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](#_Toc349428005)

[1.1. Przegląd literatury 1](#_Toc349428006)

[1.2. Motywacja i cel pracy 2](#_Toc349428007)

[1.3. Układ pracy 2](#_Toc349428008)

[2. Miary odległości i podobieństwa 3](#_Toc349428009)

[2.1. Metryki odległości 3](#_Toc349428010)

[2.2. Miara odległości kosinusowej (1) 4](#_Toc349428011)

[2.3. Wyznaczanie kosinusowego sąsiedztwa za pomocą sąsiedztwa opartego na odległości euklidesowej? 6](#_Toc349428012)

[3. Użyte algorytmy 7](#_Toc349428013)

[3.1. Grupowanie gęstościowe na przykładzie DBSCAN 8](#_Toc349428014)

[3.2. Wyszukiwanie k-sąsiadów 12](#_Toc349428015)

# 

# 1. Wprowadzenie

Współczesne systemy komputerowe agregują i generują ogromną ilość danych, która rośnie szybciej 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. Jej odkrywaniem znajduje się dziedzina informatyki zwana eksploracją danych, której ideą jest wykorzystanie komputera do znajdowania ukrytych dla człowieka prawidłowości w danych zgromadzonych w repozytoriach. Eksploracja danych jest czwartym etapem procesu odkrywania wiedzy, na który również składają się operacje selekcji, czyszczenia i transformacji danych a także analiza i interpretacja wyników.

Jednymi z najpopularniejszych metod eksploracji danych są grupowanie, klasyfikacja oraz odkrywanie asocjacji i sekwencji. 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.

Prawdopodobnie najpopularniejszym algorytmem gęstościowego grupowania danych jest DBSCAN ?? 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 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 ??, TI-NBC i PreDeCon. Dokonano również badania wpływu liczby punktów referencyjnych i strategii ich wyboru na efektywność tych algorytmów ??.

## 1.2. Motywacja i cel pracy

Grupowanie danych to proces powszechnie stosowany w porządkowaniu produktów, segmentacji klientów, organizacji obiektów czy rozpoznawaniu i nanalizie 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

Po wprowadzeniu w zagadnienia grupowania danych, gruntownie przestawiłem algorytm DBSCAN. Opis cech charakterystycznych algorytmu oraz specyficznej taksonomii zostały uzupełnione o pseudokody, do których odwołuję się w kolejnych rozdziałach, co pozwala spójnie i precyzyjnie przedstawić zmiany, które wprowadzone są w algorytmie w związku z wykorzystaniem nierówności trójkąta. Teoretycznie podstawy wprowadzanych modyfikacji przedstawiłem na początku rozdziału trzeciego.

# 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, jedne miary mogą być stosowniejsze niż inne w danym przypadku. Najpopularniejszą miarą odległości jest *odległość Euklidesowa*. *Odległość Euklidesowa* między punktami i oznaczana jest jako i definiowana w następujący sposób:

Jeśli 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:

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

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

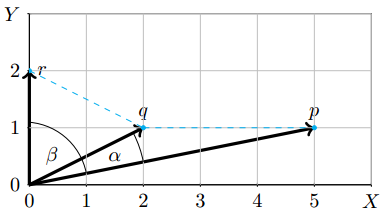
Dla uproszczenia, bez straty ogólności, w dalszych rozważaniach jako metryką odległości będę posługiwał się odległością euklidesową.

## 2.2. Miara odległości kosinusowej (1)

W wielu aplikacjach odkrywających wiedzę w danych tekstowych stosowana jest *miara podobieństwa kosinusowego* będąca funkcją konta między dwoma wektorami. *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 konta między i () jest większy niż kosinus konta między i ().

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

W pierwszej kolejności początkowy zbiór wektorów transformowany jest do ich α-znormalizowanych odpowiedników. Następnie

# 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 leżace blisko siebie jako podobne, jednak większość algorytmów jest niezależna od przyjętej miary podobieństwa. Liczność 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[[2]](#footnote-3) - 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.

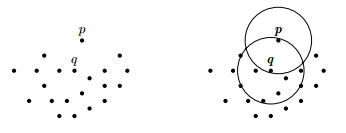
Podstawowym pojęciem używanym w kontekście algorytmu DBSCAN jest *otoczenie epsilonowe* oznaczane przez i definiowane jako zbiór takich punktów zbioru D, które są różne od i nie bardziej odległe od niż , czyli:

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. 2. 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 .

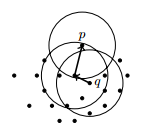


Rys. 2. Przykład ilustrujący punkty rdzeniowe i brzegowe

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.

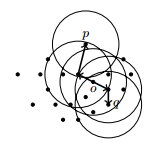
Mówimy, że punkt jest *gęstościowo połączony* z punktem względem i , jeżeli istnieje punkt taki, że i są gęstościowo osiągalne z względem i .

Na rys. 3. 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. 3. Przykład ilustrujący relację gęstościowej osiągalności

Na rys. 4. 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. 4. 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[[3]](#footnote-4) 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);

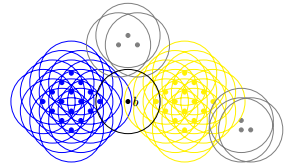
END WHILE;

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. 5..



Rys. 5. 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-sąsiadów

1. α-znormalizowanym wektorem nazywamy znormalizowany wektor przemnożony przez stałą α [↑](#footnote-ref-2)
2. Dendrogram to diagram stosowany do prezentacji związków między elementami lub grupami elementów w kształcie przypominający drzewo. [↑](#footnote-ref-3)
3. Epsilonowe sąsiedztwo jest epsilonowym otoczeniem danego punktu bez niego samego. [↑](#footnote-ref-4)