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. 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](#_Toc355125758)

[1.1. Przegląd literatury 2](#_Toc355125759)

[1.2. Motywacja i cel pracy 3](#_Toc355125760)

[1.3. Układ pracy 3](#_Toc355125761)

[2. Miary odległości i podobieństwa 4](#_Toc355125762)

[2.1. Metryki odległości 4](#_Toc355125763)

[2.2. Miara odległości Kosinusowej 5](#_Toc355125764)

[2.3. Wyznaczanie kosinusowego sąsiedztwa za pomocą sąsiedztwa opartego na odległości Euklidesowej 7](#_Toc355125765)

[3. Grupowanie gęstościowe 10](#_Toc355125766)

[3.1. Algorytm DBSCAN 11](#_Toc355125767)

[3.2. Wyszukiwanie k-najbliższych sąsiadów 16](#_Toc355125768)

[4. Szacowanie odległości 17](#_Toc355125769)

[4.1. Wykorzystanie nierówności trójkąta 17](#_Toc355125770)

[4.2. Wykorzystanie indeksu metrycznego 21](#_Toc355125771)

[4.2.1. Wgląd teoretyczny 22](#_Toc355125772)

[5. Użyte algorytmy 26](#_Toc355125773)

[5.1. Zastosowanie nierówności trójkąta w algorytmach gęstościowego grupowania danych 26](#_Toc355125774)

[5.1.1. Algorytm TI-DBSCAN 26](#_Toc355125775)

[5.1.2. Algorytm TI-DBSCAN-REF 29](#_Toc355125776)

[5.1.3. Algorytm TI-k-Neighborhood-Index 31](#_Toc355125777)

[5.1.4. Algorytm TI-k-Neighborhood-Index-Ref 34](#_Toc355125778)

[5.2. Zastosowanie rzutowania w algorytmach gęstościowego grupowania danych 36](#_Toc355125779)

[5.2.1. Algorytm DBSCAN-PROJECTION 36](#_Toc355125780)

[5.2.2. Algorytm k-Neighborhood-Index-Projection 37](#_Toc355125781)

[5.3. Zastosowanie indeksu metrycznego w algorytmie wyszukiwania k sąsiadów 37](#_Toc355125782)

[5.3.1. Algorytm k-Neighborhood-Index-Vp-Tree 38](#_Toc355125783)

[6. Szczegóły implementacji 41](#_Toc355125784)

[7. Badania eksperymentalne 44](#_Toc355125785)

[7.1. Dane testowe 45](#_Toc355125786)

[7.2. Badania algorytmu k-Neighborhood-Index 47](#_Toc355125787)

[7.2.1. Badania algorytmu TI-k-Neighborhood-Index 47](#_Toc355125788)

[Bibliografia 17](#_Toc355125789)

# 

# 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-k-Neighborhoo-Index [[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

W kolejnych rozdziałach zamiennie będę posługiwał się terminami wektor i punkt.[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 swoich rozważaniach będę posługiwał się odległością euklidesową jako metryką odległości.

W dalszych podrozdziałach, 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, chemii ,czy inżynierii biomedycznej, *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 rozumiany jest jako sekwencja komponentów , gdzie komponent jest wartością -tego wymiaru , gdzie . 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 definiowana w następujący sposób:

**Przykład 1**. Na 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 Tab. 1 zamieszczono wartości podobieństwa kosinusowego między wektorami , i . Można zauważyć, że dla podobieństwa kosinusowego warunki oraz i nie są spełnione.



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 poniższe 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 układają się na okręgu o środku w punkcie i promieniu .



Rys. 2. Przykładowy zbiór wektorów

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. 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. Grupowanie gęstościowe

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. W dendrogramie, relacja między węzłami a ich przodkami 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 zlecić człowiekowi zadanie pogrupowania punktów dwuwymiarowej przestrzeni, to 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 nich można wyszczególnić pewną gęstość punktów znacznie wyższą niż gęstość punktów poza klastrem. Zatem, do grupy należą punkty leżące w obszarze o gęstości wyraźnie większej niż punkty leżące w obszarze otaczającym ją. Tak zdefiniowanemu pojęciu metody grupowania najbliżej jest algorytmom gęstościowym.

W podrozdziałach przedstawiono opis algorytmu DBSCAN na podstawie [[1](#MEs66)] oraz opis wyszukiwania k-najbliższych sąsiadów.

## 3.1. Algorytm 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 ten wymaga podania jedynie dwóch parametrów wejściowych, które opisują najmniejszy klaster będący obiektem zainteresowania. Jest to promień wokół danego punktu, wewnątrz którego to promienia 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 , 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 , którego otoczenie epsilonowe zawiera wymaganą liczbę punktów, czyli:

Mówimy, że punkt jest *bezpośrednio gęstościowo osiągalny* z punktu względem i , jeżeli należy do otoczenia epsilonowego oraz 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 jest zbiorem punktów. Grupą względem i nazywamy niepusty zbiór 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 względem i , . *Szumem* nazywamy podzbiór punktów zbioru 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 *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, nie mające przypisanej żadnej grupy również dodawane są do nowoutworzonej grupy. Te z nich, które nie są oznaczone jako szum dodawane są do zbioru *seeds*. Na Wydruk 1 wyżej opisany algorytm został zapisany w formie pseudokodu.

Analiza kodu pozwala zauważyć, że algorytm DBSCAN jest deterministyczny z dokładnością do punktów brzegowych. Nie uwzględnia on, że punkty brzegowe znajdujące się między leżącymi blisko siebie grupami mogą należeć do kilku z grup. Taka sytuacja została przedstawiona na Rys. 7.



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

**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);

**endif**;

**endif**;

**endfor**;

**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);

**endif**;

SetOfPoints.changeClId(resultP, ClId);

**endif**;

**endfor**;

**endif**;

seeds.delete(currentP);

**endwhile**;

**return** true;

**endif**;

**end**; //ExpandCluster

Wydruk 1. Zapis algorytmu DBSCAN w formie pseudokodu

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

Kolejnym podejściem do zagadnienia gęstościowego grupowania jest poszukiwanie k najbliższych sąsiadów. Wyznaczanie 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, oraz 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.

# 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 użyte przeze mnie metody szacowania odległości między wektorami. Indeks metryczny został opisany na podstawie artykułów [[6](#PYa)] i [[7](#TBo)].

## 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 , czyli:

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 , czyli:

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 .

Zatem, 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 Rys. 8 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 do 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 pewnego zbioru punktów przestrzeni dwuwymiarowej na rysunkach Rys. 9 i Rys. 10.



Rys. 9. Dekompozycja przykładowego zbioru punktów za pomocą drzewa kd



Rys. 10. Dekompozycja przykładowego zbioru punktów za pomocą indeksu metrycznego

### 4.2.1. 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 Rys. 11. W Tab. 4znajdują się własności wektorów obserwacyjnych.



Rys. 11. Przykład rozmieszczenia 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, proporcjonalne do długości linii podziału. Stąd, najlepszym z przykładowych wektorów obserwacyjnych jest wektor , dla którego linia podziału jest najkrótsza. Z powyższego przykładu płynie intuicyjny wniosek, że wektory znajdujące się blisko rogów przestrzeni są najlepszymi wektorami obserwacyjnymi.

# 5. Użyte algorytmy

Analiza złożoności algorytmów przedstawionych w rozdziale 3 prowadzi do wniosku, że największy wpływ na wydajność procesu grupowania ma obliczanie odległości między wektorami. W celu poprawienia wydajności grupowania, czyli ograniczenia liczby obliczanych odległości, można posłużyć się metodami szacowania odległości opisanymi w rozdziale 4. W niniejszym rozdziale opisałem badane przeze mnie odmiany algorytmów grupowania gęstościowego. Odmiany te zostały oparte na uprzednio opisanych algorytmach i metodach szacowania odległości.

## 5.1. Zastosowanie nierówności trójkąta w algorytmach gęstościowego grupowania danych

Użycie nierówności trójkąta jako metody szacowania odległości poprawia wydajność algorytmów gęstościowego grupowania danych. W kolejnych podrozdziałach opisałem jej zastosowanie w algorytmie DBSCAN oraz w wyszukiwaniu k-najbliższych sąsiadów.

### 5.1.1. Algorytm TI-DBSCAN

W artykule [[3](#Kry10)] po raz pierwszy przedstawiono wykorzystanie nierówności trójkąta jako sposobu zwiększenia wydajności algorytmu DBSCAN. Opublikowaną wersję algorytmu autorzy nazwali TI-DBSCAN. Jego pseudokod został zamieszczony na Wydruk 2 i Wydruk 3. W algorytmie tym, po znanej z DBSCAN inicjalizacji punktów i oznaczeniem ich jako nieprzypisane do żadnej z grup, obliczana jest odległość każdego z punktów do uprzednio wybranego punktu referencyjnego. Następnie w oparciu o te wartości punkty grupowanego zbioru sortowane są niemalejąco.

**algorithm TI-DBSCAN**(D, Eps, MinPts)

D’ := empty set of points;

TI-Init(D);

ClusterId := nextId(NOISE);

**for** each point p in the ordered set D starting from the first point until last point in D **do**

**if** TI-ExpandCluster(D, D’, p, ClusterId, Eps, MinPts) **then**

ClusterId := nextId(ClusterId);

**endif**;

**endfor;**

**return** D’; //D’ is clustered set of points

**end;** //TI-DBSCAN

**function TI-ExpandCluster**(D, D’, p, ClusterId, Eps, MinPts)

seeds := TI-Neighborhood(D, p, Eps);

p.neighborsNr := p. neighborsNr + |seeds|;

**if** p.neighborsNr < MinPts **then**

p.clusterId := NOISE;

**for** each point q in seeds **do**

add p to q.border; q.neighborsNr := q.neighborsNr + 1;

**endfor;**

p.border := ; move p from D to D’; **return** false;

**else**

assign ClusterId to all b in p.border;

assign ClusterId to p; assign ClusterId to all q in seeds;

**for** each point q in seeds **do**

q.neighborsNr := q.neighborsNr + 1;

**endfor;**

p.border := ; move p from D to D’;

**while** |seeds| > 0 **do**

curP := first point in seeds;

curSeeds := TI-Neighborhood(D, curP, Eps);

curP.neighborsNr := curP.neighborsNr + |curSeeds|;

**if** curP.neighborsNr < MinPts **then**

**for** each point q in curSeeds **do**

q.neighborsNr := q.neighborsNr + 1;

**endfor;**

**else**

**for** each point q in curSeeds **do**

q.neighborsNr := q.neighborsNr + 1;

**if** q.clusterId = UNCLASSIFIED **then**

assign ClusterId to q; move q from curSeeds to seeds;

**else**

delete q from curSeeds;

**endif;**

**endfor;**

assign ClusterId to all b in curP.border;

**endif;**

curP.border := ; move curP from D to D’;

delete curP from seeds;

**endwhile;**

**return** true;

**endif;**

**end;** //TI-ExpandCluster

Wydruk 2. Pseudokod algorytmu TI-DBSCAN

**function TI-Init**(D)

rPoint := selectReferencePoint();

**for** each point p in D **do**

p.clusterId := UNCLASSIFIED; p.neighborsNr := 1; p.border := ;

p.dist := distance(p, rPoint);

**end for**

sort D non-decreasingly w.r.t. p.dist;

**end;** //TI-Init

**function TI-Neighborhood**(D, p, Eps)

**return** TI-Backward-Neighborhood(D, p, Eps) TI-Forward-Neighborhood(D, p, Eps);

**end;** //TI-Neighborhood

**function TI-Backward-Neighborhood**(D, p, Eps)

seeds := {};

backwardThreshold := Eps - p.dist;

**for** each point q in the ordered set D starting from the point immediately preceding point p until first point in D **do**

**if** q.dist < backwardThreshold **then**

**break;**

**endif**;

**if** distance(q,p) Eps **then**

append q to seeds;

**endif;**

**endfor**;

**return** seeds;

**end;** // TI-Backward-Neighborhood

**function TI-Forward-Neighborhood**(D, p, Eps)

Seeds := {};

forwardThreshold := Eps + p.dist;

**for** each point q in the ordered set D starting from the point immediately following point p until the last point in D **do**

**if** q.dist > forwardThreshold **then**

**break**;

**endif;**

**if** distance(p,q) Eps **then**

append q to seeds;

**endif;**

**endfor;**

**return** seeds;

**end**; // TI-Forward-Neighborhood

Wydruk 3. Pseudokod algorytmu TI-DBSCAN

Istotną różnica między algorytmem TI-DBSCAN a DBSCAN jest zastosowanie i rozszerzenie opisanej w [[8](#MKr05)] koncepcji usuwania ze zbioru D przeanalizowanych punktów. Podejście to zakłada, że każdy punkt dodatkowo przechowuje liczbę dotychczas znalezionych sąsiadów oraz zbiór punktów brzegowych. Informacje te pozwalają usuwać z analizowanego zbioru danych D wszystkie zbadane punkty.

Dzięki zastosowanemu rozwiązaniu funkcja TI-ExpandCluster iteruje jedynie po zbiorze dotychczas niezbadanych punktów. Największą korzyścią płynącą z działania na ograniczanym zbiorze punktów występuje w procesie wyznaczania sąsiedztwa punktu, ponieważ możliwe jest uniknięcie wielu obliczeń rzeczywistych odległości między punktami. Algorytm zapewnia, że operacja wyznaczania odległości między dwoma punktami zostanie wykonana najwyżej raz. Jednakże takie zapewnienie nie przychodzi bez ceny, którą jest wzrost zapotrzebowania na pamięć oraz skomplikowania algorytmu.

Kluczową modyfikacją algorytmu TI-DBSCAN względem DBSCAN jest użycie funkcji TI-Neighborhood, która dla zadanego punktu zwraca jego epsilonowe sąsiedztwo. Wynik tej funkcji stanowi sumę teoriomnogościową zbiorów będących rezultatami wywołań funkcji TI-Backward-Neighborhood i TI-Forward-Neighborhood wyszukujących punkty należące do epsilonowego sąsiedztwa danego punktu znajdujące się w indeksie odpowiednio przed i po danym punkcie. Obie funkcje przeglądają indeks odpowiednio w tył i przód, do momentu napotkania punktu, którego odległość do punktu referencyjnego różni się od odległości badanego punktu do punktu referencyjnego o więcej niż wartość Eps. Dalsze przeglądanie punktów indeksu w danym kierunku jest zbędne, ponieważ, zgodnie z teorią przedstawioną w [rozdziale 4.1](#_4.1._Wykorzystanie_nierówności), nie należą one do epsilonowego sąsiedztwa weryfikowanego punktu. Pseudokod dotyczący omówionych funkcji zamieściłem na wydruku 3.

### 5.1.2. Algorytm TI-DBSCAN-REF

TI-DBSCAN-REF jest odmianą algorytmu TI-DBSCAN opisaną w artykule [[3](#Kry10)] wykorzystującą wiele punktów referencyjnych do estymacji odległości między dwoma punktami. Dodatkowe punkty referencyjne używane są tylko wtedy gdy podstawowy punkt referencyjny, względem którego posortowany jest indeks, nie wystarcza do oszacowania czy dany punkt należy do epsilonowego otoczenia badanego punktu. Rzeczywista odległość między dwoma punktami obliczana jest tylko wtedy gdy żaden z punktów referencyjnych nie pozwala na oszacowanie przynależności do otoczenia epsilonowego. Dodatkowym kosztem wynikającym z posłużenia się wieloma punktami referencyjnymi jest wyznaczanie odległości między nimi a punktami badanego zbioru.

Na Wydruk 4 zamieściłem pseudokod funkcji składających się na algorytm TI-DBSCAN-REF różnych od funkcji algorytmu TI-DBSCAN. Szarym zaznaczeniem wyróżniłem fragmenty pseudokodu, różne od odpowiedniego pseudokodu algorytmu TI-DBSCAN.

**function TI-REF-Init**(D)

rPoints := selectReferencePoints();

**for** each point p in D **do**

p.clusterId := UNCLASSIFIED; p.neighborsNr := 1; p.border := ;

**for** each i := 1..rPoints.size() **do**

p.dists[i] := distance(p,rPoints[i]);

**endfor**;

**end for**

sort D non-decreasingly w.r.t. p.dists[0];

**end;** //TI-REF-Init

**function TI-REF-Neighborhood**(D, p, Eps)

**return** TI-REF-Backward-Neighborhood(D, p, Eps) TI-REF-Forward-Neighborhood(D, p, Eps);

**end;** //TI-REF-Neighborhood

**function TI-REF-Backward-Neighborhood**(D, p, Eps)

backwardThreshold := Eps - p.dists[0]; seeds := {};

**for** each point q in the ordered set D starting from the point immediately preceding point p until first point in D **do**

**if** q.dists[0] < backwardThreshold **then**

**break;**

**endif**;

candidateNeighbor := **true**; i := 1;

**while** candidateNeighbor and (i |p.dists|) **do**

**if** |q.dists[i] – p.dists[i]| > Eps **then**

candidateNeighbor := **false**;

**else**

i := i + 1;

**endif**;

**endwhile;**

**if** candidateNeighbor && (distance(q,p) Eps) **then**

append q to seeds;

**endif;**

**endfor**;

**return** seeds;

**end;** // TI-REF-Backward-Neighborhood

**function TI-Forward-Neighborhood**(D, p, Eps)

forwardThreshold := Eps + p.dists[0]; seeds := {};

**for** each point q in the ordered set D starting from the point immediately following point p until the last point in D **do**

**if** q.dists[0] > forwardThreshold **then**

**break**;

**endif;**

candidateNeighbor := **true**; i := 1;

**while** candidateNeighbor and (i |p.dists|) **do**

**if** |q.dists[i] – p.dists[i]| > Eps **then**

candidateNeighbor := **false**;

**else**

i := i + 1;

**endif**;

**endwhile;**

**if** candidateNeighbor && (distance(p,q) Eps) **then**

append q to seeds;

**endif;**

**endfor;**

**return** seeds;

**end**; // TI-Forward-Neighborhood

Wydruk 4. Pseudokod funkcji algorytmu TI-DBSCAN-REF rożnych od funkcji algorytmu TI-DBSCAN

### 5.1.3. Algorytm TI-k-Neighborhood-Index

W artykule [[4](#MKr11_2)] zaproponowano wykorzystanie nierówności trójkąta w wyszukiwaniu k-sąsiedztwa. Opisanemu algorytmowi twórcy nadali nazwę TI-k-Neighborhood-Index. Dla k-sąsiedztwa zastosowanie nierówności trójkąta jest bardziej zawiłe niż w przypadku DBSCAN. Skomplikowanie to wynika z faktu, że na początku wykonania algorytmu promień Eps sąsiedztwa nie jest znany, w dodatku należy wyznaczyć go niezależnie dla każdego punktu w oparciu o dystans do k-tego najbliższego sąsiada.

Pierwszym krokiem algorytmu jest wyznaczenie odległości do punktu referencyjnego dla wszystkich punktów zbioru, po czym wykonywane jest jego sortowanie w porządku niemalejącym względem obliczonych odległości. Następnie główna pętla algorytmu tworzy indeks, poprzez wyznaczenie wszystkim punktom ich k-sąsiedztwa za pomocą funkcji TI-k-Neighborhood. Pierwszym etapem wyznaczania k-sąsiedztwa jest szacowanie minimalnej wartości Eps gwarantującej znalezienie k sąsiadów. Punkty znajdujące się w promieniu o oszacowanej wartości tworzą zbiór kandydatów. Faktyczny promień k-sąsiedztwa jest mniejszy bądź równy oszacowanej wartości Eps. Aby wyznaczyć rzeczywiste k-sąsiedztwo, w kolejnym etapie dokonuje się weryfikacji zbioru kandydatów. Zaczynając od badanego punktu posortowany zbiór danych iterowany jest zarówno w przód jak i w tył, do momentu napotkania punktów, których odległość od punktu referencyjnego różni się od odległości badanego punktu do punktu referencyjnego o wartość większą niż Eps. Wartość Eps aktualizowana jest za każdym razem gdy znajdowany jest nowy sąsiad, który wstawiany jest do zbioru kandydatów. Pseudokod TI-k-Neighborhood-Index zamieściłem na Wydruk 5, Wydruk 6, Wydruk 7.

**algorithm TI-k-Neighborhood-Index**(D, k)

TI-Init(D);

**for** each point p in ordered set D starting from the first point until last point in D **do**

insert(position od point p, TI-k-Neighborhood(D, p, k))

into k-Neighborhood-Index)

**endfor**;

**end**; //TI-k-Neighborhood-Index

**function TI-Init**(D)

rPoint := selectReferencePoint();

**for** each point p in D **do**

p.dist := distance(p, rPoint);

**endfor;**

sort D non-decreasingly w.r.t. p.dist;

**end;** //TI-Init

Wydruk 5. Pseudokod algorytmu TI-k-Neighborhood-Index

**function TI-k-Neighborhood**(D, p, k)

b := p;

f := p;

backwardSearch := PrecedingPoint(D, b);

forwardSearch := FollowingPoint(D, b);

k-Neighborhood := {};

i := 0;

Find-k-Candidate-Neighbours(D, p, k, i, b, f, k-Neighborhood,

backwardSearch, forwardSearch);

p.Eps := max({e.dist | e k-Neighborhood});

Verify-k-Candidate-Neighbours-Backward(D, p, b, backwardSearch, k,

k-Neighborhood);

Verify-k-Candidate-Neighbours-Forward(D, p, f, forwardSearch, k,

k-Neighborhood);

**return** k-Neighborhood;

**end**; //TI-k-Neighborhood

**function Find-k-Candidate-Neighbours**(D, p, k, i, b, f, k-Neighborhood,

backwardSearch, forwardSearch)

**while** (backwardSearch **and** forwardSearch **and** i < k) **do**

**if** p.dis – b.dist < f.dist – p.dist **then**

dist := distance(b, p);

i := i + 1;

e := (b, dist);

insert e into k-Neighborhood holding it sorted wrt. e.dist;

backwardSearch := PrecedingPoint(D, b);

**else**

dist := distance(f, p);

i := i + 1;

e := (f, dist);

insert e into k-Neighborhood holding it sorted wrt. e.dist;

forwardSearch := FollowingPoint(D, b);

**endif**;

**endwhile**;

**while** backwardSearch **and** i < k) **do**

dist := distance(b, p);

i := i + 1;

e := (b, dist);

insert e into k-Neighborhood holding it sorted wrt. e.dist;

backwardSearch := PrecedingPoint(D, b);

**endwhile**;

**while** forwardSearch **and** i < k) **do**

dist := distance(f, p);

i := i + 1;

e := (f, dist);

insert e into k-Neighborhood holding it sorted wrt. e.dist;

forwardSearch := FollowingPoint(D, b);

**endwhile**;

**end**; //Find-k-Candidate-Neighbours

**function PrecedingPoint**(D, p)

**if** there is a point in D preceding p **then**

p := point immediately preceding p in D;

**return** **true**;

**endif**;

**return false;**

**end**; //PrecedingPoint

Wydruk 6. Pseudokod algorytmu TI-k-Neighborhood-Index

**function FollowingPoint**(D, p)

**if** there is a point in D following p **then**

p := point immediately following p in D;

**return** **true**;

**end if**;

**return false;**

**end**; //FollowingPoint

**function Verify-k-Candidate-Neighbours-Backward**(D, p, b, backwardSearch, k, k-Neighborhood)

**while** backwardSearch **and** ((p.dist – b.dist) p.Eps) **do**

dist := distance(b, p);

e := (b, dist);

**if** dist < p.Eps **then**

i := |{e k-Neighborhood | e.dist = p.Eps}|;

**if** |k-Neighborhood| - i k – 1 **then**

delete each element e with e.dist = p.Eps from k-Neighborhood;

insert e into k-Neighborhood holding it sorted wrt. e.dist;

p.Eps := max({e.dist | e k-Neighborhood});

**else**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif**;

**else**

**if** dist = p.Eps **then**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif;**

**endif;**

backwardSearch := PrecedingPoint(D, b);

**endwhile;**

**end;** //Verify-k-Candidate-Neighbours-Backward

**function Verify-k-Candidate-Neighbours-Forward**(D, p, f, forwardSearch, k, k-Neighborhood)

**while** forwardSearch **and** ((f.dist – p.dist) p.Eps) **do**

dist := distance(f, p);

e := (f, dist);

**if** dist < p.Eps **then**

i := |{e k-Neighborhood | e.dist = p.Eps}|;

**if** |k-Neighborhood| - i k – 1 **then**

delete each element e with e.dist = p.Eps from k-Neighborhood;

insert e into k-Neighborhood holding it sorted wrt. e.dist;

p.Eps := max({e.dist | e k-Neighborhood});

**else**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif**;

**else**

**if** dist = p.Eps **then**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif;**

**endif;**

forwardSearch := FollowingPoint(D, b);

**endwhile;**

**end;** //Verify-k-Candidate-Neighbours-Forward

Wydruk 7. Pseudokod algorytmu TI-k-Neighborhood-Index

### 5.1.4. Algorytm TI-k-Neighborhood-Index-Ref

Oprócz TI-k-Neighborhood-Index artykuł [[4](#MKr11_2)] proponuje jego odmianę stosującą wiele punktów referencyjnych. Odmiana ta będzie dalej nazywana TI-k-Neighborhood-Index-Ref. Zastosowanie wielu punktów referencyjnych pozwala na przyspieszenie weryfikacji zbioru kandydatów do k sąsiedztwa. Usprawnienie tą metodą wcześniejszych kroków algorytmu nie jest możliwe, ponieważ bez szacunkowej znajomości Eps nie można zaostrzyć warunku wykluczania elementów ze zbioru kandydatów. Dodatkowy warunek spełnienia nierówności trójkąta został dodany w funkcjach Verify-k-Candidate-Neighbours-Backward i Verify-k-Candidate-Neighbours-Forward. Pseudokod algorytmu TI-k-Neighborhood-Index-Ref zamieściłem na wydrukach. Szarym kolorem wyróżniłem fragmenty różne od odpowiedniego pseudokodu algorytmu TI-k-Neighborhood-Index.

**function TI-REF-Init**(D)

rPoints := selectReferencePoints();

**for** each point p in D **do**

**for** each i := 1..rPoints.size() **do**

p.dists[i] := distance(p,rPoints[i]);

**endfor**;

**endfor;**

sort D non-decreasingly w.r.t. p.dists[1];

**end;** //TI-REF-Init

**function Is-Candidate-Neighbor-By-Additional-Reference-Points**(p, q,

Eps)

candidateNeighbor := **true**;

i := 2;

**while** candidateNeighbor **and** (i |p.dists|) **do**

**if** |q.dists[i] – p.dists[i]| > Eps **then**

candidateNeighbor := **false**;

**else**

i := i + 1;

**endif**;

**endwhile**;

**return** candidateNeighbor;

**end**; //Is-Candidate-Neighbor-By-Additional-Reference-Points

Wydruk 8. Pseudokod funkcji algorytmu TI-k-Neighborhood-Index-Ref różnych od funkcji algorytmu TI-k-Neighborhood-Index

**function Verify-k-Candidate-Neighbours-Forward**(D, p, f, forwardSearch, k, k-Neighborhood)

**while** forwardSearch **and** ((f.dists[1] – p.dists[1]) p.Eps) **do**

**if** Is-Candidate-Neighbor-By-Additional-Reference-Points(p, b, p.Eps)

**then**

dist := distance(f, p);

e := (f, dist);

**if** dist < p.Eps **then**

i := |{e k-Neighborhood | e.dist = p.Eps}|;

**if** |k-Neighborhood| - i k – 1 **then**

delete each element e with e.dist = p.Eps from k-Neighborhood;

insert e into k-Neighborhood holding it sorted wrt. e.dist;

p.Eps := max({e.dist | e k-Neighborhood});

**else**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif**;

**else**

**if** dist = p.Eps **then**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif;**

**endif;**

**endif;**

forwardSearch := FollowingPoint(D, b);

**endwhile;**

**end;** //Verify-k-Candidate-Neighbours-Forward

**function Verify-k-Candidate-Neighbours-Backward**(D, p, b, backwardSearch, k, k-Neighborhood)

**while** backwardSearch **and** ((p.dists[1] – b.dists[1]) p.Eps) **do**

**if** Is-Candidate-Neighbor-By-Additional-Reference-Points(p, b, p.Eps)

**then**

dist := distance(b, p);

e := (b, dist);

**if** dist < p.Eps **then**

i := |{e k-Neighborhood | e.dist = p.Eps}|;

**if** |k-Neighborhood| - i k – 1 **then**

delete each element e with e.dist = p.Eps from k-Neighborhood;

insert e into k-Neighborhood holding it sorted wrt. e.dist;

p.Eps := max({e.dist | e k-Neighborhood});

**else**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif**;

**else**

**if** dist = p.Eps **then**

insert e into k-Neighborhood holding it sorted wrt. e.dist;

**endif;**

**endif;**

**endif;**

backwardSearch := PrecedingPoint(D, b);

**endwhile;**

**end;** //Verify-k-Candidate-Neighbours-Backward

Wydruk 9. Pseudokod funkcji algorytmu TI-k-Neighborhood-Index-Ref różnych od funkcji algorytmu TI-k-Neighborhood-Index

## 5.2. Zastosowanie rzutowania w algorytmach gęstościowego grupowania danych

Alternatywnym sposobem usprawnienia algorytmów gęstościowego grupowania jest zastosowanie rzutowania na wymiar. Łatwo zauważyć, że dla każdego wymiaru , i każdych dwóch wektorów i prawdziwe jest, że:

Stąd, jeśli , wtedy ; czyli .

Obserwacja ta prowadzi następującego twierdzenia.

**Twierdzenie 5.** Niech będzie indeksem wymiaru , gdzie , i jest zbiorem wektorów posortowanych niemalejąco względem wartości -tego wymiaru. Niech , będzie takim wektorem następującym po w , że i będzie takim wektorem poprzedzającym w , że . Wtedy:

* i wszystkie wektory następujące po w nie należą do ,
* i wszystkie wektory poprzedzające w nie należą do .

Zgodnie z twierdzeniem 2 opisanym w rozdziale 4.1 twierdzenie 5 można zastosować również do wyszukiwania k sąsiedztwa.

### 5.2.1. Algorytm DBSCAN-PROJECTION

Modyfikacja przedstawionego uprzednio algorytmu TI-DBSCAN stosująca rzutowanie na wymiar będzie w dalszej części pracy nazywana DBSCAN-PROJECTION. W stosunku do bazowego algorytmu zmianie ulega jedynie funkcja inicjalizująca dane, której pseudokod przedstawiłem na Wydruk 10. Kolorem szarym wyróżniłem fragmenty różne od odpowiedniego kodu algorytmu TI-DBSCAN.

**function Init**(D)

pDimension := selectProjectionDimension();

**for** each point p in D **do**

p.clusterId := UNCLASSIFIED; p.neighborsNr := 1; p.border := ;

p.dist := p[pDimension];

**end for**

sort D non-decreasingly w.r.t. p.dist;

**end;** //TI-Init

Wydruk 10. Pseudokod funkcji algorytmu DBSCAN-PROJECTION różnych od funkcji algorytmu TI-DBSCAN

### 5.2.2. Algorytm k-Neighborhood-Index-Projection

W dalszej części pracy modyfikacja przedstawionego uprzednio algorytmu TI-k-Neighborhood-Index stosująca rzutowanie na wymiar będzie nazywana k-Neighborhood-Index-Projection. Podobnie jak w przypadku DBSCAN-PROJECTION zmianie uległa jedynie funkcja inicjalizująca dane, której pseudokod przedstawiłem na Wydruk 11. Kolorem szarym wyróżniłem fragmenty różne od odpowiedniego kodu algorytmu TI-k-Neighborhood-Index.

**function Init**(D)

pDimension := selectProjectionDimension();

**for** each point p in D **do**

p.dist := p[pDimension];

**end for**

sort D non-decreasingly w.r.t. p.dist;

**end;** //TI-Init

Wydruk 11. Pseudokod funkcji algorytmu k-Neighborhood-Index-Projection różnych od funkcji algorytmu TI-k-Neighborhood-Index

## 5.3. Zastosowanie indeksu metrycznego w algorytmie wyszukiwania k sąsiadów

Zastosowanie indeksu metrycznego pozwala na sprawne wyszukiwanie najbliższego sąsiada danego wektora w pewnym zbiorze wektorów. W oparciu o implementację zaproponowaną w artykule [[6](#PYa)] stworzyłem algorytm gęstościowego grupowania danych poszukujący k najbliższych sąsiadów, który wykorzystuje indeks metryczny. Algorytm ten nazwałem k-Neighborhood-Index-Vp-Tree jego pseudokod zamieściłem na wydrukach Wydruk 12, Wydruk 13 i Wydruk 14.

### 5.3.1. Algorytm k-Neighborhood-Index-Vp-Tree

Pierwszym krokiem algorytmu jest budowa indeksu metrycznego w oparciu o zbiór wektorów D. Korzeń indeksu 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. Niech najmniejsza wartość większą od mediany będzie nazywana prawym ograniczeniem, a największą wartość mniejszą od mediany będzie nazywana lewym ograniczeniem. W przedstawionym algorytmie w węźle indeksu metrycznego oprócz wektora obserwacyjnego przechowywana jest wartość mediany oraz lewe i prawe ograniczenie w celu opisania metrycznej relacji między punktem obserwacyjnym a lewą i prawą podprzestrzenią. Algorytm budowy indeksu metrycznego korzysta z funkcji *select\_vp*, której celem jest obieranie lepszych niż losowe wektorów obserwacyjnych. Funkcja ta w sposób stochastyczny 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.

Kolejnym krokiem algorytmu k-Neighborhood-Index-Vp-Tree jest wyszukanie k sąsiadów, przy użyciu uprzednio zbudowanego indeksu, dla każdego wektora zbioru zapytań Q. K sąsiedztwo przechowywane jest w formie kolekcji par: (odległość między wektorem p a zapytaniem q; wektor p). Dla każdego wektora zbioru zapytań Q kolekcja k sąsiedztwa   
*k-Neighborhood* inicjalizowana jest k parami (maksymalna wartość odległości; null), a następnie przekazywana jest wraz zapytaniem i korzeniem indeksu metrycznego do funkcji *search-k-Neighborhood*. Funkcja ta przegląda w głąb indeks metryczny w poszukiwaniu dla zapytania k sąsiedztwa, które zapisywane jest w kolekcji *k-Neighborhood*.

W stworzonym algorytmie proponuję dwie metody przeszukiwania indeksu metrycznego w poszukiwaniu k sąsiedztwa. Pierwszą, która korzysta z mediany jako z kryterium przeszukiwania poddrzew indeksu. Jej pseudokod przedstawiłem na Wydruk 13. Oraz drugą, która wykorzystuje ograniczenie górne i dolne w celu wyboru poddrzewa do przeszukania. Jej pseudokod zamieściłem na Wydruk 14.

**algorithm k-Neighborhood-Index-Vp-Tree** (D, Q, k)

vp\_tree := make-Vp-Tree(D);

**for** each point q in Q **do**

k-Neighborhood := ;

**for** i := 1..k **do**

insert(MAX\_DOUBLE, null) into k-Neighborhood;

**endfor**;

search-k-Neighborhood(vp\_tree, q, k-Neighborhood)

insert(position of point q, k-Neighborhood) into k-Neighborhood-Index)

**endfor**;

**end**; //k-Neighborhood-Index-Vp-Tree

**function make-Vp-Tree**(D)

**if** **then**

**return** ;

**endif**;

vantage\_potint := select-Vp(D);

new(node);

node.p := vantage\_potint;

node.μ :=;

node.leftLimit := ;

node.rightLimit := ’

:= {dD - vantage\_point | distance(vantage\_point,d) };

:= {dD - vantage\_point | distance(vantage\_point,d) };

node.left := make-Vp-Tree();

node.right := make-Vp-Tree();

**return** node;

**end;** //mave-Vp-Tree

**function select-Vp**(D)

P := random sample of D;

best\_spread := 0;

best\_vp := null;

**for** each point p in P **do**

S := random sample of D;

μ := ;

spread := ;

**if** spread > best\_spread **then**

best\_spread := spread;

best\_p := p;

**endif;**

**endfor;**

**return** best\_p;

**end; /**/select-Vp

Wydruk 12. Pseudokod funkcji algorytmu k-Neighborhood-Index-Vp-Tree

**function search-k-Neighborhood**(vp\_tree\_p, p, k-Neighborhood)

**if** p is null **then**

**return**;

**endif**;

tau := get distance to the most distant neighbor from k-Neighborhood;

dist := distance(vp\_tree\_p.p, p);

leftBoundary := dis – tau;

rightBoundary := dist + tau;

**if** dist tau **do**

erase the most distant neighbor from k-Neighborhood;

insert(dist, vp\_tree\_p.p) into k-Neighborhood;

**endif**;

**if** rightBoundary vp\_tree\_p.μ **then**

search-k-Neighborhood(vp\_tree\_p.right, p, k-Neighborhood);

**endif**;

**if** leftBoundary vp\_tree\_p.μ **then**

search-k-Neighborhood(vp\_tree\_p.left, p, k-Neighborhood);

**endif**;

**end**; //search-k-Neighborhood

Wydruk 13. Pseudokod funkcji wyszukującej k sąsiadów danego punktu w indeksie metrycznym z wykorzystaniem mediany

**function search-k-Neighborhood**(vp\_tree\_p, p, k-Neighborhood)

**if** p is null **then**

**return**;

**endif**;

tau := get distance to the most distant neighbor from k-Neighborhood;

dist := distance(vp\_tree\_p.p, p);

leftBoundary := dis – tau; rightBoundary := dist + tau;

**if** dist tau **do**

erase the most distant neighbor from k-Neighborhood;

insert(dist, vp\_tree\_p.p) into k-Neighborhood;

**endif**;

**if** leftBoundary vp\_tree\_p.leftLimit **then**

**if** rightBoundary vp\_tree\_p.rightLimit **then**

leftBuffer := vp\_tree\_p.leftLimit – leftBoundary;

rightBuffer := rightBoundary – tree.rightLimit;

**if** leftBuffer < rightBuffer **do**

search-k-Neighborhood(vp\_tree\_p.left, p, k-Neighborhood);

search-k-Neighborhood(vp\_tree\_p.right, p, k-Neighborhood);

**else**

search-k-Neighborhood(vp\_tree\_p.right, p, k-Neighborhood);

search-k-Neighborhood(vp\_tree\_p.left, p, k-Neighborhood);

**endif**;

**else**

search-k-Neighborhood(vp\_tree\_p.left, p, k-Neighborhood);

**endif**;

**else**

**if** rightBoundary vp\_tree\_p.rightLimit **then**

search-k-Neighborhood(vp\_tree\_p.right, p, k-Neighborhood);

**endif**;

**endif**;

**end**; //search-k-Neighborhood

Wydruk 14. Pseudokod funkcji wyszukującej k sąsiadów danego punktu w indeksie metrycznym z wykorzystaniem prawego i lewego ograniczenia

# 6. Szczegóły implementacji

W niniejszym rozdziale przedstawię zaprojektowane i stworzone przeze mnie oprogramowanie służące do analizy algorytmów. Oprogramowanie to jest przeznaczone dla systemu operacyjnego Microsoft Windows i implementuje wszystkie algorytmy omówione w poprzednich rozdziałach.

**Architektura**

Zasadniczym celem implementacji jest analiza oraz testy opisanych uprzednio algorytmów, ze szczególnym zwróceniem uwagi na możliwość usprawnienia ich przez zastosowanie nierówności trójkąta lub odległości kosinusowej. Mnogość odmian implementowanych algorytmów wymusiła zastosowanie rozbudowanej hierarchii dziedziczenia w architekturze silnik algorytmów. Dziedziczenie umożliwiło uniknięcie powtarzania wspólnego kodu w implementacjach modyfikacji algorytmów, co pozwoliło na ujednolicenie wspólnych funkcji i procesów, a co najistotniejsze zmniejszyło ryzyko wpływu różnic implementacyjnych na badaną wydajność algorytmu.

**Wykorzystane technologie**

Oprogramowanie zostało napisane w całości w języku C++ w środowisku Microsoft Visual Studio. Język C++ łączy zaletę obiektowości z wydajnością, nie występują w nim niekontrolowane operacje czyszczenia pamięci (ang. garbage collection) jak w Java, które w przypadku algorytmów operujących na dużych ilościach danych mogą zakłócić wyniki testów. W celu zapewnienia efektywnej realizacji operacji przetwarzania danych takich jak wczytywanie, implementacja algorytmów, funkcji pomocniczych i zapisywanie wykorzystywana jest standardowa biblioteka szablonów STL.

**Interfejs użytkownika**

Stworzona aplikacja wykonywana jest w trybie wsadowym. W pierwszej kolejności wczytuje ona zdefiniowane przez użytkownika pliki parametrów znajdujące się w folderze *properties*. Każdy z tych plików definiuje wykonanie algorytmu z danymi parametrami na danym zbiorze danych. Rezultat wykonania pojedynczego uruchomienia algorytmu oraz raport z jego uruchomienia zapisywane są do plików tekstowych w folderze *logs*. Dodatkowo aplikacja na bieżąco tworzy plik raportu o nazwie: *ultimate\_run\_report%\_data\_i\_czas\_uruchomienia%.csv*, który zawiera wartości parametrów programu oraz czasy wykonania poszczególnych etapów algorytmów, dla wszystkich uruchomień algorytmów w danym uruchomieniu programu. Format CSV pozwala na łatwą analizę wyników dzięki narzędziom takim jak arkusz kalkulacyjny. Plik parametrów aplikacji *algorithms\_engine\_properties.txt* pozwala na skonfigurowanie jej w taki sposób, aby każdy algorytm zdefiniowany plikiem parametrów w folderze *logs* był wykonywany w serii n razy. W celu poprawy wygody pracy analityka aplikacja na bieżąco tworzy plik raportu *cleaned\_ultimate\_run\_report%\_data\_i\_czas\_uruchomienia%.csv*, który zawiera wartości parametrów programu oraz średnie czasy wykonania poszczególnych etapów algorytmów, dla wszystkich serii uruchomień algorytmów. Dzięki wprowadzeniu mechanizmu uruchamiania algorytmów seriami możliwe jest zmniejszenie wpływu losowych zakłóceń na rezultaty wykonania algorytmu poprzez badanie średnich czasów z serii.

Dokładny opis plików parametrów, możliwości konfiguracji aplikacji oraz jej uruchamiania znajduje się w dodatku ??.

# 7. Badania eksperymentalne

W przeprowadzonych przeze mnie eksperymentach badałem wydajność wyszukiwania   
k-sąsiedztwa oraz epsilonowego sąsiedztwa z zastosowaniem nierówności trójkąta oraz wpływ sposobu ich zaimplementowania. Eksperymenty przeprowadziłem na środowisku Windows 7 x64 z procesorem Intel® i7™ 950 z dostępną pamięcią RAM równą 6GB. Aplikacja, którą posłużyłem się do zebrania wyników eksperymentów została skompilowana dla środowiska WIN32 z podniesioną flagą LARGEADDRESSAWARE, która umożliwia użycie do 3GB pamięci wirtualnej. Eksperymenty uruchamiałem jako proces o priorytecie ‘Wysoki’ aby uniknąć wywłaszczenia procesu z rdzenia zmniejszając tym samym ryzyko losowego zakłócenia wyników.

Zauważyłem, że kolejne uruchomienia danego algorytmu z tymi samymi parametrami, nazwijmy je serią, wykonują się w różnym czasie. Ze względu na charakter wspomnianych różnic wyszczególniłem dwa przypadki odchyleń. Pierwszy, w którym różnice czasów wykonań algorytmów w danej serii są niewielkie oraz drugi, w którym owe różnice są znaczące. W pierwszym przypadku powodem odchyleń były losowe zakłócenia. W drugim przypadku zauważyłem pewną prawidłowość, według której pierwsze uruchomienie algorytmu w serii trwa najdłużej z wszystkich, kolejne trochę krócej a różnice czasów trwania następnych uruchomień odpowiadają tym z przypadku pierwszego. Dzieje się tak ponieważ cache procesora (w moim przypadku 8MB, L3) z każdym wykonaniem algorytmu w serii działa sprawniej do chwili osiągnięcia pewnej sprawności granicznej dla danej serii. W efekcie czas wykonania algorytmu skraca się aż do momentu gdy różnice czasów kolejnych wykonań przypominają te z przypadku pierwszego. Aby umniejszyć wpływ wyżej opisanych zjawisk na zbieranie wyników czasowych poszczególnych kroków algorytmów ich pomiar dokonywany jest w następujący sposób: spośród wszystkich uruchomień algorytmu w serii odrzucane są te, których czas wykonania był najdłuższy i najkrótszy a z pozostałych wartości obliczany jest średni czas wykonania każdego z kroków algorytmu. W prezentowanych dalej wynikach eksperymentów, czas wykonania algorytmu jest sumą średnich czasów wykonań każdego z jego kroków. W każdej tabeli prezentującej czasy wykonania algorytmów zamieszczono również średnie czasy kroków nań składających się.

**Słownik opisu rezultatów badań**

W celu zmniejszenia szerokości tabel zawierających rezultaty badań, opisy kolumn zostały dokonane skróconymi wyrażeniami, których rozwinięcia oraz wyjaśnienia znajdują się w poniższej tabeli:

Tab. 5. Słownik opisów rezultatów badań

|  |  |  |
| --- | --- | --- |
| **Skrócone wyrażenie** | **Pełne wyrażenie** | **Znaczenie** |
| **bud. ind.** | budowa indeksu | czas budowy indeksu [s] |
| **grup.** | grupowanie | czas wykonania grupowania [s] |
| **l. p.** | liczba punktów | liczność zbioru, na którym wykonywana jest eksploracja |
| **norm.** | normalizacja | czas normalizacji wektorów zbioru danych [s] |
| **obl. odl.** | obliczanie odległości | czas obliczania odległości wektorów do punktów referencyjnych lub czas rzutowania wektorów na wymiar [s] |
| **sort.** | sortowanie | czas sortowania zbioru danych [s] |
| **wysz. sąsiad.** | wyszukiwanie sąsiadów | czas wyszukiwania sąsiadów [s] |
| **wyk. alg.** | wykonanie algorytmu | czas wykonania algorytmu eksploracji danych [s] |
| **-** | brak danych | Rezultat eksploracji danych nie został osiągnięty w czasie mniejszym niż 3h lub z powodu wyczerpania pamięci operacyjnej maszyny, na której wykonywano badanie |

## 7.1. Dane testowe

W eksperymentach wykorzystałem zbiory danych stosowane w dziedzinowej literaturze. Ich odmienne charakterystyki pozwoliły na sprawdzenie działania algorytmów w różnych warunkach. Zamieszczone poniżej opisy dotyczą zbiorów, które wykorzystałem do przeprowadzenia eksperymentów.

Covtype [[9](#12Ma)] to zbiór udostępnianym przez US Forest Service. Jego pełna nazwa to Forest Cover Type. Baza ta zawiera informacje na temat gatunków drzew w amerykańskich lasach. Zbiór posiada 581012 rekordów w formacie gęstym o 55 atrybutach. 10 pierwszych atrybutów to wartości zmiennych ilościowych, kolejne 44 atrybuty posiadają wartości 0 lub 1, z kolei ostatni determinuje jeden z siedmiu gatunków drzew. Należy zwrócić uwagę, że każdy z rekordów spośród atrybutów binarnych przyjmuje wartość 1 dokładnie dwa razy, a wartość zero 42 razy. W przypadku 10 pierwszych atrybutów zróżnicowanie wartości jest znacznie większe, z których pierwsze dwa posiadają najszerszą dziedziną (ponad 7000), a pozostałe charakteryzuje znacznie mniejsza różnorodność i zakres.

KDD-Cup98 [[10](#The12)] jest zbiorem opublikowanym podczas Drugiego Międzynarodowego Konkursu Odkrywania Wiedzy i Eksploracji Danych, który miał miejsce podczas KDD-98, Czwartej Międzynarodowej Konferencji Odkrywania Wiedzy i Eksploracji Danych w 1998 roku. Zbiór posiada 96367 rekordów w formacie gęstym o 56 atrybutach o wartościach będącymi liczbami naturalnymi. Najszerszą dziedziną [0; 6000] posiadają dwa atrybuty, najwęższą dziedzinę [0; 13] również tylko dwa atrybuty, natomiast aż 44 atrybuty posiadają taką samą dziedzinę [0; 99].

Karypis\_sport to zbiór danych wchodzący w skład bazy zbiorów [[11](#GKa12)] udostępnianych na stronie profesora George Karypis z Department of Computer Science & Engineering, University of Minesota. Zbiór karypis\_sport zawiera 8580 rekordów w formacie rzadkim o 126373 atrybutach przyjmujących wartości będące liczbami naturalnymi. Baza ta zawiera informacje o dokumentach dotyczących sportu. Średnia niezerowych wartości atrybutów zbioru wynosi 1,58, natomiast najwyższa wartość to 129. Mimo, że punkty posiadają 126373 atrybutów, to średnio 129 z nich jest niezerowych. Punkt o największej liczbie niezerowych atrybutów posiada ich 1174. Liczba punktów posiadających nie więcej niezerowych atrybutów niż 56 wynosi 1478.

Karypis\_review jest zbiorem danych wchodzącym w ten sam skład bazy zbiorów co karypis\_sport [[11](#GKa12)]. Zawiera on 4069 rekordów w formacie rzadkim o 126373 atrybutach przyjmujących wartości będące liczbami naturalnymi. Baza ta skupia informacje o dokumentach dotyczących recenzji. Średnia niezerowych wartości atrybutów zbioru wynosi 1,49, natomiast najwyższa wartość to 131. Mimo, że punkty posiadają 126373 atrybutów, to średnio 191 z nich jest niezerowych. Punkt o największej liczbie niezerowych atrybutów posiada ich 1385. Liczba punktów posiadających nie więcej niezerowych atrybutów niż 56 wynosi 254.

## 7.2. Badania algorytmu k-Neighborhood-Index

[TODO]

### 7.2.1. Badania algorytmu TI-k-Neighborhood-Index

[TODO]

#### 7.2.1.1. Implementacja algorytmu

Na przykładowych zbiorach danych przetestowałem dwie wersje algorytmu TI-k-Neighborhood-Index: korzystającą ze struktury indeksu zbudowanego z iteratorów oraz opartą wyłącznie na oryginalnym zbiorze danych. W przypadku pierwszej implementacji sortowaniu poddany zostaje indeks, który następnie używany jest w celu dostępu do zbioru danych. W drugim przypadku sortowany jest zbiór danych i żaden indeks nie jest wykorzystywany w dostępie do jego elementów. W Tab. 6 zamieściłem czasy wykonania obu implementacji algorytmów wraz ze składającymi się na nie operacjami. Rys. 12 przedstawia porównanie wydajności badanych implementacji algorytmu *Tik-k-Neighborhood-Index*.

Analiza wyników z Tab. 6 pozwala stwierdzić, że koszt budowy indeksu dla implementacji z indeksem jest pomijalnie mały. Kolejną ważną cechą jest porównywalny dla obu implementacji koszt obliczania odległości do punktu referencyjnego. Na tym niestety podobieństwa się kończą. W porównaniu z implementacją bez indeksu implementacja z indeksem iteratorów radzi sobie ze znajdowaniem k-sąsiedztwa tym lepiej im więcej rekordów posiada zbiór, w którym szukane jest k-sąsiedztwo. Największa różnica w czasie wykonania algorytmu wystąpiła przy wyszukiwaniu k-sąsiedztwa dla 50000 punktów zbioru covtype. Dla tego konkretnego przypadku implementacja z bezpośrednim dostępem do danych wykonuje się blisko 7 razy wolniej niż implementacja korzystająca z indeksu.

Tab. 6. Porównanie wydajności implementacji algorytmu TI-k-Neighborhood-Index na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbiór** | **l.p.** | **TI-k-Neighborhood-Index dostęp przez indeks** | | | | | **TI-k-Neighborhood-Index bezpośredni dostęp** | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,204 | 0,204 | 0,000 | 0,000 | 0,000 | 0,235 | 0,206 | 0,004 | 0,025 |
| 2000 | 0,678 | 0,674 | 0,000 | 0,004 | 0,000 | 0,729 | 0,660 | 0,004 | 0,065 |
| 4000 | 2,824 | 2,813 | 0,000 | 0,010 | 0,001 | 2,932 | 2,755 | 0,010 | 0,166 |
| 6000 | 6,887 | 6,870 | 0,000 | 0,016 | 0,001 | 7,122 | 6,570 | 0,015 | 0,537 |
| 8000 | 13,228 | 13,202 | 0,000 | 0,023 | 0,002 | 13,440 | 12,736 | 0,024 | 0,680 |
| karypis\_review | 500 | 0,068 | 0,066 | 0,000 | 0,002 | 0,000 | 0,079 | 0,066 | 0,002 | 0,012 |
| 1000 | 0,249 | 0,246 | 0,000 | 0,003 | 0,000 | 0,279 | 0,247 | 0,003 | 0,029 |
| 2000 | 1,105 | 1,099 | 0,000 | 0,006 | 0,000 | 1,208 | 1,106 | 0,006 | 0,096 |
| 3000 | 2,394 | 2,385 | 0,000 | 0,008 | 0,001 | 2,598 | 2,380 | 0,008 | 0,210 |
| 4000 | 4,346 | 4,334 | 0,000 | 0,011 | 0,001 | 4,637 | 4,303 | 0,012 | 0,322 |
| covtype | 10000 | 0,547 | 0,391 | 0,000 | 0,153 | 0,003 | 1,364 | 0,184 | 0,058 | 1,122 |
| 50000 | 7,750 | 5,400 | 0,001 | 2,333 | 0,015 | 78,517 | 4,147 | 2,089 | 72,281 |
| 100000 | 19,989 | 15,052 | 0,002 | 4,900 | 0,035 | 79,311 | 11,357 | 4,636 | 63,318 |
| 300000 | 99,592 | 83,380 | 0,004 | 16,079 | 0,130 | 619,451 | 50,367 | 15,277 | 553,806 |
| 500000 | 261,458 | 234,340 | 0,005 | 26,861 | 0,252 | 1860,230 | 120,760 | 25,600 | 1 713,870 |
| cup98 | 10000 | 0,392 | 0,272 | 0,000 | 0,117 | 0,003 | 2,241 | 0,140 | 0,062 | 2,039 |
| 30000 | 3,073 | 2,197 | 0,000 | 0,865 | 0,010 | 3,224 | 1,191 | 0,530 | 1,503 |
| 50000 | 7,291 | 5,371 | 0,001 | 1,900 | 0,019 | 60,378 | 3,198 | 1,555 | 55,625 |
| 70000 | 12,886 | 9,563 | 0,001 | 3,293 | 0,029 | 100,792 | 5,673 | 2,772 | 92,347 |
| 90000 | 19,547 | 15,316 | 0,002 | 4,185 | 0,044 | 83,934 | 9,064 | 4,649 | 70,221 |

Kluczowym krokiem algorytmu mającym wpływ na powstałą różnicę jest sortowanie punktów. Obie implementacje do sortowania obiektów wykorzystują standardową dla języka C++ funkcję sortowania wektorów, tj. *std::sort*, która oparta jest o algorytm quicksort. Funkcja sortowania biblioteki standardowej korzysta z operacji *swap* zamieniającej miejscami elementy w wektorze, która wywołuje konstruktor kopiujący obiektu będącego elementem sortowanego wektora. W przypadku gdy głęboko kopiowany jest obiekt klasy *Punkt*, to uruchamiane są konstruktory kopiujące wszystkich obiektów będących jego polami, w szczególności kolekcji przechowujących wartości wymiarów. Dlatego, wykonanie konstruktora kopiującego iteratora jest szybsze niż wykonanie konstruktora kopiującego obiektu punktu. Im większy zbiór jest sortowany tym częściej funkcja *swap* jest wywoływana, tym samym częściej wykonywany jest konstruktor kopiujący. W skutek tego, czas sortowania zbioru obiektów punktu jest dłuższy niż czas sortowania zbioru iteratorów. Zatem mimo, że dostęp do nieposortowanych danych przez posortowany zbiór iteratorów jest wolniejszy niż bezpośredni dostęp do posortowanego zbioru, to czas sortowania bezpośrednio zbioru danych jest na tyle duży, że różnice czasowe wynikające ze sposobu dostępu do danych stają się pomijalne. W kolejnych rozważaniach będę się odnosił do algorytmu *TI-k-Neighborhood-Index* jako do implementacji korzystającej z indeksu iteratorów.

Rys. 12. Porównanie wydajności implementacji algorytmu TI-k-Neighborhood-Index na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

#### 7.2.1.2. Implementacja struktury punktu

Początkowo naturalnym podejściem do implementacji struktury punktu wydawało się przechowywanie wartości wymiarów w postaci tablicy lub wektora o liczbie elementów równej liczbie wymiarów przestrzeni danych. Pierwsze eksperymenty z danymi tekstowymi pokazały, że przy zastosowaniu wspomnianego podejścia implementacja zbioru punktów szybko wyczerpywała pamięć RAM maszyny, na której dokonywano eksperymentów. Zachowanie to wynikało z dużej liczby wymiarów przestrzeni danych. Jeden punkt przestrzeni danych tekstowych (105 wymiarowej) zajmował tyle pamięci co 5\*104 punktów przestrzeni dwuwymiarowej.

Na szczęście, przechowywanie punktów przestrzeni danych tekstowych można ulepszyć w oparciu o ich cechy charakterystyczne. Kluczową własnością danych tekstowych jest duża liczba wymiarów punktu, których wartość jest równa 0. Wartość ta nie wnosi żadnej informacji w procesie wyznaczania odległości między dwoma punktami, dlatego postanowiłem jej nie przechowywać. W tym celu zaimplementowałem rzadką reprezentację punktu w postaci listy par <*numer wymiaru*, *wartość wymiaru*> posortowanej względem numeru wymiaru.

W Tab. 7 oraz na Rys. 13 zamieszczono wyniki uruchomień algorytmu *TI-k-Neighborhood-Index* z użyciem zarówno gęstej jak i rzadkiej implementacji punktu. Rezltaty obserwacji zamieszczone w Tab. 7 pozwalają stwierdzić, że dla zbiorów cup98 i covtype, uruchomienia korzystające z implementacji punktu gęstego wykonują się szybciej niż uruchomienia posługujące się implementacją punktu rzadkiego. Odwrotne zjawisko można zauważyć dla danych tekstowych, tj. zbiorów karypis sport i karypis review, w których przypadku uruchomienia korzystające z implementacji punktu rzadkiego wykonują się o dwa rzędy wielkości szybciej niż uruchomienia korzystające z implementacji punktu gęstego. Niestety, z powodu wyczerpania pamięci RAM nie można było uzyskać rezultatów badań implementacji gęstej punktu na zbiorach danych tekstowych dla liczby rekordów większej niż 1000.

Na podstawie przeprowadzonych eksperymentów skonstruowałem następujące wnioski:

* w porównaniu z implementacją gęstego punktu, stosowanie implementacji rzadkiego punktu w uruchomieniach algorytmu pracującego na danych tekstowych przyspiesza jego wykonanie oraz zmniejsza wielkość zajętej pamięci operacyjnej;
* w porównaniu z implementacją rzadkiego punktu, stosowanie implementacji gęstego punktu w uruchomieniach algorytmu działającego na danych gęstych o niewielkim wymiarze przyspiesza jego wykonanie.

W dalszej części pracy badania algorytmów *TI-k-Neighborhood-Index* pracujących na danych tekstowych będą wykonywane z użyciem implementacji punktu rzadkiego, a działające z danymi gęstymi o niewielkim wymiarze będą wykonywane z zastosowaniem implementacji punktu gęstego.

Tab. 7. Porównanie wydajności algorytmu TI-k-Neighborhood-Index w zależności od implementacji punktu na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **TI-k-Neighborhood-Index gęsta reprezentacja punktu** | | | | | **TI-k-Neighborhood-Index rzadka reprezentacja punktu** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 22,208 | 21,959 | 0,000 | 0,249 | 0,000 | 0,204 | 0,204 | 0,000 | 0,000 | 0,000 |
| 2000 | - | - | - | - | - | 0,678 | 0,674 | 0,000 | 0,004 | 0,000 |
| 4000 | - | - | - | - | - | 2,824 | 2,813 | 0,000 | 0,010 | 0,001 |
| 6000 | - | - | - | - | - | 6,887 | 6,870 | 0,000 | 0,016 | 0,001 |
| 8000 | - | - | - | - | - | 13,228 | 13,202 | 0,000 | 0,023 | 0,002 |
| karypis\_review | 500 | 5,710 | 5,585 | 0,000 | 0,125 | 0,000 | 0,068 | 0,066 | 0,000 | 0,002 | 0,000 |
| 1000 | 22,677 | 22,428 | 0,000 | 0,249 | 0,000 | 0,249 | 0,246 | 0,000 | 0,003 | 0,000 |
| 2000 | - | - | - | - | - | 1,105 | 1,099 | 0,000 | 0,006 | 0,000 |
| 3000 | - | - | - | - | - | 2,394 | 2,385 | 0,000 | 0,008 | 0,001 |
| 4000 | - | - | - | - | - | 4,346 | 4,334 | 0,000 | 0,011 | 0,001 |
| covtype | 10000 | 0,547 | 0,391 | 0,000 | 0,153 | 0,003 | 0,348 | 0,286 | 0,000 | 0,062 | 0,000 |
| 50000 | 7,750 | 5,400 | 0,001 | 2,333 | 0,015 | 4,394 | 3,417 | 0,000 | 0,962 | 0,016 |
| 100000 | 19,989 | 15,052 | 0,002 | 4,900 | 0,035 | 16,406 | 12,995 | 0,000 | 3,380 | 0,031 |
| 300000 | 99,592 | 83,380 | 0,004 | 16,079 | 0,130 | 125,538 | 115,939 | 0,005 | 9,453 | 0,141 |
| 500000 | 261,458 | 234,340 | 0,005 | 26,861 | 0,252 | 252,783 | 235,820 | 0,016 | 16,682 | 0,265 |
| cup98 | 10000 | 0,392 | 0,272 | 0,000 | 0,117 | 0,003 | 0,670 | 0,608 | 0,000 | 0,062 | 0,000 |
| 30000 | 3,073 | 2,197 | 0,000 | 0,865 | 0,010 | 5,403 | 4,763 | 0,000 | 0,629 | 0,010 |
| 50000 | 7,291 | 5,371 | 0,001 | 1,900 | 0,019 | 13,442 | 12,054 | 0,000 | 1,367 | 0,021 |
| 70000 | 12,886 | 9,563 | 0,001 | 3,293 | 0,029 | 22,068 | 21,028 | 0,000 | 1,009 | 0,031 |
| 90000 | 19,547 | 15,316 | 0,002 | 4,185 | 0,044 | 68,218 | 65,275 | 0,000 | 2,891 | 0,052 |

Rys. 13. Porównanie wydajności algorytmu TI-k-Neighborhood-Index w zależności od implementacji punktu na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

#### 7.2.1.3. Wybór punktu referencyjnego

W poszukiwaniu właściwego punktu referencyjnego dla algorytmu *TI-k-Neighborhood-Index* przeprowadziłem badania z różnymi ich przykładami. W swoich eksperymentach skupiłem się na punktach skrajnych takich jak:

* punkt maksymalny, którego każdy z wymiarów przyjmuje maksymalną wartość z dziedziny danego wymiaru - oznaczany dalej jako *[max]*;
* punkt minimalny, którego każdy z wymiarów przyjmuje maksymalną wartość z dziedziny danego wymiaru - oznaczany dalej jako *[max]*;
* punkt, którego każdy parzysty wymiar przyjmuje minimalną wartość z dziedziny danego wymiaru, a każdy nieparzysty wymiar przyjmuje maksymalną wartość z dziedziny danego wymiaru - oznaczany dalej jako *[max\_min]*;

W rozważaniach uwzględniłem również punkt losowy, którego każdy z wymiarów przyjmuje losową wartość z dziedziny danego wymiaru - oznaczany dale jako *[rand]*.

W Tab. 8 i Tab. 9 zamieściłem czasy uruchomień algorytmu *TI-k-Neighborhood-Index.* Na Rys. 14 znajdują się wykresy czasu wykonania algorytmu w funkcji liczby punktów.

Tab. 8. Porównanie wydajności algorytmu TI-k-Neighborhood-Index w zależności od wybranego punktu referencyjnego na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **TI-k-Neighborhood-Index [min]** | | | | | **TI-k-Neighborhood-Index [max\_min]** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,171 | 0,169 | 0,000 | 0,002 | 0,000 | 0,208 | 0,203 | 0,000 | 0,005 | 0,000 |
| 2000 | 0,636 | 0,632 | 0,000 | 0,004 | 0,000 | 0,634 | 0,634 | 0,000 | 0,000 | 0,000 |
| 4000 | 2,684 | 2,675 | 0,000 | 0,008 | 0,001 | 2,719 | 2,709 | 0,000 | 0,010 | 0,000 |
| 6000 | 6,470 | 6,455 | 0,000 | 0,013 | 0,002 | 6,537 | 6,526 | 0,000 | 0,011 | 0,000 |
| 8000 | 11,770 | 11,751 | 0,000 | 0,017 | 0,002 | 12,455 | 12,439 | 0,000 | 0,016 | 0,000 |
| karypis\_review | 500 | 0,062 | 0,061 | 0,000 | 0,001 | 0,000 | 0,062 | 0,062 | 0,000 | 0,000 | 0,000 |
| 1000 | 0,241 | 0,239 | 0,000 | 0,002 | 0,000 | 0,250 | 0,250 | 0,000 | 0,000 | 0,000 |
| 2000 | 1,084 | 1,079 | 0,000 | 0,005 | 0,000 | 1,097 | 1,092 | 0,000 | 0,005 | 0,000 |
| 3000 | 2,356 | 2,348 | 0,000 | 0,007 | 0,001 | 2,371 | 2,366 | 0,000 | 0,005 | 0,000 |
| 4000 | 4,209 | 4,198 | 0,000 | 0,010 | 0,001 | 4,253 | 4,243 | 0,000 | 0,010 | 0,000 |
| covtype | 10000 | 0,551 | 0,370 | 0,000 | 0,179 | 0,002 | 0,416 | 0,333 | 0,000 | 0,083 | 0,000 |
| 50000 | 7,804 | 5,526 | 0,001 | 2,262 | 0,015 | 6,775 | 5,086 | 0,000 | 1,674 | 0,015 |
| 100000 | 19,962 | 15,025 | 0,001 | 4,901 | 0,035 | 19,568 | 15,262 | 0,000 | 4,274 | 0,031 |
| 300000 | 93,868 | 78,047 | 0,003 | 15,686 | 0,132 | 107,443 | 91,785 | 0,000 | 15,527 | 0,130 |
| 500000 | 246,482 | 219,916 | 0,005 | 26,312 | 0,249 | 277,321 | 250,843 | 0,000 | 26,229 | 0,249 |
| cup98 | 10000 | 0,388 | 0,259 | 0,000 | 0,126 | 0,003 | 0,389 | 0,296 | 0,000 | 0,093 | 0,000 |
| 30000 | 2,785 | 2,105 | 0,001 | 0,669 | 0,010 | 3,697 | 3,011 | 0,000 | 0,671 | 0,016 |
| 50000 | 6,942 | 5,287 | 0,001 | 1,635 | 0,019 | 9,486 | 7,972 | 0,000 | 1,498 | 0,016 |
| 70000 | 13,095 | 9,933 | 0,001 | 3,132 | 0,029 | 17,643 | 14,809 | 0,000 | 2,803 | 0,031 |
| 90000 | 19,041 | 14,683 | 0,001 | 4,316 | 0,041 | 27,622 | 23,228 | 0,005 | 4,347 | 0,042 |

Tab. 9. Porównanie wydajności algorytmu TI-k-Neighborhood-Index w zależności od wybranego punktu referencyjnego na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **TI-k-Neighborhood-Index [rand]** | | | | | **TI-k-Neighborhood-Index [max]** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,199 | 0,195 | 0,000 | 0,004 | 0,000 | 0,204 | 0,204 | 0,000 | 0,000 | 0,000 |
| 2000 | 0,664 | 0,660 | 0,000 | 0,004 | 0,000 | 0,678 | 0,674 | 0,000 | 0,004 | 0,000 |
| 4000 | 2,753 | 2,743 | 0,000 | 0,009 | 0,001 | 2,824 | 2,813 | 0,000 | 0,010 | 0,001 |
| 6000 | 6,576 | 6,560 | 0,000 | 0,014 | 0,002 | 6,887 | 6,870 | 0,000 | 0,016 | 0,001 |
| 8000 | 12,612 | 12,587 | 0,000 | 0,023 | 0,002 | 13,228 | 13,202 | 0,000 | 0,023 | 0,002 |
| karypis\_review | 500 | 0,066 | 0,064 | 0,000 | 0,002 | 0,000 | 0,068 | 0,066 | 0,000 | 0,002 | 0,000 |
| 1000 | 0,247 | 0,244 | 0,000 | 0,003 | 0,000 | 0,249 | 0,246 | 0,000 | 0,003 | 0,000 |
| 2000 | 1,098 | 1,092 | 0,000 | 0,006 | 0,000 | 1,105 | 1,099 | 0,000 | 0,006 | 0,000 |
| 3000 | 2,394 | 2,385 | 0,000 | 0,008 | 0,001 | 2,394 | 2,385 | 0,000 | 0,008 | 0,001 |
| 4000 | 4,340 | 4,328 | 0,000 | 0,011 | 0,001 | 4,346 | 4,334 | 0,000 | 0,011 | 0,001 |
| covtype | 10000 | 0,545 | 0,433 | 0,000 | 0,109 | 0,003 | 0,547 | 0,391 | 0,000 | 0,153 | 0,003 |
| 50000 | 8,104 | 6,078 | 0,001 | 2,009 | 0,016 | 7,750 | 5,400 | 0,001 | 2,333 | 0,015 |
| 100000 | 23,603 | 18,911 | 0,001 | 4,654 | 0,036 | 19,989 | 15,052 | 0,002 | 4,900 | 0,035 |
| 300000 | 143,614 | 127,733 | 0,000 | 15,746 | 0,135 | 99,592 | 83,380 | 0,004 | 16,079 | 0,130 |
| 500000 | 402,611 | 376,226 | 0,005 | 26,114 | 0,265 | 261,458 | 234,340 | 0,005 | 26,861 | 0,252 |
| cup98 | 10000 | 0,317 | 0,224 | 0,000 | 0,093 | 0,000 | 0,392 | 0,272 | 0,000 | 0,117 | 0,003 |
| 30000 | 3,000 | 2,335 | 0,000 | 0,650 | 0,015 | 3,073 | 2,197 | 0,000 | 0,865 | 0,010 |
| 50000 | 7,555 | 5,881 | 0,000 | 1,653 | 0,021 | 7,291 | 5,371 | 0,001 | 1,900 | 0,019 |
| 70000 | 12,672 | 9,672 | 0,000 | 2,969 | 0,031 | 12,886 | 9,563 | 0,001 | 3,293 | 0,029 |
| 90000 | 19,957 | 15,465 | 0,000 | 4,446 | 0,046 | 19,820 | 15,316 | 0,002 | 4,462 | 0,040 |

Analiza wyników pozwoliła mi zauważyć, że różnice w czasach wykonania algorytmu na danych tekstowych są zdecydowanie mniejsze niż w przypadku pozostałych zbiorów. Dla zbioru karypis\_review różnice te są na tyle małe, że nie można na jego podstawie wskazać punktu referencyjnego, który najbardziej przyspiesza wyszukiwanie k sąsiedztwa. Badania wykonane na zbiorach cup98 i covtype pozwalają wyeliminować odpowiednio punkty [max\_min] i [rand] z listy potencjalnych kandydatów punktów najbardziej przyspieszających wykonanie algorytmu. Punkt maksymalny pozwala osiągać jedne z najlepszych rezultatów w przypadku wszystkich zbiorów poza karypis\_sport, dla którego algorytm wykonuje się najdłużej z wszystkich badanych punktów referencyjnych. Punktem, dla którego w znakomitej większości eksperymentów wyszukiwanie k-sąsiedztwa wykonuje się najszybciej jest punkt minimalny.

Rys. 14. Porównanie wydajności algorytmu TI-k-Neighborhood-Index w zależności od wybranego punktu referencyjnego na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

W dalszej części pracy jako wyniki czasowe algorytmu *TI-k-Neighborhood-Index* będą prezentowane rezultaty osiągnięte przy zastosowaniu punktu minimalnego jako punktu referencyjnego.

### 7.2.2. Badania algorytmu k-Neighborhood-Index-Projection

Dla algorytmu *k-Neighborhood-Index-Projection* badałem następujące rodzaje rzutowania:

* rzutowanie na wymiar o najliczniejszej dziedzinie – oznaczane *[dmax]*;
* rzutowanie na niezerowy wymiar o najmniej licznej dziedzinie – oznaczane *[dmin]*;
* rzutowanie na losowy wymiar – oznaczane *[dmax]*;

W Tab. 10 zamieściłem czasy uruchomień algorytmu *k-Neighborhood-Index-Projection* wraz z trwaniem składających się na niego kroków. Na Rys. 15 znajdują się wykresy czasu wykonania algorytmu w funkcji liczby punktów.

Rys. 15. Porównanie wydajności algorytmu k-Neighborhood-Index-Projection w zależności od wymiaru rzutowania na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

Tab. 10. Porównanie wydajności algorytmu *k-Neighborhood-Index-Projection* w zależności od wybranego wymiaru rzutowania na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **k-Neighborhood-Index-Projection [dmax]** | | | | | **k-Neighborhood-Index-Projection [drand]** | | | | | **k-Neighborhood-Index-Projection [dmin]** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. rz.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. rz.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. rz.** | **sort.** |
|
| karypis\_sport | 1000 | 0,214 | 0,213 | 0,000 | 0,001 | 0,000 | 0,215 | 0,214 | 0,000 | 0,000 | 0,000 | 0,214 | 0,214 | 0,000 | 0,000 | 0,000 |
| 2000 | 0,809 | 0,807 | 0,000 | 0,002 | 0,000 | 0,810 | 0,809 | 0,000 | 0,001 | 0,000 | 0,812 | 0,811 | 0,000 | 0,001 | 0,000 |
| 4000 | 3,436 | 3,432 | 0,000 | 0,004 | 0,000 | 3,433 | 3,429 | 0,000 | 0,004 | 0,000 | 3,448 | 3,444 | 0,000 | 0,003 | 0,000 |
| 6000 | 7,830 | 7,823 | 0,000 | 0,006 | 0,000 | 7,829 | 7,822 | 0,000 | 0,007 | 0,000 | 7,850 | 7,844 | 0,000 | 0,006 | 0,000 |
| 8000 | 14,077 | 14,068 | 0,000 | 0,008 | 0,000 | 14,125 | 14,116 | 0,000 | 0,008 | 0,000 | 14,128 | 14,122 | 0,000 | 0,006 | 0,000 |
| karypis\_review | 500 | 0,073 | 0,073 | 0,000 | 0,000 | 0,000 | 0,073 | 0,073 | 0,000 | 0,000 | 0,000 | 0,074 | 0,074 | 0,000 | 0,000 | 0,000 |
| 1000 | 0,288 | 0,287 | 0,000 | 0,001 | 0,000 | 0,289 | 0,289 | 0,000 | 0,000 | 0,000 | 0,289 | 0,289 | 0,000 | 0,001 | 0,000 |
| 2000 | 1,258 | 1,257 | 0,000 | 0,001 | 0,000 | 1,259 | 1,258 | 0,000 | 0,001 | 0,000 | 1,263 | 1,261 | 0,000 | 0,001 | 0,000 |
| 3000 | 2,813 | 2,811 | 0,000 | 0,002 | 0,000 | 2,822 | 2,820 | 0,000 | 0,002 | 0,000 | 2,826 | 2,824 | 0,000 | 0,002 | 0,000 |
| 4000 | 5,088 | 5,085 | 0,000 | 0,003 | 0,000 | 5,074 | 5,071 | 0,000 | 0,003 | 0,000 | 5,079 | 5,076 | 0,000 | 0,003 | 0,000 |
| covtype | 10000 | 0,440 | 0,320 | 0,000 | 0,118 | 0,002 | 1,412 | 1,311 | 0,000 | 0,101 | 0,000 | 1,473 | 1,386 | 0,000 | 0,085 | 0,002 |
| 50000 | 6,376 | 4,026 | 0,001 | 2,335 | 0,013 | 45,292 | 43,381 | 0,001 | 1,907 | 0,004 | 58,654 | 56,548 | 0,001 | 2,097 | 0,008 |
| 100000 | 14,753 | 9,833 | 0,001 | 4,891 | 0,028 | 164,137 | 159,576 | 0,002 | 4,557 | 0,002 | 243,611 | 238,447 | 0,002 | 5,145 | 0,018 |
| 300000 | 97,557 | 82,223 | 0,003 | 15,230 | 0,101 | 1318,407 | 1302,450 | 0,003 | 15,947 | 0,006 | 2032,951 | 2016,690 | 0,003 | 16,179 | 0,079 |
| 500000 | 283,132 | 256,854 | 0,005 | 26,091 | 0,182 | 3559,763 | 3532,690 | 0,005 | 27,013 | 0,055 | 5290,897 | 5263,930 | 0,005 | 26,815 | 0,147 |
| cup98 | 10000 | 0,357 | 0,239 | 0,000 | 0,116 | 0,002 | 1,648 | 1,537 | 0,000 | 0,110 | 0,001 | 1,483 | 1,379 | 0,000 | 0,103 | 0,001 |
| 30000 | 2,756 | 2,094 | 0,000 | 0,653 | 0,008 | 27,807 | 27,113 | 0,001 | 0,690 | 0,004 | 20,347 | 19,668 | 0,001 | 0,677 | 0,002 |
| 50000 | 6,906 | 5,230 | 0,001 | 1,660 | 0,015 | 76,375 | 74,429 | 0,001 | 1,940 | 0,006 | 55,897 | 54,189 | 0,001 | 1,703 | 0,003 |
| 70000 | 13,061 | 10,041 | 0,001 | 2,997 | 0,022 | 145,992 | 142,618 | 0,001 | 3,362 | 0,011 | 111,705 | 108,640 | 0,001 | 3,059 | 0,004 |
| 90000 | 19,727 | 15,146 | 0,001 | 4,549 | 0,032 | 242,929 | 238,615 | 0,001 | 4,300 | 0,012 | 184,648 | 180,108 | 0,001 | 4,533 | 0,006 |

Różnice w czasach wykonania algorytmu dla strategii rzutowania wykonanych na zbiorach tekstowych są na tyle niewielkie, że nie pozwalają na wyciągnięcie wniosków na temat użyteczności danej metody. Rezultaty eksperymentów wykonanych na gęstych zbiorach danych świadczą, że projekcja na wymiar o najszerszej dziedzinie pozwala na przyspieszenie wyznaczania k sąsiedztwa o rząd wielkości w porównaniu do pozostałych strategii rzutowania. Liczność dziedziny wymiaru, na który wykonywana jest projekcja ma kluczowy wpływ na sprawność algorytmu *k-Neighborhood-Index-Projection*. Następujące wyjaśnienie jest oparte na Eps-sąsiedztwie bez straty wartości merytorycznej dla k-sąsiedztwa ponieważ problem k sąsiedztwa można sprowadzić do problemu Eps-sąsiedztwa.

Na Rys. 16 przedstawiono przykładowy zbiór dwuwymiarowej przestrzeni (, ). Przerywanymi liniami zaznaczono rzuty punktów zbioru odpowiednio na wymiary i . Zbiór punktów () będących wynikiem rzutowania zbioru na ma liczność 3, natomiast zbiór D2 () powstały w wyniku rzutowania na jest liczności 8.



Rys. 16. Zbiór punktów Z

Załóżmy, że szukamy pewnego otoczenia epsilonowego punktu (Na Rys. 16 epsilonowe otoczenie zostało oznaczone fragmentem okręgu). Gdy posłużymy się projekcją na to żaden punkt zbioru nie zostanie odrzucony w procesie wyznaczania potencjalnych sąsiadów na podstawie kryterium rzutowania na dany wymiar, ponieważ wszystkie punkty zbioru należą do otoczenia epsilonowego punktu powstałego po projekcji na (Krawędzie otoczenia epsilonowego rzutów punktu na wymiary i zostały oznaczone na Rys. 16 odpowiednio szarymi ciągłymi liniami pionowymi i linią poziomą.). Tym samym zastosowanie rzutowania w tym szczególnym przypadku nie przyspiesza wyznaczania sąsiedztwa. Gdy posłużymy się rzutowaniem na to aż połowa punktów zbioru zostanie odrzuconych w procesie wyznaczania potencjalnych sąsiadów na podstawie kryterium rzutowania na dany wymiar co znacząco przyspieszy wyznaczanie sąsiedztwa.

W algorytmie *k-Neighborhood-Index-Projection* rzutowanie tym mocniej wspiera selektywność wyznaczania potencjalnych sąsiadów im liczniejsza jest dziedzina wymiaru, na który wykonywana jest projekcja. Dlatego, w dalszej części pracy jako wyniki czasowe algorytmu *k-Neighborhood-Index-Projection* będą prezentowane rezultaty osiągnięte przy zastosowaniu rzutowania na wymiar [dmax].

### 7.2.3. Porównanie algorytmów k-Neighborhood-Index-Projection i TI-k-Neighborhood-Index

W celu porównania podejścia korzystającego z nierówności trójkąta z podejściem stosującym rzutowanie na Rys. 17 zamieściłem wykresy czasów wykonania algorytmów *k-Neighborhood-Index-Projection* i *TI-k-Neighborhood-Index* w funkcji liczby punktów we wszystkich badanych przypadkach zastosowania punktu referencyjnego i strategii rzutowania na wymiar. Dane zamieszczone na wykresach odpowiadają tym zgromadzonym w tabelach Tab. 8, Tab. 9,  
 Tab. 10.

Z rezultatów eksperymentów wynika, że dla danych tekstowych algorytm  
*k-neighborhood-Index-Projection* wykorzystujący rzutowanie wykonuje się nieznacznie wolniej niż *TI-k-Neighborhood-Index*. W przypadku gęstych zbiorów wyszukiwanie sąsiedztwa z zastosowaniem rzutowania na [drand] i [dmin] realizuje się o rząd wielkości wolniej niż w pozostałych przypadkach. Natomiast sprawność *k-neighborhood-Index-Projection* z rzutowaniem na [dmax] jest porównywalna z *TI-k-Neighborhood-Index* dla badanych punktów referencyjnych. Na niekorzyść rzutowania w porównaniu do punktu referencyjnego przemawia zależność tej metody od liczności dziedzin wymiarów.

Rys. 17. Porównanie wydajności algorytmu k-Neighborhood-Index-Projection z TI-k-Neighborhood-Index w zależności od wybranego wymiaru rzutowania i punktu referencyjnego na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

### 7.2.4. Badanie algorytmu TI-k-Neighborhood-Index-Ref – wybór dwóch punktów referencyjnych

W celu zbadania algorytmu *TI-k-Neighborhood-Index* przeprowadziłem eksperymenty testując różne pary punktów referencyjnych. W tabelach Tab. 11 i Tab. 12 zamieściłem czasy uruchomień algorytmu *TI-k-Neighborhood-Index* *-Ref* wraz z trwaniem składających się na niego kroków. Na Rys. 18 znajdują się wykresy czasu wykonania algorytmu w funkcji liczby punktów.

Rys. 18. Porównanie wydajności algorytmu TI-k-Neighborhood-Index-Ref w zależności od wybranej pary punktów referencyjnych na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

Tab. 11. Porównanie wydajności algorytmu TI-k-Neighborhood-Index-Ref w zależności od wybranej pary punktów referencyjnych na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **TI-k-Neighborhood-Index-Ref [max][min]** | | | | | **TI-k-Neighborhood-Index-Ref [max][max\_min]** | | | | | **TI-k-Neighborhood-Index-Ref [max][rand]** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,176 | 0,171 | 0,000 | 0,005 | 0,000 | 0,202 | 0,196 | 0,000 | 0,006 | 0,000 | 0,202 | 0,194 | 0,000 | 0,008 | 0,000 |
| 2000 | 0,634 | 0,629 | 0,000 | 0,005 | 0,000 | 0,647 | 0,638 | 0,000 | 0,008 | 0,000 | 0,663 | 0,655 | 0,000 | 0,008 | 0,000 |
| 4000 | 2,694 | 2,678 | 0,000 | 0,016 | 0,000 | 2,757 | 2,738 | 0,000 | 0,018 | 0,001 | 2,796 | 2,775 | 0,000 | 0,020 | 0,001 |
| 6000 | 6,427 | 6,396 | 0,000 | 0,031 | 0,000 | 6,646 | 6,617 | 0,000 | 0,028 | 0,002 | 6,927 | 6,911 | 0,000 | 0,015 | 0,002 |
| 8000 | 11,783 | 11,726 | 0,000 | 0,057 | 0,000 | 12,563 | 12,483 | 0,000 | 0,077 | 0,003 | 12,656 | 12,589 | 0,000 | 0,067 | 0,000 |
| karypis\_review | 500 | 0,062 | 0,062 | 0,000 | 0,000 | 0,000 | 0,067 | 0,064 | 0,000 | 0,003 | 0,000 | 0,062 | 0,062 | 0,000 | 0,000 | 0,000 |
| 1000 | 0,244 | 0,239 | 0,000 | 0,005 | 0,000 | 0,249 | 0,243 | 0,000 | 0,006 | 0,000 | 0,260 | 0,250 | 0,000 | 0,010 | 0,000 |
| 2000 | 1,092 | 1,076 | 0,000 | 0,016 | 0,000 | 1,103 | 1,092 | 0,000 | 0,011 | 0,000 | 1,087 | 1,076 | 0,000 | 0,010 | 0,000 |
| 3000 | 2,381 | 2,366 | 0,000 | 0,015 | 0,000 | 2,418 | 2,401 | 0,000 | 0,016 | 0,001 | 2,398 | 2,377 | 0,000 | 0,016 | 0,005 |
| 4000 | 4,238 | 4,212 | 0,000 | 0,026 | 0,000 | 4,303 | 4,280 | 0,000 | 0,022 | 0,001 | 4,285 | 4,264 | 0,000 | 0,021 | 0,000 |
| covtype | 10000 | 0,515 | 0,276 | 0,000 | 0,229 | 0,010 | 0,461 | 0,294 | 0,000 | 0,164 | 0,003 | 0,484 | 0,317 | 0,000 | 0,167 | 0,000 |
| 50000 | 6,661 | 4,227 | 0,000 | 2,418 | 0,015 | 6,775 | 4,340 | 0,001 | 2,419 | 0,015 | 6,682 | 4,222 | 0,000 | 2,444 | 0,016 |
| 100000 | 16,260 | 10,956 | 0,005 | 5,268 | 0,031 | 16,161 | 10,943 | 0,002 | 5,180 | 0,035 | 15,678 | 10,380 | 0,000 | 5,262 | 0,036 |
| 300000 | 65,057 | 48,391 | 0,005 | 16,526 | 0,135 | 64,109 | 47,364 | 0,004 | 16,609 | 0,132 | 61,146 | 44,642 | 0,000 | 16,374 | 0,130 |
| 500000 | 144,289 | 115,575 | 0,005 | 28,454 | 0,255 | 142,420 | 113,920 | 0,005 | 28,234 | 0,260 | 135,617 | 107,516 | 0,005 | 27,841 | 0,255 |
| cup98 | 10000 | 0,406 | 0,271 | 0,000 | 0,135 | 0,000 | 0,411 | 0,244 | 0,000 | 0,164 | 0,003 | 0,405 | 0,265 | 0,000 | 0,140 | 0,000 |
| 30000 | 2,783 | 1,919 | 0,000 | 0,848 | 0,016 | 2,831 | 1,953 | 0,000 | 0,867 | 0,010 | 2,876 | 2,028 | 0,000 | 0,832 | 0,015 |
| 50000 | 7,057 | 5,070 | 0,000 | 1,971 | 0,016 | 6,952 | 4,885 | 0,001 | 2,047 | 0,019 | 6,905 | 4,898 | 0,000 | 1,991 | 0,016 |
| 70000 | 12,397 | 8,829 | 0,000 | 3,536 | 0,031 | 12,060 | 8,416 | 0,001 | 3,612 | 0,031 | 12,157 | 8,429 | 0,000 | 3,697 | 0,031 |
| 90000 | 16,552 | 12,210 | 0,000 | 4,300 | 0,042 | 16,250 | 11,939 | 0,001 | 4,268 | 0,041 | 16,312 | 12,074 | 0,000 | 4,196 | 0,042 |

Tab. 12. Porównanie wydajności algorytmu TI-k-Neighborhood-Index-Ref w zależności od wybranej pary punktów referencyjnych na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10 % losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **TI-k-Neighborhood-Index-Ref [min][max]** | | | | | **TI-k-Neighborhood-Index-Ref [max\_min][max]** | | | | | **TI-k-Neighborhood-Index-Ref [rand][max]** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,174 | 0,168 | 0,000 | 0,005 | 0,001 | 0,202 | 0,196 | 0,000 | 0,006 | 0,000 | 0,202 | 0,194 | 0,000 | 0,008 | 0,000 |
| 2000 | 0,640 | 0,632 | 0,000 | 0,008 | 0,000 | 0,648 | 0,640 | 0,000 | 0,008 | 0,000 | 0,662 | 0,654 | 0,000 | 0,008 | 0,000 |
| 4000 | 2,719 | 2,698 | 0,000 | 0,019 | 0,001 | 2,746 | 2,726 | 0,000 | 0,019 | 0,001 | 2,752 | 2,732 | 0,000 | 0,019 | 0,001 |
| 6000 | 6,439 | 6,410 | 0,000 | 0,027 | 0,002 | 6,582 | 6,552 | 0,000 | 0,029 | 0,002 | 6,623 | 6,591 | 0,000 | 0,030 | 0,002 |
| 8000 | 11,795 | 11,744 | 0,000 | 0,048 | 0,002 | 12,589 | 12,483 | 0,000 | 0,103 | 0,002 | 12,684 | 12,626 | 0,000 | 0,055 | 0,003 |
| karypis\_review | 500 | 0,064 | 0,061 | 0,000 | 0,003 | 0,000 | 0,067 | 0,064 | 0,000 | 0,003 | 0,000 | 0,068 | 0,064 | 0,000 | 0,004 | 0,000 |
| 1000 | 0,245 | 0,239 | 0,000 | 0,005 | 0,001 | 0,248 | 0,242 | 0,000 | 0,006 | 0,000 | 0,250 | 0,243 | 0,000 | 0,006 | 0,000 |
| 2000 | 1,088 | 1,077 | 0,000 | 0,011 | 0,000 | 1,107 | 1,095 | 0,000 | 0,012 | 0,000 | 1,105 | 1,092 | 0,000 | 0,012 | 0,001 |
| 3000 | 2,393 | 2,377 | 0,000 | 0,016 | 0,000 | 2,471 | 2,454 | 0,000 | 0,016 | 0,001 | 2,479 | 2,461 | 0,000 | 0,017 | 0,001 |
| 4000 | 4,254 | 4,232 | 0,000 | 0,021 | 0,001 | 4,345 | 4,322 | 0,000 | 0,022 | 0,001 | 4,347 | 4,323 | 0,000 | 0,023 | 0,001 |
| covtype | 10000 | 0,481 | 0,338 | 0,000 | 0,140 | 0,003 | 0,596 | 0,381 | 0,000 | 0,211 | 0,003 | 0,571 | 0,342 | 0,000 | 0,226 | 0,003 |
| 50000 | 6,968 | 4,470 | 0,001 | 2,481 | 0,015 | 6,922 | 4,477 | 0,001 | 2,427 | 0,016 | 7,157 | 4,594 | 0,001 | 2,546 | 0,016 |
| 100000 | 16,343 | 10,977 | 0,002 | 5,328 | 0,035 | 16,451 | 11,038 | 0,002 | 5,374 | 0,036 | 16,581 | 11,310 | 0,002 | 5,232 | 0,037 |
| 300000 | 64,407 | 47,668 | 0,003 | 16,601 | 0,135 | 65,557 | 48,845 | 0,003 | 16,571 | 0,137 | 69,293 | 52,593 | 0,003 | 16,551 | 0,145 |
| 500000 | 140,108 | 111,918 | 0,005 | 27,926 | 0,259 | 143,659 | 115,806 | 0,005 | 27,585 | 0,262 | 144,438 | 116,489 | 0,005 | 27,660 | 0,285 |
| cup98 | 10000 | 0,397 | 0,254 | 0,000 | 0,140 | 0,003 | 0,440 | 0,291 | 0,000 | 0,146 | 0,003 | 0,371 | 0,251 | 0,001 | 0,117 | 0,003 |
| 30000 | 2,749 | 1,845 | 0,001 | 0,892 | 0,011 | 3,106 | 2,205 | 0,001 | 0,890 | 0,011 | 2,935 | 2,109 | 0,000 | 0,815 | 0,011 |
| 50000 | 6,734 | 4,939 | 0,000 | 1,776 | 0,019 | 7,575 | 5,525 | 0,001 | 2,030 | 0,019 | 7,049 | 5,107 | 0,001 | 1,922 | 0,019 |
| 70000 | 11,994 | 8,572 | 0,001 | 3,392 | 0,029 | 12,877 | 9,588 | 0,001 | 3,258 | 0,030 | 12,175 | 8,987 | 0,001 | 3,157 | 0,030 |
| 90000 | 16,410 | 12,269 | 0,001 | 4,098 | 0,041 | 18,446 | 14,035 | 0,001 | 4,367 | 0,042 | 17,986 | 13,539 | 0,001 | 4,404 | 0,042 |

Z rezultatów przeprowadzonych eksperymentów trudno jednoznacznie wskazać, która kombinacja punktów referencyjnych najbardziej przyspiesza wyznaczenie k sąsiedztwa. Wyniki świadczą, że najkorzystniej jako pierwszy punkt referencyjny wybrać [max], ponieważ w większości przypadków eksperymenty z punktem maksymalnym jako pierwszym punktem referencyjnym wykonują się szybciej niż dla punktu losowego [rand] czy innego punktu skrajnego [min].

Różnice w czasach wykonania algorytmu dla poszczególnych zestawów punktów referencyjnych są na tyle niewielkie, że nie pozwalają jednoznacznie stwierdzić, która z badanych kombinacji punktów referencyjnych najbardziej przyspiesza wykonanie algorytmu. Tym co wyniki badań pozwalają stwierdzić jest wniosek, że jako pierwszy punkt referencyjny najlepiej jest wybrać punkt maksymalny.

Ponieważ dla algorytmu *TI-k-Neighborhood-Index* najlepszym punktem referencyjnym okazał się być [min], dlatego w dalszej części pracy jako wyniki czasowe algorytmu *TI-k-Neighborhood-Index-Ref* będą prezentowane rezultaty osiągnięte przy zastosowaniu pary punktów referencyjnych [max][min].

### 7.2.5. Porównanie implementacji k-Neighborhood-Index

W Tab. 13, Tab. 14 i na Rys. 19 przedstawiono rezultaty badań wariacji na temat algorytmu  
*k-Neighborhood-Index*. Wyniki algorytmu *TI-k-Neighborhood-Index* zostały zebrane dla punktu referencyjnego [min], *k-Neighborhood-Index-Projection* dla [dmax], natomiast rezultaty   
*TI-k-Neighborhood-Ref* dla punktów referencyjnych [max][min]. Algorytm *k-Neighborhood-Index-Brutal* jest brutalną implementacją wyszukiwania k sąsiedztwa o złożoności kwadratowej.

Rezultaty jednoznacznie wskazują wzrosty wydajności *TI-k-Neighborhood-Index*,  
*TI-k-Neighborhood-Index-Ref* oraz *k-Neighborhood-Index-Projection* w stosunku do algorytmu  
*k-Neighborhood-Index-Brutal*. Zastosowanie nierówności trójkąta pozwala uzyskiwać wyniki o dwa rzędy wielkości szybciej. Wzrost ten jest mniej widoczny dla danych tekstowych niż dla pozostałych zbiorów z uwagi na ich rzadki charakter. Zwiększenie liczby punktów referencyjnych z jednego do dwóch nie przyniosło znaczącej poprawy sprawności wyznaczania k sąsiedztwa. Dalsze zwiększanie liczby punktów referencyjnych może spowolnić wykonanie algorytmu, ponieważ koszt obsługi wielu punktów referencyjnych może przewyższyć zysk z ich zastosowania.

Tab. 13. Porównanie wydajności odmian algorytmu k-Neighborhood-Index na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **zbioór** | **l. p.** | **k-Neighborhood-Index-Brutal** | | | |  | **TI-k-Neighborhood-Index** | | | |  |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,322 | 0,320 | 0,000 | 0,002 | 0,000 | 0,171 | 0,169 | 0,000 | 0,002 | 0,000 |
| 2000 | 1,397 | 1,394 | 0,000 | 0,004 | 0,000 | 0,636 | 0,632 | 0,000 | 0,004 | 0,000 |
| 4000 | 6,014 | 6,005 | 0,000 | 0,008 | 0,001 | 2,684 | 2,675 | 0,000 | 0,008 | 0,001 |
| 6000 | 15,178 | 15,162 | 0,000 | 0,014 | 0,002 | 6,470 | 6,455 | 0,000 | 0,013 | 0,002 |
| 8000 | 30,516 | 30,497 | 0,000 | 0,016 | 0,002 | 11,770 | 11,751 | 0,000 | 0,017 | 0,002 |
| karypis\_review | 500 | 0,093 | 0,092 | 0,000 | 0,001 | 0,000 | 0,062 | 0,061 | 0,000 | 0,001 | 0,000 |
| 1000 | 0,385 | 0,383 | 0,000 | 0,002 | 0,000 | 0,241 | 0,239 | 0,000 | 0,002 | 0,000 |
| 2000 | 2,631 | 2,626 | 0,000 | 0,005 | 0,000 | 1,084 | 1,079 | 0,000 | 0,005 | 0,000 |
| 3000 | 9,183 | 9,175 | 0,000 | 0,007 | 0,001 | 2,356 | 2,348 | 0,000 | 0,007 | 0,001 |
| 4000 | 9,803 | 9,793 | 0,000 | 0,009 | 0,001 | 4,209 | 4,198 | 0,000 | 0,010 | 0,001 |
| covtype | 10000 | 40,337 | 40,115 | 0,000 | 0,219 | 0,003 | 0,551 | 0,370 | 0,000 | 0,179 | 0,002 |
| 50000 | 844,863 | 842,475 | 0,001 | 2,372 | 0,015 | 7,804 | 5,526 | 0,001 | 2,262 | 0,015 |
| 100000 | 4575,759 | 4570,870 | 0,002 | 4,851 | 0,036 | 19,962 | 15,025 | 0,001 | 4,901 | 0,035 |
| 300000 | - | - | - | - | - | 93,868 | 78,047 | 0,003 | 15,686 | 0,132 |
| 500000 | - | - | - | - | - | 246,482 | 219,916 | 0,005 | 26,312 | 0,249 |
| cup98 | 10000 | 16,170 | 15,947 | 0,000 | 0,219 | 0,003 | 0,388 | 0,259 | 0,000 | 0,126 | 0,003 |
| 30000 | 164,166 | 163,038 | 0,001 | 1,116 | 0,011 | 2,785 | 2,105 | 0,001 | 0,669 | 0,010 |
| 50000 | 475,873 | 473,576 | 0,001 | 2,276 | 0,019 | 6,942 | 5,287 | 0,001 | 1,635 | 0,019 |
| 70000 | 996,087 | 992,978 | 0,001 | 3,078 | 0,030 | 13,095 | 9,933 | 0,001 | 3,132 | 0,029 |
| 90000 | 1717,090 | 1712,660 | 0,002 | 4,387 | 0,041 | 19,041 | 14,683 | 0,001 | 4,316 | 0,041 |

Tab. 14. Porównanie wydajności odmian algorytmu k-Neighborhood-Index na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| zbioór | **l. p.** | **k-Neighborhood-Index-Projection** | | | |  | **TI-k-Neighborhood-Index-Ref** | | | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. rz.** | **sort.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **obl. odl.** | **sort.** |
|
| karypis\_sport | 1000 | 0,214 | 0,213 | 0,000 | 0,001 | 0,000 | 0,176 | 0,171 | 0,000 | 0,005 | 0,000 |
| 2000 | 0,809 | 0,807 | 0,000 | 0,002 | 0,000 | 0,634 | 0,629 | 0,000 | 0,005 | 0,000 |
| 4000 | 3,436 | 3,432 | 0,000 | 0,004 | 0,000 | 2,694 | 2,678 | 0,000 | 0,016 | 0,000 |
| 6000 | 7,830 | 7,823 | 0,000 | 0,006 | 0,000 | 6,427 | 6,396 | 0,000 | 0,031 | 0,000 |
| 8000 | 14,077 | 14,068 | 0,000 | 0,008 | 0,000 | 11,783 | 11,726 | 0,000 | 0,057 | 0,000 |
| karypis\_review | 500 | 0,073 | 0,073 | 0,000 | 0,000 | 0,000 | 0,062 | 0,062 | 0,000 | 0,000 | 0,000 |
| 1000 | 0,288 | 0,287 | 0,000 | 0,001 | 0,000 | 0,244 | 0,239 | 0,000 | 0,005 | 0,000 |
| 2000 | 1,258 | 1,257 | 0,000 | 0,001 | 0,000 | 1,092 | 1,076 | 0,000 | 0,016 | 0,000 |
| 3000 | 2,813 | 2,811 | 0,000 | 0,002 | 0,000 | 2,381 | 2,366 | 0,000 | 0,015 | 0,000 |
| 4000 | 5,088 | 5,085 | 0,000 | 0,003 | 0,000 | 4,238 | 4,212 | 0,000 | 0,026 | 0,000 |
| covtype | 10000 | 0,440 | 0,320 | 0,000 | 0,118 | 0,002 | 0,515 | 0,276 | 0,000 | 0,229 | 0,010 |
| 50000 | 6,376 | 4,026 | 0,001 | 2,335 | 0,013 | 6,661 | 4,227 | 0,000 | 2,418 | 0,015 |
| 100000 | 14,753 | 9,833 | 0,001 | 4,891 | 0,028 | 16,260 | 10,956 | 0,005 | 5,268 | 0,031 |
| 300000 | 97,557 | 82,223 | 0,003 | 15,230 | 0,101 | 65,057 | 48,391 | 0,005 | 16,526 | 0,135 |
| 500000 | 283,132 | 256,854 | 0,005 | 26,091 | 0,182 | 144,289 | 115,575 | 0,005 | 28,454 | 0,255 |
| cup98 | 10000 | 0,357 | 0,239 | 0,000 | 0,116 | 0,002 | 0,406 | 0,271 | 0,000 | 0,135 | 0,000 |
| 30000 | 2,756 | 2,094 | 0,000 | 0,653 | 0,008 | 2,783 | 1,919 | 0,000 | 0,848 | 0,016 |
| 50000 | 6,906 | 5,230 | 0,001 | 1,660 | 0,015 | 7,057 | 5,070 | 0,000 | 1,971 | 0,016 |
| 70000 | 13,061 | 10,041 | 0,001 | 2,997 | 0,022 | 12,397 | 8,829 | 0,000 | 3,536 | 0,031 |
| 90000 | 19,727 | 15,146 | 0,001 | 4,549 | 0,032 | 16,552 | 12,210 | 0,000 | 4,300 | 0,042 |

Rys. 19. Porównanie wydajności odmian algorytmu k-Neighborhood-Index na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

Na Rys. 20 przedstawiono rezultaty z pominięcie algorytmu *k-Neighborhood-Index* dla uwidocznienia różnic między pozostałymi eksperymentami. Wyniki wykazują, że zastosowanie rzutowania nie przyspiesza wyszukiwania k sąsiedztwa bardziej niż wykorzystanie nierówności trójkąta.

Rys. 20. Porównanie wydajności odmian algorytmu k-Neighborhood-Index na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach danych dla 10% losowo wybranych punktów

## 7.3. Badanie algorytmu k-Neighborhood-Index-Vp-Tree

W następującym rozdziale przedstawiłem rezultaty badań implementacji algorytmu   
k-Neighborhood-Index-Vp-Tree. W swoich eksperymentach skupiłem się na metodach przeszukiwania indeksu metrycznego w celu wyznaczenia k sąsiedztwa danego zapytania. Badałem również wpływ implementacji punktu na wydajność algorytmu w zależności od charakteru zbioru danych, na którym działał.

### 7.3.1. Implementacja algorytmu

Na przykładowych zbiorach danych przetestowałem dwie metody wyszukiwania k sąsiedztwa w oparciu o indeks metryczny:

* pierwszą, której kryterium przeszukiwania kolejnych gałęzi indeksu metrycznego stanowi mediana odległości punktów do punktu obserwacyjnego, zwaną dalej *metodą mediany*;
* drugą, której kryterium przeszukiwania kolejnych gałęzi indeksu metrycznego stanowią lewe i prawe ograniczenie, zwaną dalej *metodą ograniczeń*.

W ?? i na ?? zamieściłem czasy uruchomień algorytmu *k-Neighborhood-Index-Vp-Tree* dla obu metod wyszukiwania k sąsiedztwa. Z rezultatów badań wynika, że *metoda ograniczeń* zapewnia szybsze wyszukiwanie k sąsiedztwa niż *metoda mediany*. Największa różnica w czasach wykonania algorytmów występuje dla wyszukiwania k-sąsiedztwa spośród 500000 punktów zbioru covtype. Implementacja korzystająca z *metody ograniczeń* wykonuje się blisko 1.5 razy szybciej niż implementacja stosująca *metodę mediany*.

Wartym podkreślenia jest fakt, iż mimo, że w obu rozpatrywanych przypadkach indeks metryczny budowany był w ten sam sposób oraz mimo, że prezentowane są uśrednione wyniki, to różnice w czasie budowy drzewa są znaczące. Zachowanie to wynika z heurystyki zastosowanej w procesie wyboru punktu obserwacyjnego. Pewna kombinacja punktów obserwacyjnych pozwala zbudować drzewo szybciej niż inna, stąd różnice w czasie budowy drzewa.

Jednakże wariancja ta nie wpływa decydująco na różnice w czasie wykonania obu implementacji. Powodem tych różnic jest kryterium wyszukiwania. Następujące wyjaśnienie jest oparte na Eps-sąsiedztwie bez straty wartości merytorycznej dla k sąsiedztwa ponieważ problem k-sąsiedztwa można sprowadzić do problemu Eps-sąsiedztwa.

Tab. 15. Porównanie wydajności algorytmu k-Neighborhood-Index-VP-Tree w zależności od implementacji metody przeszukiwania indeksu metrycznego na przykładowych zbiorach danych. Tabela zawiera czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach dla 10% losowo wybranych punktów

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| zbioór | **l. p.** | **VP-Tree-Index mediana** | | | **VP-Tree-Index ograniczenia** | | |
| **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** | **wyk. alg.** | **wysz. sąsiad.** | **bud. ind.** |
|
| karypis\_sport | 1000 | 0,614 | 0,241 | 0,373 | 0,296 | 0,218 | 0,078 |
| 2000 | 1,562 | 0,883 | 0,678 | 0,988 | 0,822 | 0,166 |
| 4000 | 4,976 | 3,747 | 1,229 | 3,879 | 3,510 | 0,369 |
| 6000 | 10,276 | 8,483 | 1,793 | 8,684 | 8,107 | 0,577 |
| 8000 | 15,944 | 15,039 | 0,905 | 15,543 | 14,768 | 0,774 |
| karypis\_review | 500 | 0,392 | 0,084 | 0,308 | 0,270 | 0,078 | 0,192 |
| 1000 | 0,911 | 0,318 | 0,593 | 0,675 | 0,296 | 0,379 |
| 2000 | 2,530 | 1,385 | 1,145 | 1,986 | 1,326 | 0,660 |
| 3000 | 4,650 | 3,089 | 1,561 | 3,630 | 2,949 | 0,681 |
| 4000 | 7,344 | 5,462 | 1,882 | 5,840 | 5,247 | 0,593 |
| covtype | 10000 | 1,858 | 0,366 | 1,491 | 0,859 | 0,298 | 0,561 |
| 50000 | 9,037 | 3,029 | 6,008 | 5,010 | 2,241 | 2,770 |
| 100000 | 27,845 | 10,687 | 17,158 | 14,204 | 7,703 | 6,501 |
| 300000 | 126,437 | 67,943 | 58,494 | 106,525 | 70,103 | 36,422 |
| 500000 | 346,216 | 219,689 | 126,527 | 249,375 | 180,813 | 68,562 |
| cup98 | 10000 | 2,477 | 0,494 | 1,983 | 0,728 | 0,307 | 0,421 |
| 30000 | 8,068 | 4,054 | 4,014 | 4,178 | 2,812 | 1,366 |
| 50000 | 19,508 | 11,432 | 8,076 | 10,215 | 7,766 | 2,450 |
| 70000 | 29,371 | 18,886 | 10,485 | 20,500 | 16,790 | 3,710 |
| 90000 | 44,818 | 29,800 | 15,018 | 40,305 | 27,149 | 13,156 |

Na ?? przedstawiono sytuację znajdowania Eps-sąsiedztwa punktu P (obszar pokryty czarno-białą kratą) w węźle indeksu metrycznego *metodą mediany*. Fragmentem okręgu zaznaczono medianę odległości punktów od punktu obserwacyjnego. Algorytm znajdowania Eps-sąsiedztwa punktu *metodą mediany* odwiedzając węzeł indeksu metrycznego oblicza odległość punktu obserwacyjnego od a następnie:

* przeszukuje lewe poddrzewo węzła indeksu metrycznego jeśli i
* przeszukuje lewe poddrzewo węzła indeksu metrycznego jeśli i ;
* przeszukuje lewe i prawe poddrzewo węzła indeksu metrycznego jeśli i ;

Zatem w przypadku przedstawionym na ?? *metoda mediany* przeszuka zarówno lewe jak i prawe poddrzewo mimo, że żaden z punktów prawego poddrzewa (, , , ) nie należy do Eps-sąsiedztwa punktu .

Rys. 21. Porównanie wydajności algorytmu k-neighborhood-Index-VP-Tree w zależności od implementacji metody przeszukiwania indeksu metrycznego na przykładowych zbiorach danych. Wykresy zawierają czasy wykonania (w sekundach) poszukiwań k=5 sąsiadów w przykładowych zbiorach dla 10% losowo wybranych punktów

?? przedstawia sytuację znajdowania Eps-sąsiedztwa punktu (obszar pokryty czarno-białą kratą) w węźle indeksu metrycznego *metodą ograniczeń.* Fragmentami okręgów zostały oznaczone ograniczenie górne i ograniczenie dolne . Algorytm znajdowania Eps-sąsiedztwa punktu  *metodą ograniczeń* odwiedzając węzeł indeksu oblicza odległość a następnie:

* przeszukuje lewe poddrzewo węzła indeksu metrycznego jeśli i ;
* przeszukuje prawe poddrzewo węzła indeksu metrycznego jeśli 0 ;
* przeszukuje lewe i prawe poddrzewo węzła VP indeksu metrycznego jeśli i ;

Zatem, w przypadku przedstawionym na ?? *metoda ograniczeń* przeszuka jedynie lewe poddrzewo *vp-drzewa*.

*Metoda ograniczeń* szybciej znajduje k-sąsiedztwo niż *metoda mediany*, ponieważ znajomość i pozwala w pewnych przypadkach wcześniej zdecydować o braku konieczności eksploracji aktualnie rozpatrywanej gałęzi indeksu metrycznegoniż *metoda mediany*.

# 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] | Yanilos P., "Data Struvtures and Algorithms of Nearest Neighbor Search in General Metric Spaces,". |
| [7] | Bozkaya T. and Ozsoyoglu M., "Distance based indexing for high-dimensional metric spaces,". |
| [8] | Kryszkiewicz M. and Skonieczny Ł., "Faster Clustering with DBSCAN," Gdańsk, Materiały z IIPWM'05, p. 605-614 2005. |
| [9] | US Forest Service. (2012, Marzec) The Forest CoverType dataset. [Online]. <http://ftp.ics.uci.edu/pub/machine-learning-databases/covtype/covtype.info> |
| [10] | The Second International Knowledge Discovery and Data Mining Tools Competition. (2012, Marzec) KDD Cup 1998 Data. [Online]. <http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html> |
| [11] | Karypis G. (2012, Marzec) The various datasets used in evaluating the performance of CLUTO's clustering algorithms. [Online]. <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download> |

x

1. kosinusowe sąsiedztwo jest sąsiedztwem opartym 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)