wektorów, tj. metrykach odległości oraz metryce podobieństwa kosinusowego.

## 2.1. Metryki odległości

Metryką odległości (lub krócej odległością) w zbiorze wektorów jest miara podobieństwa , która spełnia następujące warunki dla dowolnych wektorów , oraz w :

1. ;
2. ;
3. ;

Warto zauważyć, że istnieje wiele miar odległości. W zależności od zastosowania, w danym przypadku, jedne miary mogą być stosowniejsze niż inne. Najpopularniejszą miarą odległości jest *odległość Euklidesowa*. Odległość Euklidesowa między punktami i oznaczana jest jako , i definiowana jako:

Gdy stosowana jest odległość Euklidesowa to otoczenie punktu przyjmuje sferyczny kształt.

Innym przykładem popularnej miary odległości jest *odległość Manhattan*. Odległość Manhattan między punktami i oznaczana jest jako , i definiowana w następujący sposób:

Gdy stosowana jest odległość Manhattan to otoczenie punktu przyjmuje prostokątny kształt.

Zarówno odległość Manhattan jak i odległość Euklidesowa są szczególnymi przypadkami *odległości Minkowskiego*. Odległość Minkowskiego rzędu między punktami i oznaczana jest jako , i definiowana jako:

Dla uproszczenia, bez straty ogólności, w dalszych rozważaniach będę posługiwał się odległością euklidesową jako metryką odległości.

W dalszej części rozdziału, na podstawie [[5](#MKr12)], przedstawiono miarę odległości Kosinusowej oraz wyznaczanie kosinusowego sąsiedztwa za pomocą sąsiedztwa opartego na odległości Euklidesowej.

## 2.2. Miara odległości Kosinusowej

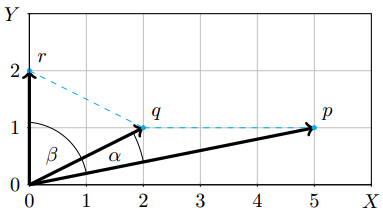
W wielu aplikacjach, w szczególności odkrywających wiedzę w danych tekstowych , *miara podobieństwa kosinusowego* stosowana jest w celu znajdowania obiektów podobnych danemu.

W dalszej części pracy zakładam, że obiekty reprezentowane są przez wektory przestrzeni wymiarowej. Każdy wektor jest rozumiany jako sekwencja komponentów , gdzie komponent jest wartością -tego wymiaru , . Wektor o wszystkich wymiarach równych 0 będzie nazywany *wektorem zerowym*. W inny przypadku, będzie nazywany *wektorem niezerowym*.

Miara podobieństwa kosinusowego między wektorami i oznaczana jest jako i zdefiniowana w następujący sposób:

**Przykład 1**. Na rysunku Rys. 1 przedstawiono trzy wektory , i . Należy zwrócić uwagę, że odległość między i jest większa niż odległość euklidesowa między i . Z drugiej strony, w sensie podobieństwa kosinusowego, jest bardziej podobne do niż , ponieważ kosinus kąta między i () jest większy niż kosinus kąta między i ().

W tabeli Tab. 1 zamieszczono wartości podobieństwa kosinusowego między wektorami , i . Można zauważyć, że dla podobieństwa kosinusowego warunki oraz i są niespełnione.



Rys. 1. Odległość euklidesowa i podobieństwo kosinusowe

Tab. 1. Podobieństwo kosinusowe wektorów z rys. 1.

|  |  |
| --- | --- |
| **(u,v)** | **cosSim(u,v)** |
| (p,q) | 0,965 |
| (p,r) | 0,196 |
| (q,r) | 0,447 |

Z powyższego przykładu płyną następujące wnioski:

1. Nierówność trójkąta nie jest spełniona dla dla dowolnego zbioru wektorów.
2. Nierówność trójkąta nie jest spełniona dla dla dowolnego zbioru wektorów.
3. Nierówność trójkąta nie jest spełniona dla dla dowolnego zbioru wektorów.

Ponieważ podobieństwo kosinusowe między niezerowymi wektorami i opiera się wyłącznie na kącie zawartym między nimi i nie zależy od ich długości, stąd obliczanie może być wyznaczone w oparciu o znormalizowane wektory i , tj. i . Z powyższego spostrzeżenia wynikają następujące własności:

1. ;
2. ;
3. .

## 2.3. Wyznaczanie kosinusowego sąsiedztwa za pomocą sąsiedztwa opartego na odległości euklidesowej

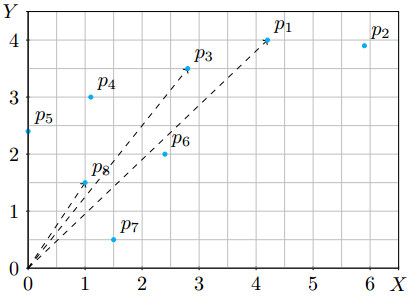
Artykuł [[5](#MKr12)] dowodzi, że kosinusowe sąsiedztwo[[1]](#footnote-2) w zbiorze wektorów może zostać wyznaczone za pomocą odpowiedniego sąsiedztwa opartego na odległości euklidesowej w zbiorze wektorów składającym się z α-znormalizowanych[[2]](#footnote-3) wektorów z . Stąd, autorka artykułu proponuje następujące podejście do wyznaczania sąsiedztwa opartego na podobieństwie kosinusowym.

W pierwszej kolejności początkowy zbiór wektorów transformowany jest do - zbioru α-znormalizowanych wektorów z . Następnie kosinusowe -sąsiedztwo (lub kosinusowe k-sąsiedztwo) w zbiorze ustanawiane jest jako oparte na odległości euklidesowej -sąsiedztwo (lub alternatywnie k-sąsiedztwo) w zbiorze , gdzie . Warto zwrócić uwagę, że w przeciwieństwie do kosinusowego -sąsiedztwa, -sąsiedztwo spełnia własność nierówności trójkąta.

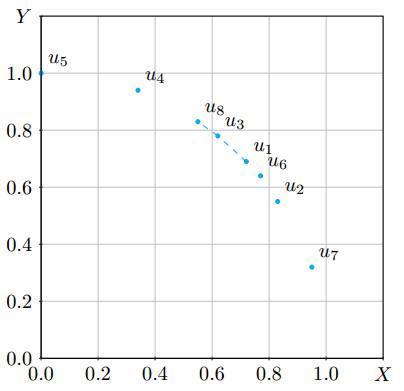
**Przykład 2.** W Tab. 2 zdefiniowano a na Rys. 2 przedstawiono przykładowy zbiór . Na Rys. 3 zamieszczono zbiór α-znormalizowanych wektorów z , gdzie . Warto zauważyć, że znormalizowany wektory zbioru mają długość równa , a punkty opisywane wektorami zbioru D’ układają się na okręgu o środku w punkcie i promieniu .

Tab. 2. Przykładowy zbiór wektorów

|  |  |  |
| --- | --- | --- |
| **Wektor** |  |  |
|  | 4.20 | 4.00 |
|  | 5.90 | 3.90 |
|  | 2.80 | 3.50 |
|  | 1.10 | 3.00 |
|  | 0.00 | 2.40 |
|  | 2.40 | 2.00 |
|  | 1.50 | 0.50 |
|  | 1.00 | 1.50 |



Rys. 2. Przykładowy zbiór wektorów



Rys. 3. Zbiór wektorów α-znormalizowanych

Załóżmy, że chcemy znaleźć wektory najbardziej podobne do wektora względem kosinusowego podobieństwa. Z Rys. 2 widać, że są nimi wektory i . Analogicznie przedstawia się relacja między α-znormalizowanymi wektorami , i , tj. , i . Najbardziej podobnymi wektorami do względem kosinusowego podobieństwa są wektory i . Co więcej wektory i są również najbardziej podobne do względem podobieństwa opartego na odległości euklidesowej, co zaznaczono na Rys. 3 błękitną przerywaną linią.

# 3. Użyte algorytmy

Istnieje wiele rozwiązań problemu grupowania danych czyli wyznaczania zbiorów obiektów podobnych przy zachowaniu właściwości maksymalizacji podobieństwa obiektów należących do tych samych grup i minimalizacji podobieństwa obiektów z różnych grup. Popularnym przykładem miary podobieństwa jest odległość euklidesowa klasyfikująca obiekty leżące blisko siebie jako podobne, jednak większość algorytmów jest niezależna od przyjętej miary podobieństwa. Mnogość zastosowań grupowania częstokroć o odmiennych wymaganiach co do rezultatu oraz specyficznych danych wejściowych (np. o różnej liczności, rozkładzie bądź liczbie atrybutów) prowadzi do dużej liczby wyspecjalizowanych algorytmów. W każdym z nich można doszukać się wad oraz zalet, jednakże nie znaleziono dotychczas uniwersalnego algorytmu. Często trudno porównywać algorytmy grupowania danych ponieważ ze względu na charakterystyczne podejście do rozwiązywanego problemu różnią się one nie tylko sposobem grupowania ale także definicją grupy.

Najpopularniejsza klasyfikacja algorytmów grupowania dzieli je na algorytmy oparte na podziale i algorytmy hierarchiczne. W przypadku pierwszej klasy kluczowym elementem jest znalezienie najlepszego podziału zbioru na z góry zadaną liczbę możliwie najbardziej jednorodnych grup. Początkowy podział odpowiednio ze zdeterminowaną strategią optymalizowany jest w kolejnych iteracjach zgodnie z przyjętą funkcją celu. Przykładami metod podziału są algorytmy k-średnich i k-medoidów. Wynikiem drugiej klasy algorytmów grupowania jest dendrogram[[3]](#footnote-4) - drzewo, które iteracyjnie dzieli zbiór danych na coraz to mniejsze podzbiory dopóki każdy podzbiór składa się z jednego obiektu. W takiej hierarchii każdy węzeł drzewa reprezentuje klaster zbioru danych. Relacja między węzłami a ich przodkami w dendrogramie odpowiada relacji między podgrupami a grupami. Dendrogramy mogą być tworzone od liści w górę do korzenia (*podejście aglomeracyjne*) lub od korzenia w dół do liści (*podejście podziału*) poprzez scalanie lub podział klastrów z każdym krokiem algorytmu. Obie wymienione klasy algorytmów grupowania posiadają pewne wady. W przeciwieństwie do algorytmów opartych na podziale algorytmy hierarchiczne nie oczekują arbitralnie zadanej liczby klastrów, jednakże wymagają zdefiniowania *warunku zakończenia* wskazującego kiedy proces podziału lub scalania powinien się zakończyć.

Wyniki wyżej wymienionych metod rzadko odpowiadają oczekiwaniom. Taki stan rzeczy można tłumaczyć nienaturalnym dla człowieka mechanizmem grupowania. Gdyby zadać człowiekowi zadanie pogrupowania punktów dwuwymiarowej przestrzeni okazałoby się, że nie dzieliłby on zbioru hierarchicznie na kolejne podzbiory czy też nie próbowałby podzielić go na z góry określoną liczbę podzbiorów. Ludzie z łatwością rozpoznają klastry o dowolnych kształtach oraz szum. Głównym powodem, dla którego rozpoznajemy klastry jest fakt, iż wewnątrz każdego z klastrów można wyszczególnić pewną gęstość punktów znacznie wyższą niż poza klastrem. Zatem do grupy należą punkty leżące w obszarze o gęstości wyraźnie większej niż w obszarze otaczającym ją. Tak zdefiniowanemu pojęciu metody grupowania najbliżej jest algorytmom gęstościowym, których przykładem jest DBSCAN opisany w kolejnym rozdziale.

## 3.1. Grupowanie gęstościowe na przykładzie DBSCAN

DBSCAN czyli Density Based Spatial Clustering of Applications with Noise jest jednym z najpopularniejszych algorytmów gęstościowego grupowania danych. Zaproponowany w 1996 roku wciąż jest sztandarowym algorytmem gęstościowym będącym punktem odniesienia w wielu pracach naukowych dotyczących tematyki grupowania oraz prezentujących nowe rozwiązania lub algorytmy.

Oprócz grup, czyli zbioru punktów o dużej gęstości punktów DBSCAN rozpoznaje również szum, do którego należą punkty leżące w obszarze o małej gęstości. Algorytm wymaga podania jedynie dwóch parametrów wejściowych, które opisują najmniejszy klaster będący obiektem zainteresowania. Jest to promień wokół danego punktuy, wewnątrz którego znajduje się minimalna liczba punktów. Para parametrów i stanowi intuicyjną definicję najmniejszej gęstości tym samym definiując minimalną liczność wykrywanych grup.

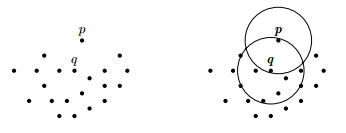
Podstawowym pojęciem używanym w kontekście algorytmu DBSCAN jest *otoczenie epsilonowe* oznaczane przez i definiowane jako zbiór takich punktów zbioru D, które są różne od i nie bardziej odległe od niż , czyli:

W DBSCAN wyróżnia się dwa rodzaje punktów wchodzące w skład klastra: punkty wewnątrz grupy zwane *punktami rdzeniowymi* oraz punkty leżące na obrzeżach klastra – *punkty brzegowe*.

*Punktem rdzeniowym* nazywamy taki punkt p, którego otoczenie epsilonowe zawiera wymaganą liczbę punktów, czyli:

Mówimy, że punkt jest *bezpośrednio gęstościowo osiągalny* z punktu względem i , jeżeli należy do otoczenia epsilonowego p, oraz q jest punktem rdzeniowym:

Relacja *bezpośredniej gęstościowej osiągalności* jest symetryczna tylko dla punktów rdzeniowych. Na Rys. 4 przedstawiono klaster, w którym zaznaczono pewien punkt brzegowy , punkt rdzeniowy , okręgami zaznaczono otoczenie epsilonowe równe a wynosi 5. Rysunek prezentuje asymetryczny przypadek, w którym jest bezpośrednio gęstościowo osiągalny z , natomiast nie jest bezpośrednio gęstościowo osiągalny z .

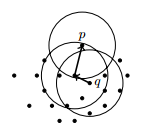


Rys. 4. Przykład ilustrujący punkty rdzeniowe i brzegowe

Mówimy, że punkt jest *gęstościowo osiągalny* z punktu względem i , jeżeli istnieje sekwencja punktów takich, że jest *bezpośrednio gęstościowo osiągalny* z . Relacja ta to kanoniczne rozszerzenie relacji bezpośredniej gęstościowej osiągalności, jest tranzytywna lecz nie jest symetryczna. Z tego powodu została wprowadzona symetryczna relacja gęstościowego połączenia.

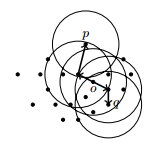
Mówimy, że punkt jest *gęstościowo połączony* z punktem względem i , jeżeli istnieje punkt taki, że i są gęstościowo osiągalne z względem i .

Na Rys. 5 przedstawiono klaster, w którym zaznaczono pewien punkt brzegowy , punkt rdzeniowy , okręgami zaznaczono otoczenia epsilonowe Eps pewnych punktów a wynosi 5. Analiza rysunku pozwala zauważyć, że punkt jest gęstościowo osiągalny z , natomiast jest gęstościowo osiągalny z .



Rys. 5. Przykład ilustrujący relację gęstościowej osiągalności

Na Rys. 6 przedstawiono klaster, w którym zaznaczono punkty brzegowe i oraz punkt . Okręgami wyznaczono otoczenia epsilonowe punktów, a wynosi 5. Studium rysunku pozwala spostrzec, że punkty i są gęstościowo osiągalne z , czli punkty i są gęstościowo połączone.



Rys. 6. Przykład ilustrujący relację gęstościowej łączności

Wszystkie terminy niezbędne do przedstawienia gęstościowego pojęcia grupy zostały już wprowadzone. Niech D jest zbiorem punktów. Grupą G względem i nazywamy niepusty zbiór D spełniający następujące warunki:

1. : jeśli i jest gęstościowo osiągalne z względem i , wtedy .
2. : jest gęstościowo połączone z względem i .

Niech będą grupami zbioru punków D względem i , . *Szumem* nazywamy podzbiór punktów zbioru D nie należących do żadnej z grup czyli:

Algorytm DBSCAN iteruje wejściowy zbiór punktów i uruchamia procedurę wyznaczania nowej grupy *ExpandCluster* dla każdego punktu, który nie został jeszcze przypisany do którejś z grup lub zidentyfikowany jako szum. *ExpandCluster* w pierwszej kolejności wyznacza otoczenie epsilonowe danego punktu i buduje nową grupę jeśli ów punkt jest punktem rdzeniowym, w przeciwnym przypadku oznacza go jako szum. Proces tworzenia nowej grupy rozpoczyna się od dodania do niej punktów należących do otoczenia epsilonowego danego punktu. Następnie wszystkie punkty epsilonowego sąsiedztwa[[4]](#footnote-5) dodawane są do zbioru ziaren *seeds* zawierającego punkty, które potencjalnie mogą rozszerzyć budowaną grupę. Algorytm iteruje zbiór seeds wyznaczając epsilonowe otoczenie dla każdego jego punktu. Jeżeli dany punkt jest punktem rdzeniowym, to wszystkie punkty należące do jego otoczenia epsilonowego, które nie mają przypisanej żadnej grupy również dodawane są do nowoutworzonej grupy. Te z nich, które nie są oznaczone jako szum dodawane są do zbioru *seeds*, ponieważ mogą rozszerzyć tworzoną grupę. Na wydruku 1 wyżej opisany algorytm został zapisany w formie pseudokodu.

Wydruk 1

**DBSCAN**(SetOfPoints, Eps, MinPts)

// SetOfPoints is UNCLASSIFIED

ClusterId := nextId(NOISE);

FOR i FROM 1 TO SetOfPoints.size DO

Point := SetOfPoints.get(i);

IF Point.ClId = UNCLASSIFIED THEN

IF ExpandCluster(SetOfPoints, Point, ClusterId, Eps, MinPts) THEN

ClusterId := nextId(ClusterId);

END IF;

END IF;

END FOR;

END; //DBSCAN

**ExpandCluster**(SetOfPoints, Point, ClId, Eps, MinPts) : Boolean

seeds := SetOfPoints.regionQuery(Point, Eps);

IF seeds.size < MinPts THEN // no core point

SetOfPoints.changeClId(Point, NOISE);

RETURN false;

ELSE // all points in seeds are density reachable from Point

SetOfPoints.changeClId(seeds, ClId);

seeds.delete(Point);

WHILE seeds <> Empty DO

currentP := seeds.first();

result := SetOfPoints.regionQuery(currentP, Eps);

IF result.size >= MinPts THEN

FOR I FROM 1 TO result.size DO

resultP := result.get(i);

IF resultP.ClId IN (UNCLASSIFIED, NOISE) THEN

IF resultP.ClId = UNCLASSIFIED THEN

seeds.append(resultP);

END IF;

SetOfPoints.changeClId(resultP, ClId);

END IF;

END FOR;

END IF;

seeds.delete(currentP);

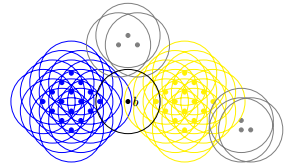
END WHILE;

RETURN true;

END IF;

END; //ExpandCluster

Dokładna analiza kodu pozwala zauważyć, że algorytm DBSCAN jest deterministyczny z dokładnością do punktów brzegowych. Nie bierze on pod uwagę, że punkty brzegowe znajdujące się między leżącymi blisko siebie grupami mogą należeć do więcej niż jednej z grup. Taka sytuacja została przedstawiona na Rys. 7.



Rys. 7. Ilustracja sytuacji, w której przynależność do jednej z grup (niebieskiej bądź żółtej) punktu brzegowego b zależy od kolejności w jakiej DBSCAN będzie badał punkty

Wynik wykonania algorytmu DBSCAN zależy od kolejności przeglądania punktów, ponieważ punkt brzegowy zakwalifikowany do pewnej grupy, w rezultacie rozbudowy kolejnych grup, może zostać przypisany do innych grup. Problem ten może zostać rozwiązany poprzez przechowywanie w każdym punkcie zamiast jednego identyfikatora *clusteId* zbioru identyfikatorów. Jednakże podobnie do autorów algorytmu, problem ten uznaję za pomijalny.

## 3.2. Wyszukiwanie k-najbliższych sąsiadów

Poszukiwanie k najbliższych sąsiadów jest zagadnieniem optymalizacyjnym znajdowania najbliższych punktów w przestrzeni metrycznej. Problem ten definiowany jest w następujący sposób.

Niech będzie punktem zbioru , a odległość między punktami i będzie wyrażana jako . Zbiór wszystkich punktów w , które są różne od i bliższe niż będzie oznaczany przez ; czyli:

K sąsiedztwo punktu , oznaczane przez , jest definiowane jako zbiór wszystkich punktów w D, gdzie , takich, że liczba punktów różnych od i bliższych niż jest mniejsza niż ; czyli:

W większości przypadków k sąsiedztwo wyznaczane jest w n wymiarowej przestrzeni euklidesowej a odległość mierzona jest odległością euklidesową lub odległością manhattan.

Problem wyszukiwania najbliższych sąsiadów pojawia się na wielu polach, wśród których znajdują się:

* rozpoznawanie wzorców,
* sekwencjonowanie DNA,
* systemy rekomendacji,
* analiza skupień.

Istnieje niemało metod rozwiązań problemu k najbliższych sąsiadów. Użyteczność oraz jakość tych algorytmów determinowana jest przez złożoność czasową zapytań jak również koszt utrzymania potrzebnych struktur danych. Najprostszym z nich jest obliczanie odległości punktu zapytania do wszystkich punktów zbioru , śledząc dotychczasowo najlepszych punktów.

W niniejszej pracy problem k sąsiedztwa rozpatrywany jest w kontekście analizy skupień.

# 4. Szacowanie odległości

Szacowanie odległości jest przybliżonym określaniem jej wartości. Działanie to pozwala uniknąć wielokrotnego jej obliczania między pewnym wektorem a wszystkimi wektorami danego zbioru wektorów . W następujących podrozdziałach opisałem wykorzystane przeze mnie metody szacowania odległości między wektorami.

## 4.1. Wykorzystanie nierówności trójkąta

Na początek warto przypomnieć nierówność trójkąta.

**Twierdzenie 1.** Dla dowolnych wektorów , i :

Stąd, odległości i do arbitralnie wybranego wektora (czyli różnica ) zapewniają pesymistyczne oszacowanie odległości między i . Owa pesymistyczna odległość między wektorami i w odniesieniu do wektora będzie oznaczana jako . Wektor niezbędny do wyznaczania pesymistycznego oszacowania będzie nazywany *wektorem referencyjnym*. Oczywiście, wartość pesymistycznego oszacowania zależy od wyboru wektora referencyjnego.

Załóżmy, że dla każdego rozważanego wektora odległość do punktu referencyjnego została już obliczona. W takiej sytuacji określenie pesymistycznego oszacowania odległości między wektorami i jest niezwykle szybkie, jako że wymaga jedynie odjęcia uprzednio obliczonych wartości od . Jeżeli szukane są wektory nie dalsze od danego wektora niż , to gdy pesymistyczne oszacowanie jest większe niż , wtedy odległość wektora od jest również większa niż . W takim przypadku nie jest konieczne obliczanie odległości między wektorami i , którego złożoność zależy liniowo od liczby wymiarów, aby upewnić się, że jest większa od .

Rozważmy wektory , i takie, że . Wtedy, pesymistyczne oszacowanie odległości między wektorami i : . Stąd, implikuje . Zatem, jeśli , to bez żadnych dodatkowych obliczeń wiadomo, że , z czego wynika, że odległość między i jest większa od .

Analogicznie, rozważmy wektory , i takie, że . Wtedy, . Stąd, implikuje . Czyli, jeżeli , wtedy beż żadnych dodatkowych obliczeń wiadomo, że ., z czego wynika, że odległość między i jest większa od .

Powyższe obserwacje prowadzą do następującego wniosku.

**Wniosek 1.** Niech będzie dowolnym wektorem, a zbiorem wektorów posortowanych niemalejąco względem ich odległości do . Niech będzie dowolnym wektorem z , będzie wektorem następującym po w takim, że , a będzie wektorem poprzedzającym w takim, że . Wtedy:

1. i wszystkie wektory następujące po w nie należą do otoczenia epsilonowego w ;
2. i wszystkie wektory poprzedzające w nie należą do otoczenia epsilonowego w .

Tak więc, warto jest uporządkować wektory zbioru względem odległości do punktu referencyjnego , ponieważ umożliwia to prostą eliminację potencjalnie licznego podzbioru wektorów nie należących do otoczenia epsilonowego rozpatrywanego wektora.

**Przykład 3.** Niech będzie wektorem referencyjnym o współrzędnych (0,0). Na Rys. 8 przedstawiono zbiór wektorów przestrzeni dwuwymiarowej. Tabela Tab. 3 przedstawia zbiór wektorów uporządkowany niemalejąco względem odległości jego wektorów do wektora referencyjnego . Rozważmy wyznaczenie sąsiedztwa wektora o . Odległość do wektora referencyjnego jest równa 7,07 (). Pierwszym wektorem następującym po wektorze w takim, że jest wektor . Natomiast pierwszym wektorem poprzedzającym wektor w takim, że jest wektor . Przez Wniosek 1, tylko wektory następujące po i poprzedzające w (czyli wektory , , , , ) mogą należeć do . Zatem, , , oraz są jedynymi wektorami, dla których należy obliczyć odległość do wektora w celu właściwego wyznaczenia otoczenia . Przestrzeń potencjalnych sąsiadów wektora wyznaczona w oparciu o wektor referencyjny została oznaczona na Rys. 8 jako pole ograniczone przez okręgi o środkach w . Otoczenie zostało oznaczone na rys1 jako koło o środku w , pokryte szachownicą.

Tab. 3. Zbiór wektorów , wraz z odległościami do wektora i wektora referencyjnego

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nazwa punktu** | **X** | **Y** | **odl. do R** | **odl. do P** |
| A | 3 | 1 | 3,16 | 4,47 |
| B | 3 | 4 | 5,00 | 0,24 |
| C | 5 | 4 | 6,40 | 1,00 |
| P | 5 | 5 | 7,07 | 0,00 |
| D | 2 | 7 | 7,28 | 3,61 |
| E | 8 | 1 | 8,06 | 5,00 |
| F | 6 | 6 | 8,48 | 1,41 |
| G | 6 | 8 | 10,00 | 3,16 |
| H | 9 | 5 | 10,30 | 4,00 |
| I | 9 | 9 | 12,73 | 5,66 |



Rys. 8. Zbiór wektorów

Nierówność trójkąta można również zastosować w określaniu sąsiedztwa dowolnego wektora w zbiorze wektorów ponieważ problem ten można sprowadzić do wyznaczania otoczenia epsilonowego. Dla każdego wektora , można określić wartość w taki sposób, że . Najmniejsza wartość taka, że , będzie nazywana *promieniem* .

**Twierdzenie 2.** Niech . Wtedy i jest promieniem .

**Twierdzenie 3.** Jeżeli , to .

W praktyce odległość , w zasięgu której gwarantowane jest znalezienie sąsiadów wektora , jest zmniejszana w trakcie obliczania odległości między a kolejnymi wektorami z , różnymi od .

## 4.2. Wykorzystanie indeksu metrycznego

Dany jest skończony podzbiór przestrzeni zwany *zbiorem wektorów*, zadaniem jest zlokalizowanie, dla dowolnego *zapytania* należącego do przestrzeni, elementu zbioru wektorów najbliższego zapytaniu.

Przykładowym narzędziem pozwalającym na rozwiązanie postawionego problemu jest drzewo kd. Drzewo kd jest drzewem BSP[[5]](#footnote-6) tworzonym poprzez rekurencyjną bisekcję zbioru wektorów na podstawie ich położenia względem hiperpłaszczyzny tnącej. W każdym przebiegu rekurencji zbiór wektorów dzielony jest na podzbiory względem mediany rozkładu tworzonego przez rzutowanie zbioru wektorów na k-ty wymiar. Numer wymiaru, na który dokonywane jest rzutowanie, zmienia się cyklicznie i ma taką samą wartość na danym poziomie drzewa kd. Niestety struktura ta podatna jest na przekleństwo wymiaru. Gdy liczba wymiarów wzrasta, wyszukiwanie w drzewie kd szybko zaczyna odwiedzać wszystkie węzły drzewa.

Podobnie jak drzewo kd, indeks metryczny jest drzewem BSP. Każdy węzeł indeksu metrycznego dzieli przestrzeń na dwie podprzestrzenie. W procesie podziału zamiast korzystać ze współrzędnych, indeks metryczny posługuje się odległością od wybranego *wektora obserwacyjnego*. Wektory bliskie wektorowi obserwacyjnemu tworzą *lewą/wewnętrzną* podprzestrzeń, podczas gdy *prawa/zewnętrzna* podprzestrzeń składa się z dalszych wektorów. Rekurencyjne stosowanie wyżej opisanego podziału prowadzi do utworzenia drzewa binarnego. Każdy węzeł tego drzewa zawiera punkt obserwacyjny danej przestrzeni, odległość progową, na podstawie której dokonano podziału na podprzestrzenie, oraz wskazania na punkty obserwacyjne podprzestrzeni – swoich potomków.

W procesie budowy indeksu metrycznego przestrzeń metryczna dekomponowana jest przy użyciu sferycznych cięć o środkach w punktach obserwacyjnych. Rozwiązanie to kontrastuje z wykorzystaniem podziału hiperpłaszczyznami w drzewie kd. Obie metody dekompozycji zostały zilustrowane na przykładzie przestrzeni dwuwymiarowej na rysunkach (?) i (?).

**Wgląd teoretyczny**

Dana jest pewna przestrzeń metryczna oraz skończony podzbiór reprezentujący zbiór wektorów, wśród których wyszukiwane jest najbliższe sąsiedztwo. Dla wektora problem najbliższego sąsiedztwa sprowadza się do znalezienia wektora najmniej odległego od i należącego do . Operacja ta będzie dalej oznaczana jako . Ponieważ wektor najbliższy wektorowi może być od niego dość odległy, warto wprowadzić odległość progową , poza którą nie jesteśmy zainteresowani istnieniem sąsiadów . Należy zwrócić uwagę, że w czasie obliczania wartość może być redukowana z każdym kolejnym napotkanym bliższym sąsiadem . Wyszukiwanie sąsiedztwa ograniczane w wyżej wymieniony sposób będzie oznaczane przez .

W dalszej części rozważań załóżmy, że zasięg funkcji odległości przestrzeni jest równy przedziałowi . Ponieważ każdy metryczny zasięg może być sprowadzony do przedziału bez wpływu na relację sąsiedztwa[[6]](#footnote-7), obostrzenie to może zostać wprowadzone bez straty ogólności.

Niech będzie ograniczoną przestrzenią metryczną . Dla danego wektora oraz :

Funkcja jest symetryczna oraz spełnia nierówność trójkąta, stąd , a konsekwencją tej relacji jest implikacja . Czyli jeśli w procesie poszukiwania napotkano już wektor w odległości od , to w dalszej części poszukiwań nie należy brać pod uwagę elementów, dla których .

Dla pewnego wektora rozważmy przeciwdziedzinę dziedziny w . Przez oznaczymy medianę dzielącą na i . Pierwszy z tych przedziałów leży wewnątrz sfery , natomiast drugi z nich składa się z punktów leżący na powierzchni oraz poza sferą. Dziedziny pierwszego i drugiego przedziału oznaczymy odpowiednio przez i . Innymi słowy wektor obserwacyjny dzieli zbiór wektorów na podzbiory (lewy/wewnętrzny) i (prawy/zewnętrzny).

Niech oznacza liczność podzbioru a liczność podzbioru . W ogólności niewiele można powiedzieć o relacji między i nie czyniąc żadnych założeń co do natury przestrzeni metrycznej. Wiadomo jednak, że podział wektorów z jest najlepszy gdy , czyli gdy nie więcej niż jeden z wektorów leży na sferze .

Na tym etapie rozważań powinno już być zrozumiałe, że jedne wektory obserwacyjne mogą być lepsze od innych. Jako przykład rozważmy dwuwymiarową przestrzeń unormowaną, w której znajduje się równomiernie rozłożony zbiór wektorów. W zadanej sytuacji należy wybrać w taki sposób aby fragment powstałego wycinka koła zajmował połowę powierzchni przestrzeni. Rozważmy trzy przykładowe wektory obserwacyjne , i . Rozmieszczenie wektorów obserwacyjnych wraz z przynależnymi im liniami podziału zilustrowano na rysunku. W tabeli znajdują się własności wektorów obserwacyjnych.

Tab. 4. Własności przykładowych wektorów obserwacyjnych

|  |  |  |
| --- | --- | --- |
| Wektor obserwacyjny | Promień linii podziału | Długość linii podziału |
|  | 0,7979 | 1,2533 |
|  | 0,5225 | 1,338 |
|  | 0,3989 | 2,5066 |

Wiemy już, że najlepszy wektor obserwacyjny , to taki, dla którego co najwyżej jeden z wektorów leży na sferze . Oczywistym jest, że prawdopodobieństwo położenia wektora na powierzchni jest proporcjonalne do powierzchni , a dla rozpatrywanego przypadku długości linii podziału. Stąd, najlepszym z przykładowych wektorów obserwacyjnych jest wektor . Z powyższego przykładu płynie intuicyjny wniosek, że wektory znajdujące się blisko rogów przestrzeni są najlepszymi wektorami obserwacyjnymi.

**Algorytm**

Korzeń indeksu metrycznego odnosi się do całej rozpatrywanej przestrzeni wektorów. Jego punkt obserwacyjny dzieli przestrzeń na lewą i prawą podprzestrzeń, które odpowiadają lewemu i prawemu potomkowi korzenia. Każdy kolejny węzeł drzewa nawiązuje do coraz to mniejszych podprzestrzeni.

Algorytm budowy indeksu metrycznego korzysta z funkcji *Select\_vp*, której celem jest obieranie lepszych niż losowe wektorów obserwacyjnych. Funkcja ta losowo konstruuje zbiór kandydatów do wektora obserwacyjnego. Następnie, dla każdego wektora zbioru kandydatów losowo konstruowany jest podzbiór przestrzeni, dla którego wyznaczana jest mediana oraz odchylenie standardowe. Spośród zbioru kandydatów , wybierany jest ten o największym odchyleniu standardowym.

**MAKE\_VP\_TREE()**

**if** **then return** ;

new(node);

node.p:=SELECT\_VP();

node.μ:=;

:=;

:=;

node.left=MAKE\_VP\_TREE();

node.right=MAKE\_VP\_TREE();

**return** node;

**SELECT\_VP()**

:=;

best\_spread:=0;

**for**

:=;

μ:=;

spread:=;

if spread>best\_spread

best\_spread:=spread;

best\_p:=p;

**return** best\_p;

W przedstawionym algorytmie w węźle indeksu metrycznego oprócz wektora obserwacyjnego przechowywana jest wartość mediany w celu opisania metrycznej relacji między punktem obserwacyjnym a lewą i prawą podprzestrzenią.

# Bibliografia

x

|  |  |
| --- | --- |
| [1] | Ester M., Kriegel H.P., Sander J., and Xu X., "A Density-Based Algorithm of Discovering Clusters in LargeSpatial Database with Noise," , Portland, 1966. |
| [2] | Elkan C., "Using the triangle inequality to accelerate k-Means," 2003. |
| [3] | Kryszkiewicz M. and Lasek P., "TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality," 2010. |
| [4] | Kryszkiewicz M. and Lasek P., "A Neighborhood Based Clustering by Means of the Triangle Inequality and Reference Points," 2011. |
| [5] | Kryszkiewicz M., "Determining Cosine Similarity Neighborhoods by Means of the Euclidean Distance," 2012. |
| [6] | Kryszkiewicz M., "Efficient Determination of Neighborhoods Defined in Terms of Codine Similarity Measure," 2011. |
| [7] | Yanilos P., "Data Struvtures and Algorithms of Nearest Neighbor Search in General Metric Spaces,". |

x

1. kosinusowe sąsiedztwo to sąsiedztwo oparte na podobieństwie kosinusowym [↑](#footnote-ref-2)
2. α-znormalizowany wektor to znormalizowany wektor przemnożony przez stałą α [↑](#footnote-ref-3)
3. Dendrogram to diagram stosowany do prezentacji związków między elementami lub grupami elementów w kształcie przypominający drzewo. [↑](#footnote-ref-4)
4. Epsilonowe sąsiedztwo jest epsilonowym otoczeniem danego punktu bez niego samego. [↑](#footnote-ref-5)
5. BSP – (ang. Binary Search Partitioning) metoda dokonująca rekurencyjnego podziału przestrzeni na podprzestrzenie za pomocą hiperpłaszczyzn. Podział ten tworzy reprezentację obiektów w przestrzeni zwaną drzewem BSP. Wyszukiwanie w drzewie BSP jest wyszukiwaniem binarnym. [↑](#footnote-ref-6)
6. Ograniczone przestrzenie metryczne mogą zostać w prosty sposób przeskalowane. Nieograniczone przestrzenie metryczne mogą zostać dostosowane dzięki zastosowaniu wzoru: [↑](#footnote-ref-7)