Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Informatyki

Rok akademicki 2012/2013

Praca dyplomowa magisterska

inż. Bartłomiej Jańczak

Tytuł pracy

Opiekun pracy:
prof. nzw. dr hab. Marzena Kryszkiewicz

Ocena .....

.....
Podpis Przewodniczącego
Komisji Egzaminu Dyplomowego

Specjalność: Inżynieria Systemów Informatycznych

Data urodzenia: 3 marca 1988 r.

Data rozpoczęcia studiów: 1 października 2011 r.

Życiorys

Nazywam się . . . .

. . . . . .
podpis studenta

Egzamin dyplomowy

Złożył egzamin dyplomowy w dn. . . . . .

Z wynikiem . . . . . .

Ogólny wynik studiów . . . . . .

Dodatkowe wnioski i uwagi Komisji . . . . . .

. . . . . .

Streszczenie

Praca ta prezentuje . . .

Słowa kluczowe: *słowa kluczowe.*

Abstract

Title: *Thesis title.*

This thesis describes . . .

Key words: *key words.*

Spis treści

1. Wprowadzenie

Współczesne systemy komputerowe agregują i generują ogromną ilość danych, która rośnie szybciej niż przewidywano jeszcze kilka lat temu. Ich gromadzenie i przechowywanie na nośnikach pamięci masowej nie stanowi problemu 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 wilekiej wartości, jednakże rozsądnie wykorzystane mogą stać się cennym źródłem szczególnej wiedzy. Jej odkrywaniem zajmuje się dziedzina informatyki zwana eksploracją danych, której ideą jest wykorzystanie komputera do znajdowania ukrytych dla człowieka prawidłowości w danych zgromadzonych w repozytoriach. Eksploracja danych jest czwartym etapem procesu odkrywania wiedzy, na który również składają się operacje selekcji, czyszczenia i transformacji danych a także analiza i interpretacja wyników.

Jednymi z najpopularniejszych metod eksploracji danych są grupowanie, klasyfikacja oraz odkrywanie asocjacji i sekwencji. Każda z metod odkrywa różnego rodzaju korelacje pomiędzy danymi, z czego wynika ich odmienne zastosowanie. W tym artykule skoncentrowano sie na zagadnieniu grupowania i klasyfikacji danych. Klasyfikacja definiowana jest jako problem przydzielenia danej obserwacji do jednej z kategorii w oparciu o zbiór obserwacji z nadanymi kategoriami zwany zbiorem trenującym. Algorytm implementujący klasyfikację nazywany jest klasyfikatorem. Grupowanie danych 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 łatwo o jej opis w literaturze. W przypadku algorytmów, na których skupiłem się w niniejszej pracy, wyjątkowo przydane okazują się artykuły naukowe.

Prawdopodobnie najpopularniejszym algorytmem gęstościowego grupowania danych jest DBSCAN ?? stanowiący często punkt odniesienia dla porównań z innymi algorytmami gęstościowych grupowań. [TODO]

Nową koncepcją zwiększenia wydajności wyżej wymienionych algorytmów jest wykorzystanie nierówności trójkąta do redukcji liczby kosztownych operacji wyznaczania podobieństwa obiektów. Na przykładnie algorytmu k-środków przedstawiane już były próby wykorzystania nierówności trójkąta w algorytmach grupowania danych. Natomiast po raz pierwszy została ona użyta w celu porządkowania dostępu do danych w algorytmach gęstościowego grupowania TI-DBSCAN ??, I-NBC ?? i PreDeCon ??. W pracach naukowych można również znaleźć wpływ

liczby punktów referencyjnych i stretegii ich wyboru na efektywność algorytmów ??. [TODO]

1.2. Motywacja i cel pracy

Grupowanie i klasyfikacja danych to procesy powszechnie stosowane 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 i klasyfikacji 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, to z kolei umożliwia przetwarzanie zboirów danych z większą liczbą obiektów lub atrybutów. Niekiedy może to oznaczać sposobność użycia tych algorytmów w nieosiągalnych dotychczas obszarach.

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

Celem pracy jest ... [TODO]

1.3. Układ pracy

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

2. Użyte algorytmy

Istnieje wiele rozwiązań problemu grupowania danych czyli wyznaczania zbiorów podobnych przy zachowaniu własciwości 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, jendak większość algorytmów jest niezależna od przyjętej miary podobieństwa. Liczność zastosowań grupowania częstokroć o odmiennych wymaganiach co do rezultatu oraz specyficznych danych wejściowych (np. o różnej liczności, rozkładzie bądź liczbie atrybutów) prowadzi do dużej liczby wyspecjalizowanych algorytmów. W każdym z nich można doszukać się wad oraz zalet, jednakże nie znaleziono dotychczas uniwersalnego algorytmu. Często trudno porównywać algorytmy grupowania danych ponieważ ze względu na charakterystyczne podejście do rozwiązywanego problemu różnią się one nie tylko sposobem grupowania ale także definicją grupy.

Najpopularniejsza klasyfikacja algorytmów grupowania dzieli je na algorytmy hierarchiczne i algorytmy oparte na podziale. Wynikiem pierwszej klasy są dendrogramy[1], których liście to klastry końcowe, a węzły reprezentują grupy. Relacja między węzłami a ich przodkami w dendrogramie odpowiada relacji między podgrupami a grupami. Dendrogramy powstają w wyniku rekurencyjnego podziału grup poczynając od zbioru danych lub odwrotnego procesu łączenia w grupy, począwszy od pojedynczych obiektów. W metodach opartych na podziale kluczowym elementem jest znalezienie najlepszego podziału zbioru na z góry zadaną liczbę możliwie najbardziej jednorodnych grup. Poczatkowy podział odpowiednio ze zdeterminowaną strategią optymalizowany jest w kolejnych iteracjach zgodnie z funkcją celu. Przykladami metod podziału są algorytm k-średnich i algorytm k-medoidów. Obie klasy algorytmów grupowania posiadają pewne wady. W przypadku algorytmów opartych na podziale jest to konieczność z góry zadania liczby grup. Z kolei w przypadku metod hierarchicznych głównym problemem jest określenie warunku stopu podzialu alby algorytm nie dzielił zbioru na niepotrzebnie małe lub duże grupy.

Wyniki wyżej wymienionych metod rzadko odpowiadają oczekiwaniam. taki stan rzeczy można tłumaczyć ninaturalnym dla człowieka mechanizmem grupowania. Gdyby zadać człowiekowi zadania pogrupowania punktów dwuwymiarowej przestrzeni okazałoby się, że nie dzieliłby on zbioru hierarchicznie na kolejne podzbiory czy też nie próbowałby podzielić go na zgóry określona liczbę podzbiorów. Ludzie grupuja punkty w zbiory o dowolnym kształcie, gdzie podstawą przydziału punktu do grupy jest odległość między punktami. Dokładniej rzecz ujmując do grupy należą punkty leżące w obszarze o gęstości wyraźnie większej niż w obszarze otaczającym ją. Tak zdefiniowanemu pojeciu metody grupowania najbliżej jest algorytmom gęstościowym, których przykładem jest DBSCAN opisany w kolejnym podrozdziale.

[1] Dendrogram to diagram stosowany do prezentacji związków między elementami lub grupami elementów w kształcie przypominający drzewo.

2.1. DBSCAN

Bibliografia

[1] Parsa I. The UCI KDD Archive. Kdd cup 1998 data. Luty 1999.

[2] Ozsoyoglu M. Bozkaya, T. Distance-based indexing for high-dimensional metric spaces. Raport instytutowy, Case Western Reserve University.

[3] Karypis G. The various datasets used in evaluating the performance of cluto's clustering algorithms. Sierpień 2002.

[4] Lasek P. Kryszkiewicz, M. Ti-dbscan: Clustering with dbscan by means of the triangular inequality. Raport instytutowy, Warsaw University of Tehcnology, Kwiecień 2010.

[5] P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metrics spaces. Raport instytutowy, The NEC Research Institute.