Implementacja wybranych algorytmów eksploracji danych na Oracle 10g

Sławomir Skowyra, Michał Rudowski

Instytut Informatyki Wydziału Elektroniki i Technik Informacyjnych,
Politechnika Warszawska
S.Skowyra@stud.elka.pw.edu.pl,
M.Rudowski@ii.pw.edu.pl

Streszczenie. Duże systemy baz danych i hurtownie danych zawierają w sobie bardzo użyteczną i niewidoczną dla człowieka wiedzę, ukrytą pod postacią wzorów, trendów, regularności i wyjątków. Tradycyjne metody analizy danych tracą zastosowanie, nie będąc w stanie przetworzyć bardzo dużych ilości gromadzonych danych. Spowodowało to rozwój dziedziny zwanej eksploracją danych (ang. data mining), obejmującej metody i algorytmy automatycznej analizy, takie jak: klasyfikacja, predykcja, regresja, określanie ważności atrybutów, grupowanie obiektów podobnych, znajdowanie reguł asocjacyjnych oraz eksploracja dokumentów tekstowych. Celem pracy jest analiza wybranych algorytmów, wsparcia jakie daje Oracle dla technik eksploracji danych, własna implementacja części z nich oraz skonfrontowanie wyników z narzędziami jakie dostarcza Microsoft Server 2005. Wynikiem pracy będzie porównanie wymienionych rozwiązań.

Słowa kluczowe: eksploracja danych, klasyfikacja, grupowanie, reguły asocjacyjne, regresja

1 Wstęp

Rozwój technologii systemów baz danych, magazynów danych, sieci komputerowych, automatycznych narzędzi do gromadzenia danych, z jednej strony, z drugiej, upowszechnienie systemów informatycznych związane ze wzrostem świadomości użytkowników i znaczącym spadkiem cen sprzętu komputerowego, zaowocowały nagromadzeniem olbrzymich wolumenów danych przechowywanych w bazach danych, hurtowniach danych i różnego rodzaju repozytoriach danych. Postęp technologiczny w zakresie cyfrowego generowania i gromadzenia informacji doprowadził do przekształcenia się baz danych wielu przedsiębiorstw, urzędów i placówek badawczych w zbiorniki ogromnych ilości danych. Odpowiedz na pytanie Śkąd biorą sie takie olbrzymie ilości danych?"jest bardzo prosta, codziennie w bankach, ubezpieczalniach, firmach, szpitalach, sieciach handlowych (nawet niewielkie supermarkety rejestrują codziennie sprzedaż tysięcy artykułów), wykonuje sie tysiące operacji handlowych (transakcje bankowe), raportów (sprzedaży) i gdzie generuje sie ogromne ilości danych eksperymentalnych w niemalże każdej

dziedzinie naukowej np. fizyka, astronomia, biologia, bioinformatyka itd. Niezbędna jest analiza przechowywanych danych, dzięki której można otrzymać informacje (ukrytą wiedzę) w nich zawartych. Inaczej przechowywanie ogromnych ilości danych i samo ich magazynowanie nie ma najmniejszego sensu. Odpowiedzią na potrzebę bardziej zaawansowanej i automatycznej analizy danych przechowywanych w bazach i hurtowniach danych jest technologia Eksploracji Danych (ang. Data Mining). Można postawić pytanie: 'Czym jest eksploracja danych?'. Zadaniem metod eksploracji danych jest automatyczne odkrywanie nietrywialnych, dotychczas nieznanych, zależności, związków, podobieństw lub trendów – ogólnie nazywanych wzorcami (ang. patterns) – w dużych repozytoriach danych. Odkrywane w procesie eksploracji danych wzorce mają, najczęściej postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień, wykresów, itp. Celem eksploracji najogólniej mówiąc jest analiza danych i procesów w celu lepszego ich poznania i zrozumienia. Automatyczna eksploracja danych otwiera nowe możliwości w zakresie interakcji użytkownika z systemem bazy i magazynem danych. Przede wszystkim umożliwia formułowanie zapytań na znacznie wyższym poziomie abstrakcji niż pozwala na to standard SQL. Termin 'eksploracja danych' jest często używany jako synonim terminu 'odkrywanie wiedzy' w bazach i magazynach danych. W istocie należy rozróżnić dwa pojęcia: odkrywanie wiedzy i eksploracja danych. Zgodnie z definicja, termin 'odkrywanie wiedzy' ma charakter ogólniejszy i odnosi sie do całego procesu odkrywania wiedzy, który stanowi zbiór kroków transformujących zbiór danych 'surowych' w zbiór wzorców, które następnie mogą być wykorzystane w procesie wspomagania podejmowania decyzji.

Można postawić trywialne pytanie 'Co można eksplorować?'. Odpowiedz jest równie trywialna jak pytanie, eksplorować możemy dowolny zbiór danych w postaci relacyjnych baz danych, hurtowni danych, repozytorium danych czy innych zaawansowanych systemów informatycznych w postaci obiektowych czy obiektowo-relacyjnych baz danych, przestrzennych baz danych, przebiegów czasowych i temporalnych baz danych, WWW, i innych. Najważniejszy jest odpowiedni dobór metody eksploracji do analizowanego zbioru informacji.

2 Techniki eksploracji danych

Techniki eksploracji danych można ogólnie podzielić na dwie zasadnicze kategorie. Techniki predykcyjne starają się, na podstawie odkrytych wzorców, dokonać uogólnienia i przewidywania (np. wartości nieznanego atrybutu, zachowania i cech nowego obiektu, itp.). Przykładami zastosowania technik predykcyjnych mogą być: identyfikacja docelowych grup klientów, ocena ryzyka ubezpieczeniowego związanego z klientem, lub oszacowanie prawdopodobieństwa przejścia klienta do konkurencyjnego usługodawcy. Techniki deskrypcyjne mają na celu wykorzystanie wzorców odkrytych w danych do spójnego opisu danych i uchwycenia ogólnych cech danych. Typowe przykłady technik deskrypcyjnych obejmują odkrywanie grup podobnych klientów, znajdowanie zbiorów produktów często kupowanych razem, lub identyfikacja osobliwości występujących w danych. Inny podział technik eksploracji danych jest związany z charakterem danych wejściowych. W przypadku technik uczenia nadzorowanego (ang. supervised learning) dane wejściowe zawierają tzw. zbiór uczący, w którym przykładowe instancje danych są powiązane z prawidłowym rozwiązaniem. Na podstawie zbioru uczącego dana technika potrafi ńauczyć sięódróżniać przykłady należące do róż-

nych klas, a zdobyta w ten sposób wiedza może być wykorzystana do formułowania uogólnień dotyczących przyszłych instancji problemu. Najczęściej spotykanymi technikami uczenia nadzorowanego są techniki klasyfikacji (drzewa decyzyjne , algorytmy bazujące na n najbliższych sąsiadach , sieci neuronowe , statystyka bayesowska) oraz techniki regresji. Drugą klasą technik eksploracji danych są techniki uczenia bez nadzoru (ang. unsupervised learning), gdy algorytm nie ma do dyspozycji zbioru uczącego. W takim przypadku algorytm eksploracji danych stara się sformułować model najlepiej pasujący do obserwowanych danych. Przykłady technik uczenia bez nadzoru obejmują techniki analizy skupień (ang. clustering), samoorganizujące się mapy oraz algorytmy maksymalizacji wartości oczekiwanej (ang. expectation-maximization).
Terminy ęksploracja danych"i ódkrywanie wiedzy w bazach danychśą często stosowane wymiennie, choć drugi termin posiada dużo szersze znaczenie. Odkrywanie wiedzy to cały proces akwizycji wiedzy, począwszy od selekcji danych źródłowych a skończywszy na ocenie odkrytych wzorców. Zgodnie z tą definicją, eksploracja danych oznacza zastosowanie konkretnego algorytmu odkrywania wzorców na wybranych danych źródłowych i stanowi jeden z etapów składowych całego procesu odkrywania wiedzy. Na cały proces składają się : sformułowanie problemu, wybór danych, czyszczenie danych, integracja danych, transformacja danych, eksploracja danych, wizualizacja i ocena odkrytych wzorców, i wreszcie zastosowanie wzorców. Postać uzyskanych wzorców zależy od zastosowanej techniki eksploracji danych. Poniżej przedstawiono opisy najpopularniejszych technik eksploracji. Z konieczności nie jest to lista wyczerpująca, uwzględniono tylko te metody eksploracji danych, które zostały zaimplementowane w pakiecie Oracle Data Mining [3].

2.1 Reguły asocjacyjne

Odkrywanie reguł asocjacyjnych polega na znalezieniu w dużej kolekcji zbiorów korelacji wiążącej współwystępowanie podzbiorów elementów. Znalezione korelacje są prezentowane jako reguły postaci X to Y (wsparcie, ufność), gdzie X i Y są rozłącznymi zbiorami elementów, wsparcie oznacza częstotliwość występowania zbioru X to Y w kolekcji zbiorów, zaś ufność reprezentuje prawdopodobieństwo warunkowe P(Y—X). Na gruncie analizy ekonomicznej reguły asocjacyjne są najczęściej stosowane do analizy koszyka zakupów. W takim przypadku wejściowa kolekcja zbiorów odpowiada bazie danych koszyków zakupów klientów, a odkryte reguły asocjacyjne reprezentują zbiory produktów, które są często nabywane wspólnie. Przykładowo, reguła asocjacyjna odkryta w bazie danych transakcji sklepowych mogłaby mieć postać (chleb,kiełbasa)to(musztarda) (3%,75%) a jej interpretacja byłaby następująca: 3% klientów sklepu kupiło chleb, kiełbasę i musztardę w trakcie pojedynczej transakcji, przy czym 75% transakcji zawierających chleb i kiełbasę, zawierało również musztardę. Odkryte reguły asocjacyjne mogą być wykorzystane do organizowania promocji i sprzedaży wiązanej, do konstruowania katalogów wysyłkowych, ustalania rozmieszczenia towarów na półkach, itp. Reguły asocjacyjne doczekały się wielu rozwinięć i

modyfikacji. Najbardziej znane przykłady takich algorytmów to Apriori oraz Eclat [3].

2.2 Wzorce sekwencji

Sekwencja jest to uporządkowany ciąg zbiorów elementów, gdzie każdy zbiór posiada dodatkowo znacznik czasowy. Sekwencja może reprezentować zbiory produktów kupowanych przez klientów podczas kolejnych wizyt w sklepie, filmy wypożyczane podczas kolejnych wizyt w wypożyczalni wideo, czy rozmowy telefoniczne wykonywane w określonych przedziałach czasu. Problem znajdowania wzorców sekwencji polega na znalezieniu, w bazie danych sekwencji, podsekwencji występujących częściej niż zadany przez użytkownika próg częstości, zwany progiem minimalnego wsparcia (ang. minsup). Przykładem wzorca sekwencji znalezionego w bazie danych księgarni może być następujący wzorzec: ('Ogniem i mieczem')to('Potop')to('Pan Wołodyjowski') (1,5%). Dodatkowo, użytkownik może sformułować ograniczenia dotyczące maksymalnych interwałów czasowych między kolejnymi wystąpieniami elementów sekwencji. Podobnie jak w przypadku reguł asocjacyjnych, także wzorce sekwencji doczekały się rozwinięć (np. uogólnione wzorce sekwencji) oraz efektywnych algorytmów eksploracji, takich jak GSP. Domeny potencjalnego zastosowania wzorców sekwencji praktycznie pokrywają się z regułami asocjacyjnymi i obejmują, między innymi: telekomunikację, handel detaliczny, zastosowania bankowe, ubezpieczenia, analizę dzienników serwerów WWW, i wiele innych [3].

2.3 Klasyfikacja

Klasyfikacja (ang. classification) jest jedną z najpopularniejszych technik eksploracji danych. Polega na stworzeniu modelu, który umożliwia przypisanie nowego, wcześniej niewidzianego obiektu, do jednej ze zbioru predefiniowanych klas. Model umożliwiający takie przypisanie nazywa się klasyfikatorem. Klasyfikator dokonuje przypisania na podstawie doświadczenia nabytego podczas trenowania i testowania na zbiorze uczącym. W trakcie wieloletnich prac prowadzonych nad klasyfikatorami i ich zastosowaniem w statystyce, uczeniu maszynowym, czy sztucznej inteligencji, zaproponowano bardzo wiele metod klasyfikacji. Najczęściej stosowane techniki to klasyfikacja bayesowska, klasyfikacja na podstawie k najbliższych sąsiadów , drzewa decyzyjne, sieci neuronowe, sieci bayesowskie, czy algorytmy SVM (ang. support vector machines). Popularność technik klasyfikacji wynika przede wszystkim z faktu szerokiej stosowalności tego modelu wiedzy. Klasyfikatory mogą być wykorzystane do oceny ryzyka związanego z udzieleniem klientowi kredytu, wyznaczeniem prawdopodobieństwa przejścia klienta do konkurencji, czy znalezienia zbioru klientów, którzy z największym prawdopodobieństwem odpowiedzą na ofertę promocyjną. Podstawową wadą praktycznie wszystkich technik klasyfikacji jest konieczność starannego wytrenowania klasyfikatora i trafnego wyboru rodzaju klasyfikatora w zależności od charakterystyki przetwarzanych danych. Te czynności mogą wymagać od użytkownika wiedzy technicznej, zazwyczaj wykraczającej poza sferę kompetencji analityków

i decydentów. Techniką podobną do klasyfikacji jest regresja (ang. regression). Różnica między dwiema technikami polega na tym, że w przypadku klasyfikacji przewidywana wartość jest kategoryczna, podczas gdy w regresji celem modelu jest przewidzenie wartości numerycznej [3].

2.4 Odkrywanie cech

Wiele przetwarzanych zbiorów danych charakteryzuje się bardzo dużą liczbą wymiarów (atrybutów). Niczyjego zdziwienia nie budzą tabele z danymi wejściowymi zawierające setki atrybutów kategorycznych i numerycznych. Niestety, efektywność większości metod eksploracji danych gwałtownie spada wraz z rosnącą liczbą przetwarzanych wymiarów. Jednym z rozwiązań tego problemu jest wybór cech (ang. feature selection) lub odkrywanie cech (ang. feature extraction). Pierwsza metoda polega na wyselekcjonowaniu z dużej liczby atrybutów tylko tych atrybutów, które posiadają istotną wartość informacyjną. Druga metoda polega na połączeniu aktualnie dostępnych atrybutów i stworzeniu ich liniowych kombinacji w celu zmniejszenia liczby wymiarów i uzyskania nowych źródeł danych. Wybór i generacja nowych atrybutów może odbywać się w sposób nadzorowany (wówczas wybierane są atrybuty, które umożliwiają dyskryminację między wartościami atrybutu decyzyjnego), lub też bez nadzoru (wówczas najczęściej wybiera się atrybuty powodujące najmniejszą utratę informacji) [3].

3 Grupowanie

Grupowaniu poswięciłem oddzielny rozdział, gdyż postanowiłem bliżej przedstawić ten temat, a takaże w późniejszej częsci artykułu pokazać wyniki swoich prac związanych z grupowaniem.
Grupowanie (klasteryzacja) jest to jedna z dziedzin eksploracji danych. Nie można mówić o grupowaniu nie wyjaśniając pojęcia klastera. Klaster to kolekcja obiektów podobnych do siebie w ramach jednego klastra i jednocześnie nie podobnych do obiektów należących do innych klastrów. Jest to uczenie bez nadzoru bez zdefiniowanych żadnych klas. Typowe zastosowania to narzędzia do analizy rozmieszczenia obiektów oraz przetwarzanie wstępne w różnych algorytmach. Jak również rozpoznawanie obrazów, analiza danych przestrzennych, analiza rynku, przetwarzanie obrazów, wyszukiwanie informacji, diagnostyka medyczna czy klasyfikacja dokumentów. W praktyce stosuje się klasteryzację do rozpoznawania grup klientów ich preferencji ,w ubezpieczeniach przy tworzeniu nowych form ubezpieczeń, czy przy planowaniu miast do rozpoznawania grup domów. Dobra klasteryzacja powinna cechować się wysokim podobieństwem wewnątrz klastrowym i niskim zewnątrz klastrowy. Jakoś klasteryzacji w stosowanych metodach zależy od miary podobieństwa. Atrybuty miar mogą być numeryczne, binarna bądź symboliczna. Przy atrybutach numerycznych wykorzystuje się funkcje odległości. Najbardziej popularna to odległość Minkowsikiego [2]

$$d(i,j) = \sqrt[q]{\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \ldots + \left|x_{ip} - x_{jp}\right|^q}$$

$$d(i,j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \ldots + \left|x_{ip} - x_{jp}\right|$$

Gdzie q to dodatnia liczba naturalna. Dla q = 1 otrzymujemy odległość Manhattan.

Dla q = 2 uzyskujemy odległość Euklidesową.

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \ldots + \left|x_{ip} - x_{jp}\right|^2}$$

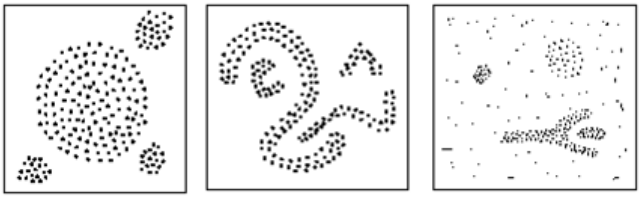
Atrybuty binarne przyjmują dwie wartości 0, 1 (true, false), przy porównywaniu obiektów, dla każdych cech zlicza się pokrycie w cechach. Natomiast atrybuty symboliczne są uogólnieniem atrybutów binarnych i mogą przyjmować kilka wartości np. czarny, biały, czerwony itd.

Główne metody klasteryzujące możemy podzielić na cztery kategorie: metody hierarchiczne, partycjonujące, metody wyszukiwania gęstości oraz metody gridowe. Metody hierarchiczne są szybkie w wykonaniu i produkują klastry o strukturze hierarchicznej, dzięki czemu możliwa jest obserwacja na różnych poziomach szczegółowości. Przykłady metod hierarchicznych to: SAHN, BIRCH i CURE. Kolejna grupa metod czyli metody partycjonujące. Konstruują one partycje dla n obiektów w postaci z góry określonej liczby klastrów. Wykorzystuje się w nich niej metody heurystyczne , gdyż sprawdzenie wszystkich podzbiorów nie jest możliwe. Przykładem jest metoda k-środków (k-means). Metody wyszukiwania gęstości bazują natomiast na wyszukiwaniu punktów gęsto ułożonych. Tworzą klastry o dowolnych kształtach i dobrze sobie radzą z szumem i punktami oddalonymi. Używa ich się często do klasteryzacji danych przestrzennych, gdyż wymagana jest przestrzeń metryczna. Przykładowe algorytmy to DBSCAN, którą omówię bardziej szczegółowo w dalszej części artykułu oraz OPTICS. Ostatnia grupą są metody gridowe, używające siatkowych struktur danych o wielu poziomach dokładności. Przykładami są algorytmy STING i WaveCluster [2].

3.1 Metody wyszukiwania gęstości - DBSCAN

Jak już powiedziałem wcześniej metody wyszukiwania gęstości bazują na wyszukiwaniu punktów gęsto ułożonych i tworzą klastry o dowolnych kształtach.

Ze zbiorami pokazanymi powyżej metody oparte na gęstości radzą sobie doskonale. Każdy klaster zawiera punkty o znacznie większym zagęszczeniu niż poza klastrem oraz lażdy klaster składa się z blisko położonych grup punktów gęsto ułożonych. Metody wyszukiwania gęstości przyjmują na wejściu dwa



Rysunek 1. Przykładowe zbiory punktów [1]

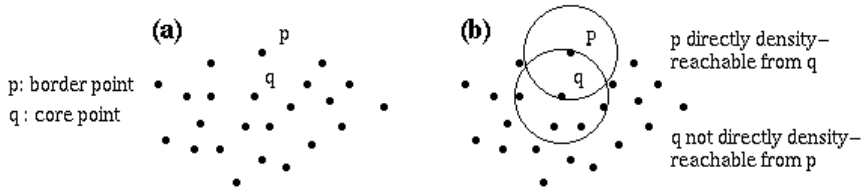
parametry: E - maksymalny promień sąsiedztwa oraz minPts - minimalna ilość punktów w E-sąsiedztwie danego punktu. Sąsiedztwo określane jest jako

$$N_{\varepsilon} = \left| y \in X; d(x,y) \leq \varepsilon \right|$$

Poniżej przedstawię podstawowe pojęcia związane z metodami gęstościowymi.

Obiekt rdzenny - punkt w którego sąsiedztwie znajduje się conjamniej minPts punktów.

Punkt y jest bezpośrednio osiągalny z rdzennego obiektu x jeżeli należy do jego sąsiedztwa N.

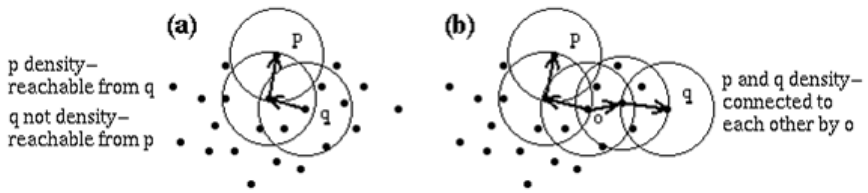


Rysunek 2. Bezpośrednie sąsiedztwo [1]

Punkt y jest osiągalny z rdzennego obiektu x jeśli istnieje ścieżka p1=x, p2, ..., pn=y taka, że pi+1 należy do N ?(pi)

Punkty x i y są gęstościowo połączone jeśli istnieje rdzenny punkt, taki że zarówno x jak i y są z niego osiągalne

DBSCAN (Density Based Spatial Clustering of Applications with Noise) Klaster zdefiniowany jest jako maksymalny zbiór gęstościowo połączonych punktów. Punkty niepołączone z żadnym klasterem to tzw. punkty oddalone (outliers). Punkty wewnątrz klastra, które nie są rdzenne stanowią granicę klastra. Algorytm wymaga zdefiniowania dwóch parametrów (minPts, E)

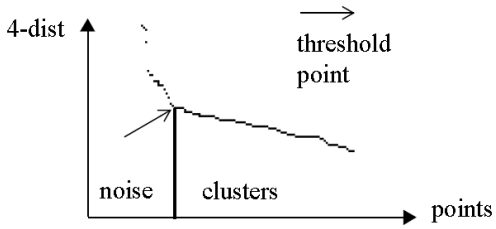


Rysunek 3. Osiągalność i gęstościowe połączenie [1]

Uproszczony opis algorytmu:

- Wybierz dowolny punkt p
- Wyszukaj wszystkie punkty osiągalne z p dla ustalonych MinPts, Jeśli p jest rdzennym punktem to tworzymy klaster; Jeśli p jest punktem granicznym, to żaden punkt nie jest osiągalny z p; Jeśli p jest punktem oddalonym to zaznaczamy go jako takiego;
- Powtarzaj proces do momentu, aż wszystkie punkty zostaną przetworzone

Do wyznaczani E można zastosować heurystykę. Mianowicie, dla każdego punktu wyznacza się odległość do k-najbliższego sąsiada (z reguły k = 4), następnie sortuje się te odległości w kierunku malejącym. Obrazuje to poniższy rysunek:



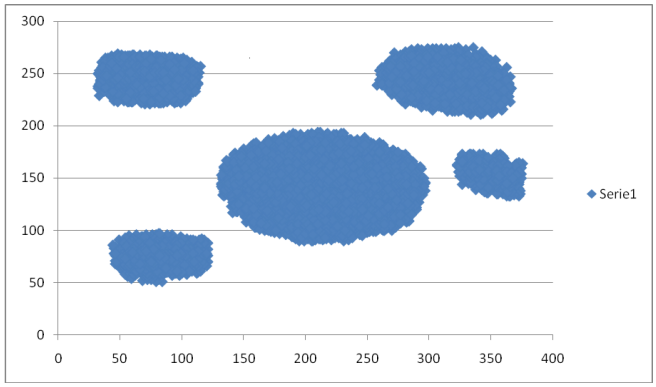
Rysunek 4. Wykres odległości do k-sasiedztwa [1]

Dla tak wyznaczonych odległości należy ustalić próg. Wszystkie punkty o odległości mniejszej lub równej od progu będą punktami rdzennymi.

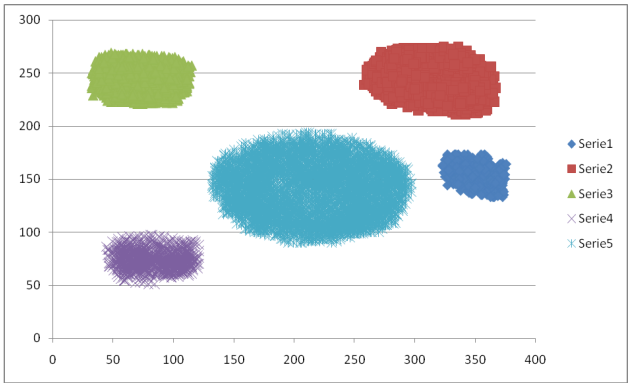
3.2 Praktyczny przykład grupowania

Do analizy działania algorytmów grupowania użyłem prostego dwuwymiarowego przykładu dla którego łatwo jest zobrazować wyniki grupowania (Rysunek 5).

Dla algorytmu DBSCAN napisałem własną implementację w języku Java, wynik działania algorytmu na przykładowym zbiorze danych jest przedstawiony na rysunku (Rysunek 6)



Rysunek 5. Badany zbiór danych dwuwymiarowych



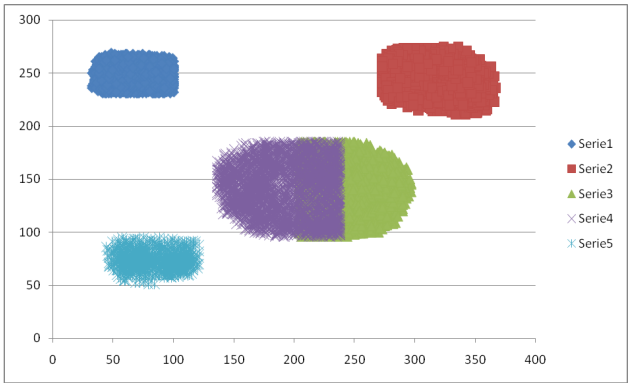
Rysunek 6. Badany zbiór danych dwuwymiarowych

Jak widać algorytm DBSCAN doskonale poradził sobie z badanym przykładowym zbiorem danych. Parametry wejściowe algorytmu DBSCAN dla badanego przypadku ustawiłem następująco: E = 13, minPts = 4. Badania wykonałem także z wykorzystaniem narzędzi i algorytmów dostarczanych przez Oracle i SQL Serwer 2005. W bazie Oracle wykorzystałem do grupowania algorytmu k-means. Wynik działania pokazuje (Rysunek 7).

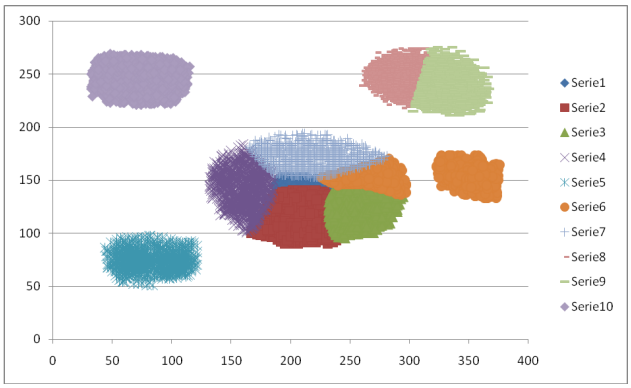
W algorytmie k-means należy z góry określić liczbę klastrów na jaka ma zostać podzielony zbiór wejściowy. Ponieważ znałem testowy przypadek ustawiłem ten parametr na wartość 5. Algorytm podzielił zbiór na 5 klastrów, ale jeden z nich został pominięty, a jeden który wydaje się być pojedynczym klastrem został podzielony na dwa.

SQL Serwer dokonał podziału w następujący sposób (Rysunek 8).

Gołym okiem widać, że badany przypadek testowy nie zgrupował się dobrze w SQL Serwerze 2005.



Rysunek 7. Grupowanie algorytmem k-means w Oracle



Rysunek 8. Grupowanie z wykorzystaniem SQL Server 2005

3.3 Podsumowanie

Wraz z rozwojem dziedziny metod eksploracji danych rozwijają się narzędzia do eksploracji danych. Główni producenci systemów baz danych takie jak Oracle, czy Microsoft dostarczają takich narzędzi i w pełni wspierają eksploracje danych. Można w nich znaleźć wszystkie główne metody i algorytmy. Dostarczane narzędzia są proste i wygodne w obsłudze oraz charakteryzują się wysoką wydajnością. Wykorzystując te gotowe narzędzia można przeprowadzić prawie cały proces odkrywania wiedzy, począwszy od selekcji i wstępnego przetwarzania danych źródłowych, aż po wygenerowanie wzorców. Ścisła integracja technik eksploracji danych z bazą danych umożliwia wykorzystanie technik eksploracji w aplikacjach, ułatwia pielęgnację aplikacji, oferuje ogromnie wzbogaconą funkcjonalność aplikacji. Ponadto użytkownicy mają dostęp do wyczerpujących dokumentacji i tutoriali.

Literatura

1. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Institute for Computer Science, University of Munich.
2. Adam Lessnau: Klasteryzacja, (2005).
3. Mikołaj Morzy: Oracle Data Mining – odkrywanie wiedzy w dużych wolumenach danych, Instytut Informatyki Politechniki Poznańskiej.