Systemy do dogłębnej analizy danych - porównanie.

Tomasz Winiarski

Rok akademicki 2000/2001

Praca licencjacka
pod kierunkiem prof. dr. hab. W. Ducha

Spis treści

Rozdział 1

Wstęp

W krótkim okresie czasu, podczas którego świat uległ komputeryzacji, ludzie zaczęli używać różnego rodzaju maszyn liczących w celu przechowywania danych. Podyktowane to zostało wygodą - szybkim dostępem do informacji, możliwością ich graficznego przedstawienia. Jednak dość wcześnie okazało się, że komputerom brakuje możliwości wykrywania podobieństw i zależności. Zaczęły powstawać systemy uczące się - modele sieci neuronowych, drzewa decyzji. Drążenie danych (data mining) jest dziedziną, która wykorzystuje zdolność maszyn do uczenia się. Systemy do dogłębnej analizy danych (zwane także systemami data mining) można rozpatrywać jako naturalną ewolucję w technologii obsługi baz danych, która prowadziła przez zbieranie danych i tworzenie baz danych. Następnie powstały systemy potrzebne do zarządzania i analizy danych. Data mining wydaje się być następnym krokiem w procesie przetwarzania informacji, jest próbą odkrywania *wiedzy* ukrytej w danych. Dziedzina ta wywodzi się głównie z uczenia maszynowego oraz statystyki. Duży postęp w dziedzinie obróbki informacji, powstanie relacyjnych baz danych oraz języków zapytań (Query Languages takich jak np.: SQL) miały niebagatelny wpływ na jej rozwój. Ogromne ilości danych zgromadzone przez różne instytucje, były poddawane obróbce w celu wydobycia użytecznej wiedzy. Firmy piszące oprogramowanie do tworzenia i obsługi wymyślały różnego rodzaju “standardy”. Powstanie hurtowni danych (data warehouses) rozwiało poniekąd te problemy - hurtownie mają inną architekturę niż zwykłe składnice danych, pozwalają na znaczne zmniejszenie ilości zbieranych informacji, ograniczając się do pewnego tylko typu danych. Te właśnie cechy spowodowały, iż są one teraz jednym z ważniejszych źródeł poddawanym obróbce w celu pozyskania wiedzy. Podstawową motywacją eksploracji danych jest nadal wzrost ilości zbieranych informacji w bazach danych oraz potrzeba przekształcenia ich w wiedzę. Wiele z systemów wspomagania decyzji bazuje na technikach używanych przez systemy drążenia danych.

Tematem niniejszej pracy są systemy wydobywania wiedzy z danych. Moim celem jest porównanie programów do data mining wykonanych przez różnych producentów. Dane porównawcze zawarłem w postaci tabelarycznej w aneksie. Drugi rozdział poświęcony jest przybliżeniu czytelnikowi terminu *data mining*. Zaznaczyć należy, że terminologia polska nie jest ustalona, stąd pojawiają się kalki językowe: drążenie danych, dogłębna analiza danych, Także w tym rozdziale zostanie przedstawiona typowa architektura systemów oraz różne możliwo-

ści zastosowania. Rozdział trzeci opisuje typy składnic danych, na jakich pracują najczęściej programy do data mining. W następnym rozdziale przestawiam rodzaje używanych analiz oraz metody, jakie stosuje się, by systemy były w stanie wydobyć wiedzę. Rozdział piąty to pokazanie, w jaki sposób można klasyfikować programy służące do ekploracji danych. W rozdziale szóstym wskazuję ważniejsze problemy, na które natykają się projektanci oprogramowania. Rozdział siódmy zawiera opis przykładowych programów komercyjnych. W zakończeniu podsumowuję problematykę przedstawioną w pracy.

Rozdział 2

Drążenie danych

2.1 Pochodzenie terminu

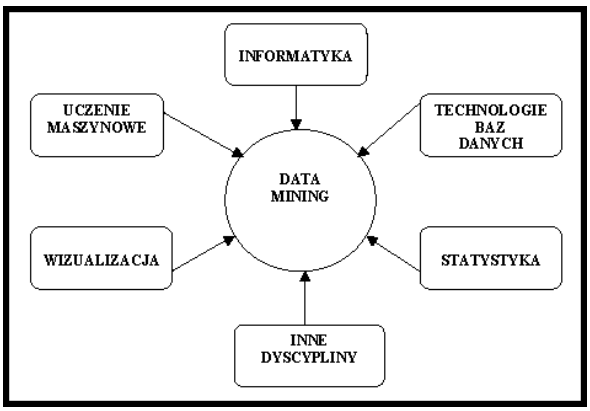
Termin data mining odzwierciedla dogłębną analizę wraz z otrzymywaniem wiedzy z ogromnych zbiorów danych - oznacza **odkrywanie ukrytych zależności**. Przypomina to wydobywanie cennego surowca (jak złota ze skał czy piasku) z "gąszcza" innych niepotrzebnych rzeczy. Dlatego proces ten ma wiele nazw takich jak np. "wykopywanie wiedzy". Najbardziej rozpowszechnioną nazwą tego procesu jest data mining, ale spotkać także można inne:

- "wydobywanie wiedzy z baz danych" (knowledge extraction from data bases),
- "ekstrakcja wiedzy" (knowledge extraction),
- "analiza wzorców" (data/pattern analysis),
- "archeologia danych" (data archeology),
- "data dredging".

Drążenie danych jest dziedziną interdyscyplinarną, ponieważ zawiera w sobie wiele technik pochodzących z różnych dyscyplin, takich jak: bazy danych, statystyka, uczenie maszynowe, wysoko wydajne obliczenia komputerowe, obróbkę obrazów i sygnałów, oraz analizę danych przestrzennych. Poza tym używa metody rozpoznawania wzorców, sieci neuronowych, wizualizacyjnych metod eksploracji danych, logiki rozmytej i przybliżonej, by otrzymywać informacje i wizualizować wiedzę.

2.2 Procesy składające się na analizę danych

Data mining jest tylko jednym z kroków w procesie odkrywania wiedzy. Całość można przedstawić w następującej kolejności:



Rysunek 2.1: Data mining jako dziedzina interdyscyplinarna.

1. Czyszczenie danych. Zadaniem tego kroku jest wyzbycie się szumów z danych tzn. pozbycie się braków w danych i poprawienie danych niekonsekwentnych. Krok ten jest ważny, ponieważ jeśli będziemy mieli w bazie nieodpowiednie dane, to nie uzyskamy zadawalających wyników.

2. Konsolidacja danych. Proces ten jest potrzebny, by połączyć dane pochodzące z różnych źródeł. W celu znalezienia jakiś zależności, system potrzebuje znormalizowanych danych w jednolitej formie.

   Popularny trend w przemyśle informacji polega na czyszczeniu i konsolidacji danych, a następnie umieszczeniu rezultatów w hurtowniach.

3. Selekcja danych. Każdy model wymaga specyficznych danych, w oparciu o które jest on budowany. Przydatne tu mogą być narzędzia służące do wizualizacji danych i zależności między nimi - mogą one wskazać dane niezależne. Możliwa jest selekcja danych, np. przez odrzucenie niepasujących elementów. Czasami takie dane mogą zawierać ważne informacje, dlatego ich wybieranie jest bardzo delikatnym procesem mającym wpływ na wynik analizy. Często przy obróbce dużych baz stosuje się próbkowanie danych, by zapobiec utracie informacji. W związku z tym próbkowane dane potrzebne do analizy powinny być wybierane w sposób zupełnie przypadkowy.

4. Transformacja danych. Wybrane dane mogą zostać poddane transformacji ze względu na aktualne potrzeby. Przykładem może być bankowość, gdzie potrzebna jest ocena ryzyka udzielenia kredytu, dlatego preferowane będzie używanie stosunku zadłużenia do przychodów zamiast użycia tych dwóch zmiennych jako danych niezależnych.

5. Wydobywanie wiedzy. Jest to proces, który za pomocą inteligentnych metod pozwala na znajdowanie ukrytych zależności między danymi. Odkrywa on powiązania między pozornie niezależnymi zmiennymi. Buduje modele bazując na dwóch rodzajach uczenia się: z nadzorem (takich jak: klasyfikacja, regresja) oraz uczenia bez nadzoru (takich jak: klastering, asocjacja, detekcja powtórzeń).

6. Ocena wzorców. System data mining ma potencjalną możliwość generowania dużej ilości wzorców i reguł. Potrzebny jest zatem proces, który zredukuje liczbę stworzonych modeli i oceni, które ze wskazanych wzorców są interesujące. Reguły produkowane przez system muszą być łatwe do zrozumienia i interpretacji przez użytkownika.

7. Prezentacja wiedzy. Krok, w którym używane są metody wizualizacji, w celu przedstawiania odkrytej wiedzy użytkownikowi w sposób jak najbardziej czytelny (np. wykresy 2D i 3D). Od tego procesu zależy późniejsza interpretacja, gdyż zbyt zagmatwany sposób pokazywania zależności może skuteczne zatrzeć wyrazistość otrzymanych wyników.

2.3 Typowa architektura

Typowa architektura systemów data mining może zawierać następujące elementy:

- Hurtownie, bazy i inne rodzaje składnic danych - bazy lub zbiory baz danych, hurtownie, arkusze kalkulacyjne lub inne rodzaje składnic informacji są źródłem danych podlegających analizie. Czyszczenie i konsolidacja danych mogą często mieć miejsce już podczas zbierania informacji.

- Serwer hurtowni lub baz danych - jest on odpowiedzialny za przechwytywanie i przenoszenie istotnych danych; opiera się na żądaniach (zapytaniach) użytkownika systemu data mining.

- Baza wiedzy - to jest domena wiedzy, która jest używana do przeprowadzania poszukiwań czy wypróbowania i oceny istoty odkrytych wzorców. Zawarta w niej wiedza zawiera koncepcję hierarchii, użytej by zorganizować atrybuty zmiennych na różnych poziomach abstrakcji. Tu może być zawarta wiedza użytkownika, by zapobiec znajdowaniu zależności już znanych.

- Moduł data mining - esencja procesu odkrywania wiedzy, która składa się ze zbioru modułów dla następujących zadań: charakteryzacja, asocjacja, klasyfikacja, analiza klastrowa oraz analiza ewolucyjna i dewiacyjna.

- Moduł oceny wzorców - to komponent, który sprawdza istotę modelu i ingeruje w proces analizy tak, by ten skupiał się tylko na interesujących wzorcach. Moduł ten może być zintegrowany z modułem drążącym. W celu zwiększenia efektywności analizy nie należy drążyć tak głęboko, jak tylko to możliwe, ale jedynie w obrębie interesujących wzorców.

- Graficzny interfejs użytkownika - to moduł służący do komunikowania się z użytkownikiem. Pozwala na ingerencję użytkownika w system przez zadawanie pytań lub zadań systemowi. Dostarcza informacji potrzebnych do skupienia się na szukaniu oraz umożliwia eksplorację. Przedstawia i wizualizuje użytkownikowi wydobytą wiedzę.

2.4 Zastosowania systemów

Systemy wykonujące tak skomplikowaną analizę znalazły wiele zastosowań. Najbardziej rozpowszechnione są one w dziedzinach:

- bankowość
- ubezpieczenia
- zarządzanie biznesem,
- kontrola produkcji,
- analiza sprzedaży,
- projektowanie,
- telekomunikacja,
- farmacja,
- medycyna,
- badania naukowe (fizyka, chemia, astronomia),
- technika.

Systemy do dogłębnej analizy danych są stosowane przez duże przedsiębiorstwa, takie jak *Mobil Oil* (przechowywanie ponad 100 terabajtów danych związanych z wydobyciem ropy naftowej), czy instytucje takie jak *NASA* (generuje w ciągu każdej godziny dziesiątki gigabajtów danych obrazowych).

Pośród “zwyczajnych” zastosowań wymienić można detekcję anomalii - próby wyłudzenia ubezpieczenia i oszustwa podatkowe. Do innych, mniej rozpowszechnionych, dołączyć należy próby zastosowania sieci neuronowych do sterowania lotu satelitą czy helikopterem, albo znalezienie przez system analizujący nowego rodzajów galaktyk, który był wcześniej przeoczony przez naukowców.

2.4.1 Oszacowanie zbierania paliwa jądrowego i kategoryzacja obrazu

Projekt ten dotyczy dodania kamer cyfrowych i obróbki otrzymanych z nich obrazów do celów inspekcyjnych dla Międzynarodowej Komisji do Energii Atomowej. Jego zadaniem jest wynalezienie systemu do automatycznego oszacowania zbiorów paliwa jądrowego przez analizę obrazów otrzymanych z cyfrowego detektora promieniowania Czerenkowa.

Celem jest tu dostarczenie nie tylko samego obrazu z kamery CCD czy innego urządzenia do przetwarzania obrazu, ale przeprowadzenie analizy i wyciągnięcie przydatnych informacji dotyczących zbioru paliwa w czasie rzeczywistym. Na podstawie tego projektu ukazały się ciekawe podprojekty skupiające się na kategoryzacji obrazu.

Promieniowanie Czerenkowa wytwarzane w wodzie otaczającej paliwo jest silne w ultrafiolecie i dlatego inspektorowie używają wzmacniaczy czułych w zakresie UV. Tego typu urządzenia używane są, by sprawdzić dystrybucję (rozkład) emitowanego promieniowania od zbiornika. Rozkład taki powinien być inny dla zbiornika z prawdziwym paliwem od rozkładu promieniowania wadliwych prętów nie nadających się na paliwo. Efekt ten jednakże jest bardzo subtelny i trudny do wykrycia. Wymyślono więc kamerę CCD służącą do detekcji promieniowania Czerenkowa (Cerenkov Viewing Device - CVD). Urządzenie to nie tylko dostarcza obraz lecz pozwala na jego obróbkę w czasie rzeczywistym (lub prawie), co jest bardzo ważne wówczas, gdy inspektorzy mają 10-20 sekund na sprawdzenie każdego ze składów paliwa. Kontrolujący poznać może nawet właściwości sprawdzanego zbiornika.

2.4.2 Utrzymanie sprawności maszyn

Specjaliści w Hong Kongu używają rozmytych systemów sieci neuronowych,aby wykryć prawdopodobieństwo usterki maszyny wydającej bilety na stacjach kolejowych. Zapobiegają w ten sposób przestojom i opóźnieniom, które były spowodowane przez popsute automaty.

2.4.3 Porównanie własności materiału przed i po spaleniu (microgravity combustion experiments)

Ważne jest, aby dokonać charakterystyki zmienionej po spaleniu powierzchni. Eksperyment ten został wykonany, by lepiej zrozumieć, jak płomień rozprzestrzenia się w warunkach mikrograwitacji takiej, jaka istnieje w przestrzeni kosmicznej. Ten rodzaj informacji jest kluczowym dla bezpieczeństwa astronautów przebywających w stacjach kosmicznych. Aby otrzymać precyzyjną informację o stanie przed i po spaleniu powierzchni użyta została do tego celu naddźwiękowa technika obrazowa.

Profil powierzchni spalonej w warunkach mikrograwitacji okazuje się znacząco różny od tego, który możemy zobaczyć w warunkach normalnych.[1]

[1] Dr Don J. Roth, NASA Lewis Research Center

2.4.4 Pomiar kolorów a stan chorobowy roślin

Aby wykryć stan martwoty, choroby czy innych dolegliwości wykazywanych przez rośliny należy użyć informacji o kolorach ich liści. *SigmaScan Pro* potrafi określać ilościowo obrazy na podstawie zawartości kolorów. Najpierw tworzony jest obraz za pomocą kamery CCD, a następnie za pomoc histogramu sprawdzana jest intensywność poszczególnych barw. Po całkowitej obróbce możliwa jest diagnoza stanu chorobowego rośliny.

2.4.5 Ocena wielkości kropli oleju

SigmaScan Pro and *TableCurve 2D*0 zostały użytem, aby scharakteryzować wielkość i rozłożenie kropli oleju.(oil droplets). *SigmaScan* zmierzył promień kropli oleju, a *TableCurve* znalazł funkcję, która najlepiej charakteryzuje rozkład ich wielkości. Badane krople zawieszone były w cieczy i rejestrowane przez konwencjonalną kamerę CCD. Obróbka obrazu była wykonana na komputerze klasy PC.

Obraz był kalibrowany za pomocą dwóch punktów, a następnie, poprzez dalszą obróbkę, powiększony zostaje kontrast, aby operator mógł łatwiej zobaczyć krople oleju. Po skończonej obróbce program pozwala na zmierzenie promienia każdej kropli.

2.4.6 Medycyna

Najbardziej spektakularnym zastosowaniem w medycynie jest diagnoza chorych na raka. Metoda polega na obróbce i analizie obrazów w celu znalezienia anomalii, które świadczą o stanie chorobowym. W trakcie testów klinicznych, system oparty na sieciach neuronowych wykazał się średnio 97% stopniem trafności. (dotychczas stosowany system miał 70%).

2.4.7 Detekcja prób oszustw ubezpieczeniowych

Większość oszustw ubezpieczeniowych jest niezauważana. Firmy ubezpieczeniowy i zdrowotne wykrywają około $10\% - 20\%$ prób wyłudzeń różnego typu. Zastosowanie inteligentnych metod detekcji pozwala na wskazanie potencjalnych podejrzanych i znaczne zmniejszenie prawdopodobieństwa oszustwa.

2.4.8 Kalkulowanie terminu przydatności

Program *SigmaPlot*[2] posiada makro, “Shelf Life”, które może być użyte, by otrzymać dokładne przewidywanie terminu przydatności artykułów farmaceutycznych. Nie potrzebuje więc on ani wizualnej ani numerycznej interpolacji. Makro może być uruchamiane wielokrotnie na zbiorach danych dla różnych opakowań, temperatur czy wilgotności. Redukuje czas potrzebny

[2]program SigmaPlot jest programam głównmie do rysowania wykresów, zawiera jednak w sobie metody inteligencji obliczeniowej

do przygotowania wejścia nowych leków do sprzedaży. Algorytm sprawdzony został na zbiorach danych przekraczających spektrum przypadków, które mogą być obserwowane w praktyce. Program znacznie upraszcza wprowadzenie nowych leków do obiegu.

2.4.9 Optymalizacja

Systemy oparte na sieciach neuronowych wspomagają procesy optymalizacji. Stosuje się je z sukcesem do sterowania silnikami w samolotach takich jak Concord. Możliwe jest także ich użycie w telekomunikacji w celu zidentyfikowania błędnych modułów zawartych w oprogramowaniu.

Ogólnie sieci neuronowe potrafią minimalizować koszty produkcji oraz zmniejszać emisję substancji szkodliwych do środowiska przez wybieranie materiałów i sposobów ich obróbki.

2.4.10 Polityka

W trakcie kampanii wyborczej Billa Clintona, jego sztab odkrył, że aby został on wybrany na drugą kadencję, powinien zwrócić szczególną uwagę na niezdecydowaną część elektoratu, czyli na rodziny, które grają w kręgle. Doradcy prezydenta używali sieci neuronowych w celu odkrycia tej zależności.

Rozdział 3

Typy używanych baz danych

W ogólności data mining może być wykonany w każdym rodzaju składnicy danych, tzn. na relacyjnych bazach danych, hurtowniach danych, transakcyjnych bazach danych, zaawansowanych systemach baz danych, “płaskich” plikach (np. ASCII) i najlepiej sprzedającym się dziś World Wide Web (WWW).

3.1 Relacyjne bazy danych

Relacyjne bazy danych są nazywane także systemami zarządzania bazami danych (database management system - DBMS). System ten składa się ze zbioru powiązanych ze sobą tabel, które nazywane są bazami danych. Programy zawierają mechanizm definiujący strukturę bazy, w celu przechowywania, dzielenia (share), dystrybucji i jednoczesnego dostępu do danych. Zapewniają także konsystencję i bezpieczeństwo danych, nawet pomimo utraty stabilności i prób nieautoryzowanego dostępu. Relacyjna baza danych jest zbiorem tablic, którym przypisane są niepowtarzalne nazwy. Model ER (entity-relationship) opiera swoje relacje bazując na zbiorach jednostek i ich powiązań (np. infromacja o kliencie składa się ze zbioru atrybutów, cech etc. i jest on opisany przez tablicę relacji: klient, przedmiot - sprzedany artykuł, nazwisko, nr id., adres, wiek, stan cywilny, przychód, informacje o kredytach itp.) Podobieństwa każdej relacji składają się ze zbioru atrybutów opisujących ich własności. Bazy danych relacyjnych mogą być dostępne przez zapytania (SQL) lub przy pomocy interfejsu graficznego.

3.2 Zaawansowane systemy baz danych

Systemy te stworzone zostały po to, by przechowywać dane przestrzenne (np. mapy), dane niezbędne do projektowania inżynierskiego (jak budowa budynków, składników systemu, obwodów), World Wide Web (ogromne ilości danych w internecie), dane czasowe (np. dane historyczne), multimedialne bazy danych i wiele innych. Można wyróżnić tu następujące systemy baz danych:

- obiektowe,
- obiektowo-relacyjne,
- bazy danych przestrzennych,
- tekstowe,
- multimedialne,
- czasowe (time-series)

3.3 Hurtownie danych

Hurtownie danych są jednym z ważniejszych źródeł danych w systemach pozyskujących wiedzę, dlatego przedstawię pokrótce ich ważniejsze cechy:

3.3.1 Tematyczność

Systemy transakcyjne są opracowywane z myślą o obsłudze bieżących zadań, takich jak: sprzedaż, wystawianie faktur czy gospodarka magazynowa. Pod tym kątem są też projektowane tablice baz danych, które zapewniają efektywną obsługę transakcji, ale w niewielkim stopniu pozwalają na analizowanie gromadzonych w nich danych. W hurtowniach danych stosuje się więc inne podejście: dane są grupowane względem pewnych kategorii np. klientów, produktów, dostawców. Taka systematyzacja pozwala na formułowanie bardziej przekrojowych zapytań i ułatwia analizę np. stanu przedsiębiorstwa. W hurtowniach umieszczane są tylko te dane, które dotyczą określonego tematu, np. klientów czy sprzedaży. Nie jest to zatem centralna baza wszystkich danych. Dane zostają starannie wybrane z różnych źródeł ze względu na przydatonść w analizie.

3.3.2 Integralność

Faktem jest, że dane zbierane są z różnych źródeł (np. z różnych systemów transakcyjnych), stąd pojawia się problem ich integracji (każde źródło może mieć je zakodowane w inny sposób); jak wiadomo, w celu przeprowadzenia wspólnej analizy tych danych, trzeba je wpierw sprowadzić do ujednoliconego sposobu zapisu i kodowania. Integracja obejmuje także normalizację wartości pól, polegającą na przeliczeniu ich na jedną wybraną jednostkę miary i identyfikację tożsamych ciągów i znaków (np. dopasowywanie nazwisk itp.).

3.3.3 Oznaczenia czasowe

Dane gromadzone w hurtowniach danych są sygnowane podczas ładowania znacznikiem czasu. Takie rozwiązanie daje możliwość odwzorowania kolejnych stanów baz operacyjnych. Mogą one zostać uaktualnione, ale historia zmian będzie zapamiętana w hurtowni. Można ją potem wykorzystać do analizy trendów zachodzących w czasie. Oznaczenie czasowe pozwala na gromadzenie danych i archiwizowanie ich przez bardzo długi (zwykle kilkuletni) okres czasu, co z kolei umożliwia analizę długoterminową, potrzebną do podejmowania decyzji.

3.3.4 Niezmienność

Dane zapisane w hurtowni charakteryzuje niezmienność. W przypadku operacyjnych baz danych operacje te dotyczą wstawiania, usuwania lub zmiany już istniejących zapisów; związane jest to z funkcjonowaniem tego rodzaju systemów. w przypadku hurtowni danych dozwolone są tylko dwie operacje: ładowania i dostępu do danych. Raz załadowane dane stanowią obraz przedsiębiorstwa w danej chwili, zatem nigdy już nie mogą zostać zmienione. Jeśli część danych, którą tam umieszczono, ulegnie aktualizacji, to wszelkie zmiany zostaną uwzględnione przy następnym załadowaniu hurtowni i będą miały inny znacznik czasu. Ta procedura pozwala odtwarzać zmiany zachodzące w bazie danych z perspektywy czasu. Operacja usuwania zapisów w odniesieniu do hurtowni praktycznie nie istnieje (wyjątek stanowi np. przebudowa hurtowni).

Do branży najczęściej korzystających z hurtowni danych zaliczamy: bankowość, ubezpieczenia, telekomunikację, energetykę, handel i biznes internetowy. Z punktu widzenia hurtowni danych data mining może być postrzegane jako zaawansowane stadium on-line analitical processing (OLAP) czyli analizy danych przeprowadzanej na bieżąco. Jednakże data mining idzie dalej, poza proste sumowanie danych, jego celem jest **zrozumienie danych**.

Rozdział 4

Rodzaje i metody analizy danych

Rozdział ten poświęcony jest rodzajom analizy danych oraz stosowanym w tym celu metodom.

4.1 Rodzaje analizy danych

- Analiza asocjacyjna

Analiza asocjacyjna jest to odkrywanie skojarzonych danych i ukazywanie zależności atrybut - wartość, które często występują w zbiorze danych (np. klienci w wieku od 20 do 29 lat i przychodach 20tys. - 29tys. zł. kupują odtwarzacze CD. Istnieje więc 60% prawdopodobieństwo, że klient w tym wieku dokona takiego właśnie zakupu). Tak więc w wyniku analizy asocjacyjnej odkrywane są zależności miedzy kilkoma atrybutami.

- Klasyfikacja i predykcja

Klasyfikacja polega na znajdowaniu zbiorów modeli (lub funkcji), które opisują i rozróżniają koncepcje i klasy danych. Celem jest tu możliwość użycia pewnego modelu do przewidywań pewnych klas i obiektów, które dotychczas nie zostały nigdzie zaklasyfikowane. Model bazuje tu na analizie zbiorów “danych treningowych”, czyli na danych, które już zostały wcześniej sklasyfikowane. Dostarczony typ może być reprezentowany w następujących formach: klasyfikacja (reguły if-then), drzewa decyzyjne, formuły matematyczne lub sieci neuronowe. Klasyfikacja pozwala, poprzez bazowanie na dostępnych danych, na “odgadywanie” danych brakujących, predykcję i identyfikację trendów.

- Analiza klastrowa

W odróżnieniu od poprzedniego przypadku, klastering analizuje obiekty danych bez sprawdzenia nazwy klas, przez co może być stosowany do ich generowania. Ogólnie rzecz ujmując, etykiety klas nie są obecne w danych treningowych z tej prostej przyczyny, że nie są znane. Obiekty są dzielone na klastry lub grupowane ze względu na minimalne albo maksymalne podobieństwa obiektów przypisanych do jednego skupienia. Klastering umożliwa taksonomizację

- ustawia wyników obserwacji w klasy o określonej hierarchii, które stanowią grupy podobnych przypadków.

- Analiza wyrzutków

Baza danych może zawierać obiekty, które nie współgrają z resztą danych. Większość metod data mining odrzuca je, traktując jako szum. Jednakże w niektórych zastosowaniach, takich jak detekcja oszustw, te rzadkie przypadki są szczególnie interesujące, wręcz ważniejsze niż reszta danych. Owe niepasujące dane mogą być wykrywane za pomocą testów statystycznych, które zakładają pewien model rozkładu lub prawdopodobieństwa.

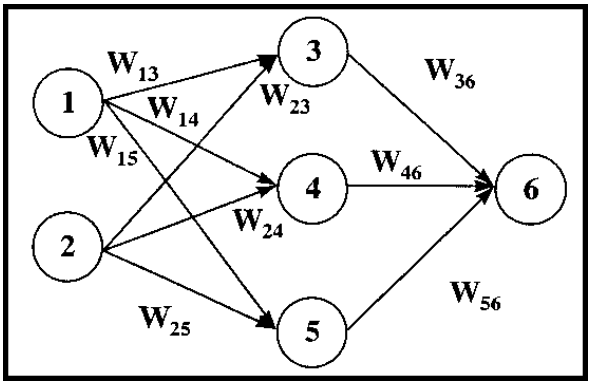
- Analiza ewolucyjna

Ten typ analizy opisuje i modeluje regulacje lub trendy dla obiektów, których zachowanie zmienia się w czasie, mimo że może zawierać charakteryzację, dyskryminację, asocjację czy klastering.

4.2 Metody

Opiszę teraz metody analizy używane w systemach dogłębnej analizy danych:

- sieci neuronowe
- drzewa decyzyjne
- metodę *k* najbliższych sąsiadów i rozumowanie bazujące na pamięci
- algorytmy genetyczne
- detekcję powtórzeń i asocjacji
- regresję logistyczną
- analizę dyskryminacyjną
- Ogólne modele addytywne (GAM)
- Multivariate Adaptive Regression Splines (MARS)



Rysunek 4.1: Sieć neuronowa z jedną warstwą ukrytą.

4.2.1 Sieci neuronowe

Wynalezienie sieci neuronowych odgrywa znaczącą rolę w badaniach prowadzonych nad sztuczną inteligencją. Inspiracją stworzenia sieci neuronowych była biologia, co tłumaczy, dlaczego sieci są przedstawiane jako proste neuropodobne procesory. Należy stwierdzić, że owe biologiczne sieci neuronowe posiadają nieporównywalnie bardziej złożoną budowę niż ich sztuczne odpowiedniki. Sieci neuronowe są jednak niezwykle interesujące, ponieważ oferują rozwiązania skomplikowanych problemów, mogą używać bardzo dużych ilości niezależnych wejść - zmiennych. Można używać ich do klasyfikacji (np. wyjście jest jasną zmienną) lub regresji (np. zmienne wyjściowe są ciągłe).

Sieć neuronowa (rysunek 1) zaczyna się od warstwy wejściowej, w której każdy węzeł jest związany z niezależną zmienną (nazywaną także wejściem lub predykatem). Te węzły są połączone z innymi węzłami w warstwie ukrytej. Każdy węzeł wejściowy jest zazwyczaj połączony z każdym z węzłów znajdujących się w warstwie ukrytej, te zaś mogą być połączone z innymi węzłami w innych warstwach ukrytych lub w warstwie wyjściowej. Na rysunku widać także, jak po przejściu przez warstwę wejściową na każdy węzeł ma wpływ zbiór wejść mnożąc je przez wagi połączeń W_{xy} (np. waga z węzła 1 do węzła 3 jest W_{13}) dodając je razem i używając na nich funkcji aktywacji i przekazuje wynik do węzłów w następnej warstwie. Dla przykładu rozważmy wartość, którą otrzymamy w węźle 3.Jest to Funkcja aktywacji zastosowana na W_{14} z węzła pierwszego (1) + W_{24} z węzła drugiego (2).:

$$\sigma(x) = \sigma\left(\sum_i^2 W_{i3} x_i\right)$$

gdzie $W_{x_1 x_2}$ to wagi miedzy odpowiednimi połączeniami.

Wagi oczywiście są nieznane i są wyliczane podczas uczenia sieci. Najczęściej stosowaną metodą do trenowania sieci była i jest sieć oparta na *wstecznej propagacji* błędu. Stosuje się także inne metody.Każda z metod ma szereg własnych parametrów, które kontrolują różne aspekty treningu - np. zapobieganie wpadaniu w minima lokalne. W sieciach neuronowych

ważną rzeczą jest znalezienie odpowiedniego modelu, który posiadałby odpowiednią liczbę wejść i wyjść, odpowiednią liczbę węzłów ukrytych, połączeń oraz funkcji aktywacji.

Algorytm wstecznej propagacji to typ algorytmu, który próbuje zredukować wartość docelową (minimalizacja błędu). Działa on w następujący sposób: aktywacja węzła wyjściowego obliczana jest na podstawie aktywacja węzła wejściowych i zbioru wag początkowych (ta część nazywana jest feed-forward). Następnie zaczyna się część wstecznej propagacji, w której błąd obliczony na wyjściu jest różnicą między wyliczonym wynikiem a wynikiem spodziewanym (np. wartości znajdujących się w zbiorze treningowym). Pozwala to wyliczyć błąd dla każdego węzła w warstwie ukrytej oraz wyjściowej. Następnie błąd każdego z węzłów jest użyty do ustawienia wag w taki sposób, by ten błąd został zminimalizowany. Taki proces powtarzany jest dla wszystkich rzędów zbioru treningowego. Prezentacja całego zbioru treningowego nazywa się epoką.

Przy trenowaniu sieci neuronowej należy uważać na to, by jej nie przeuczyć, gdyż taka sieć stanie się nieprzydatna ze względu na zbytnie dopasowanie się do danych treningowych.

Wady sieci neuronowych:

- dają wyniki trudne do zinterpretowania
- mają tendencję do przetrenowania
- potrzebują dużo czasu na naukę
- wymagają oczyszczenia danych

Zalety używania sieci neuronowych:

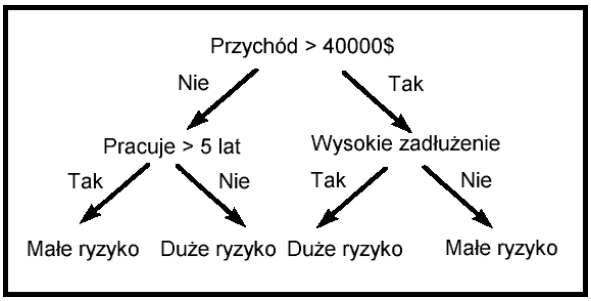
- można stosować je z łatwością na komputerach równoległych
- można je stosować dla danych ciągłych bez dodatkowych procedur dyskretyzujących
- stosuje się je do modeli o bardzo wielu wejściach
- wytrenowane dają szybko wyniki
- działają najlepiej, gdy jest dużo danych i stosunkowo wysoki poziom szumu względem danych.

4.2.2 Drzewa decyzyjne

Drzewa decyzyjne[1] są sposobem na reprezentację serii reguł, które prowadzą do klas lub wartości. Typowym zastosowaniem drzew decyzyjnych jest klasyfikacja, ale stosuje się je również do regresji.

Rozróżnić można tu dwie grupy - jedną małego, drugą dużego ryzyka. Każdy węzeł wykonuje jakiś test i - w związku z jego wynikiem - drzewo dzieli się na gałęzie. Najczęściej

[1] por. Systemy uczące się, Paweł Cichosz, Warszawa 2000, 138 i następne



Rysunek 4.2: Drzewo klasyfikacyjne

spotykanymi drzewami są drzewa binarne, czyli takie, które w wyniku działania dostają odpowiedź “tak” lub “nie”. Do takich właśnie algorytmów należy np. *CART*. Każdy węzeł prowadzi do następnych dwóch węzłów. Drzewa decyzyjne tworzą się w wyniku podziału danych na dyskretne grupy, których celem jest zmaksymalizowanie “dystansu” pomiędzy każdą z grup. Drzewa decyzyjne które używane są do przewidywania zmiennych ciągłych nazywamy drzewami regresyjnymi, a te drzewa, które klasyfikują dane - drzewami klasyfikacyjnymi.

Wady drzew decyzyjnych:

- taka sama dokładność wyników dla różnych drzew (tzn. powoduje problemy z interpretacją)
- każdy podział wartości atrybutu jest zależny od poprzedniego(jeśli zaczniemy od innego podziału otrzymamy inne drzewo)
- trudno jest wykrywać korelację między niezależnymi danymi

Zalety używania drzew decyzyjnych:

- stosunkowo łatwa interpretacja przewidywań
- można stosować je dla danych o dużej ilości zmiennych (wymaga dużej liczby wektorów treningowych, bo im niższa gałąź w drzewie tym mniej danych do niej dociera a liczba gałęzi zależy przede wszystkim od liczby atrybutów)
- szybko budują modele (można stosować do dużych ilości danych)
- możliwość przycinania drzewa
- dają sobie radę z danymi nienumerycznymi

4.2.3 K-nn i MBR

K-nn[2] to metoda k-najbliższych sąsiadów, zaś MBR - rozumowanie oparte na pamięci. Metoda K-nn decyduje, do jakiej klasy przynależy dany wektor, poprzez sprawdzenie jego otoczenia - k najbliższych (w sensie - podobnych) sąsiadów. Zostaje tu przeliczona liczba przypadków przypadających na każdą z klas i na tej podstawie przypisuje się nowy przypadek do klasy, do której należy większość jego sąsiadów.

Wady K-nn:

- złożoność obliczeń na etapie optymalizacji funkcji odległości lub liczby sąsiadów proporcjonalna jest do kwadratu liczby wektorów treningowych, a na etapie rozpoznawania wymaga obliczenia odległości od wszystkich wektorów treningowych

Zalety K-nn:

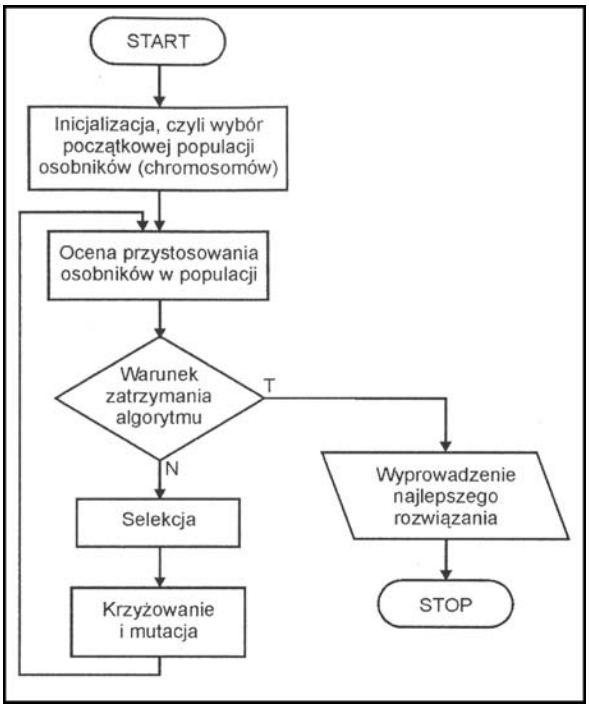
- radzi sobie z danymi nienumerycznymi
- potrafi wykryć asocjacje (może brać dowolny atrybut za klasę)
- łatwy w użyciu,
- nie ma parametrów,
- prosty, często działa bardzo dobrze,
- zastosowanie probabilistycznych miar odległości pozwala na używanie danych symbolicznych

4.2.4 Algorytmy genetyczne

Ten typ algorytmów - używany jest nie do znajdowania wzorców, lecz do przeprowadzania procesu uczenia w algorytmach takich, jak sieci neuronowe. Można powiedzieć, że zajmuje się on poszukiwaniami najlepszych modeli w przestrzeni rozwiązań. Każdy poprzedni model zostawia coś w spadku poprzedniemu, aż do momentu znalezienia najlepszego modelu. Algorytm ten posiada więc inspiracje biologiczne. Algorytmy genetyczne są metoda optymalizacji globalnej. Można ją zastosować do uczenia się innych modeli. Działanie podstawowego algorytmu genetycznego przedstawić można za pomocą schematu blokowego:

Inicjalizacja algorytmu polega na utworzeniu początkowej populacji osobników poprzez losowy wybór żądanej liczby chromosomów reprezentowanych przez ciągi binarne o określonej

[2]por. Duch W., Grudziński K., Sieci Neuronowe i Uczenie Maszynowe: próba integracji [w:] Biocybernetyka 2000, Tom 6: Sieci neuronowe pod red. W. Ducha, J. Korbicza, L. Rutkowskiego i R. Tadeusiewicza), Warszawa 2000, s. 666-667



Rysunek 4.3: Schemat blokowy działania algorytmu genetycznego.

długości. Ocena przystosowania osobników polega na obliczeniu pewnej funkcji zwanej funkcją przystosowania (musi być odpowiednio zdefiniowana w zależności od rozwiązywanego problemu). Warunek zatrzymania algorytmu określony może być na różne sposoby, np. dojście do żądanej wartości optymalnej, z określoną dokładnością lub przez zadanie czasu działania czy w sytuacji gdy jego działanie nie powoduje polepszenia uzyskanej wartości. W ostatnim kroku algorytmu (po spełnieniu jednego z warunków zatrzymania) jest wprowadzenie najlepszego chromosomu - czyli podania najlepszego rozwiązania.[3]

Zalety algorytmów genetycznych:

- algorytmy genetyczne pozwalają optymizować modele

4.2.5 Detekcja asocjacji i powtórzeń

Jej zadaniem jest znalezienie reguł, które występują razem. Detekcja powtórzeń jest podzbiorem asocjacji, gdyż tu występuje asocjacja powiązana z okresem czasu. Regułą asocjacyjną może być stwierdzenie: osoby kupujące soczewki kontaktowe nabywają także płyn do soczewek. Przyczyną jest tutaj kupno soczewek, skutkiem zaś nabycie płynu do ich pielęgnacji. Częstość, z jaką powtarza się dany związek w bazie danych mówi o jego wadze. Rozważmy

[3]por. Rutkowska D., Algorytmy genetyczne i ewolucyjne [w:] Biocybernetyka 2000, Tom 6: Sieci neuronowe pod red. W. Ducha, J. Korbicza, L. Rutkowskiego i R. Tadeusiewicza, Warszawa 2000, s. 691 i następne

przykładową bazę:

Ogólna liczba transakcji: 1000
Ilość pozycji zawierających “młotek” : 50
Ilość pozycji zawierających “gwoździe” : 80
Ilość pozycji zawierających “drewno” : 20
Ilość pozycji zawierających “młotek i gwoździe” : 15
Ilość pozycji zawierających “gwoździe i drewno” : 10
Ilość pozycji zawierających “młotek i drewno” : 10
Ilość pozycji zawierających “młotek gwoździe i drewno” : 5

Po przeliczeniu otrzymujemy:
Wagę dla “młotek + gwoździe” = $1,5\%(15/1000)$
Wagę dla “młotek + gwoździe + drewno” : $0,5(5/1000)$
Powiązanie “młotek” $\Rightarrow$ “gwoździe” : $30\%(15/50)$
Powiązanie “gwoździe” $\Rightarrow$ “młotek” : $19\%(15/80)$
Powiązanie “młotek, gwoździe” $\Rightarrow$ “drewno” : $33\%(5/15)$
Powiązanie “drewno” $\Rightarrow$ “młotek + gwoździe” : $25\%(5/20)$

Innym sposobem określania znaczenia poszczególnych asocjacji jest *lift* - parametr liczony jako stosunek:

$$Lift = \frac{\text{zaufanie do wynikania B z A}}{\text{częstość występowania B}}$$

Lift z “młotek $\Rightarrow$ gwoździe” : $3,72\%(30\%/8\%)$
Lift z “młotek + gwoździe” $\Rightarrow$ “drewno” : $16,5\%(33\%/2\%)$

Im większy jest ten parametr, tym większe jest prawdopodobieństwo wystąpienia czynnika B po pojawieniu się czynnika A.

4.2.6 Analiza dyskryminacyjna

Analiza dyskryminacyjna jest najstarszą matematyczna technika klasyfikacyjna. Dyskryminacja liniowa (LDA, Linear Discrimination Analysis) stosowana jest do klasyfikacji wektorów na dwie grupy, rozdzielane za pomocą hiperpłaszczyzny. Oprócz dyskryminacji liniowej stosuje się dyskryminację kwadratową (podział za pomocą powierzchni 2-stopnia). Dyskryminacja logistyczna stosuje również hiperpłaszczyznę rozdzielającą, zmieniając nieco kryterium optymalizujące sposób podziału.
Zalety:

- modele łatwe do interpretacji
- łatwy trening

- wrażliwość na wzorce (często stosuje się go w medycynie, biologii etc.)

Wady:

- do separacji używane są tylko proste i płaszczyzny

4.2.7 Regresja logistyczna

Regresja logistyczna posiada cechy podobne do regresji liniowej, jest ulepszoną wersją regresji, przewiduje zarówno ciągłe, jak i binarne wartości. Używana jest do przewidywania zmiennych binarnych (tzn. takich wartości jak tak/nie czy $0/1$). Zmienne, które są binarne nie mogą być modelowane przez regresję liniową. Model wówczas budowany jest za pomocą logarytmu z różnicy trafności. Ta transformacja nazywana jest logarytmem różnicowym. Jest on odpowiednikiem funkcji aktywacji w sieciach neuronowych. Stosunek:

$$\frac{\text{prawdopodobieństwo wystąpienia zdarzenia}}{\text{prawdopodobieństwo niewystąpienia zdarzenia}}$$

interpretuje się tak samo, jak szanse na wygranie w grach. Regresja przydatna jest w prognozowaniu np. zmienności szeregów czasowych lub przewidywaniu zmian wartości jakiejś funkcji.

4.2.8 Ogólne Modele Addytywne - Generalized Additive Models (GAM)

Jest to klasa rozszerzonych modeli: regresji logistycznej i logicznej. Jej model można zapisać jako sumę funkcji nieliniowych, po jednej na każdy predykat. Ogólne Modele Addytywne stosowane są w regresji oraz klasyfikacji

4.2.9 Multivariate Adaptive Regression Splines (MARS)

Podstawowym założeniem MARS-a jest zamiana nieciągłych gałęzi na węzły posiadające ciągły model przejściowy kształtowany przez parę prostych. Przy końcu procesu budowania modelu każda z linii jest zastępowana przez gładką funkcję sklejaną zwaną “spline”. Przez te modyfikacje ów model stracił cenną strukturę, która pozwalała CART-owi na podawanie reguł. MARS za to automatycznie znajduje i pokazuje listę ważnych predykatów oraz zależności miedzy nimi. Oczywiście, tak jak sieci neuronowe i drzewa decyzyjne, także MARS ma tendencję do przetrenowywania. Aby temu zaradzić stosuje się walidację skrośną. MARS stosowany jest zarówno do klasyfikacji jak i do regresji.

Rozdział 5

Klasyfikacja systemów data mining

W zależności od rodzaju danych przeznaczonych do drążenia lub w zależności od zastosowań, systemy data mining mogą wykorzystywać (integrować) techniki pochodzące z analizy danych przestrzennych, otrzymywania informacji, rozpoznawania wzorców, ekonomii, biznesu, bioinformatyki lub psychologii. W związku z różnorodnością dyscyplin, zaistniała potrzeba klasyfikacji tych systemów. Systemy drążenia danych można podzielić na różne sposoby:

- Klasyfikacja w zależności od rodzajów drążonych baz danych.
  Rodzaje drążonych baz danych można podzielić według różnych kryteriów, takich jak: modele danych, typy danych, ich zastosowania. Jeśli dokonamy klasyfikacji według modeli danych, wówczas otrzymamy następujące typy: relacyjne, transakcyjne, obiektowo zorientowane, obiektowo-relacyjne czy "hurtownicze" systemy drążenia danych. Jeśli podzielić DMS (data mining systems) według typów danych, na których one pracują, otrzymamy następujące systemy drążące: przestrzenne, time-series (czasowe), tekstowe, multimedialne czy World Wide Web.

- Klasyfikacja ze względu na rodzaj "wydobywanej" wiedzy.
  Możliwe jest utworzenie kategorii ze względu na rodzaj "wydobywanej" wiedzy, poprzez bazowanie na funkcjonalności drążenia, np. na charakteryzacji, dyskryminacji, klasyfikacji, analizy danych odstających i analizy ewolucyjnej. Dane mogą być przedstawiane na różnym poziomie abstrakcji - jako pewna generalizacja (wysoki poziom abstrakcji), wiedza prymitywna (na poziomie "surowych" danych), wiedza o wielu poziomach (rozważana na kilku poziomach abstrakcji). Zaawansowane systemy eksploracji danych powinny zdobywać wiedzę dotyczącą kilku poziomów abstrakcji. Inną możliwość stanowi rozróżnienie według regularności danych (często spotykane wzorce) lub przeciwnie - ze względu na wyjątki bądź dane odstające od normy.

- Klasyfikacja w zależności od używanych technik.
  Systemy można podzielić według poziomu interakcji z użytkownikiem (np. systemy autonomiczne, interaktywne systemy eksplorujące, systemy pracujące na zapytaniach), metod analizy danych (technik zorientowanych na konkretne typy baz danych, wizualizacji, uczenia maszynowego, statystyki, rozpoznawania wzorców, sieci neuronowych itd.).

- Klasyfikacja ze względu na zastosowania.
  Podział ten istnieje ze względu na dziedziny, w których te systemy znalazły zastosowanie. Jako przykład podać można systemy dopasowane do wymogów branży finansowej, telekomunikacyjnej, DNA, przemysłowej, e-mail, WWW (!). Dlatego często typowe systemy służące “do wszystkiego” nie nadają się do zastosowań specjalistycznych.

Rozdział 6

Ważniejsze zagadnienia związane z drążeniem danych

Jak w każdej dziedzinie także w drążeniu danych istnieje spora ilość problemów z którą muszą dać sobie radę programiści piszący oprogramowanie. Specjaliści powinni zwrócić szczególną uwagę na pewne problemy, gdyż od nich zależy funkcjonalność i użyteczność oprogramowania.

6.1 Analiza i interakcja z użytkownikiem

Ważniejszymi aspektami analizy i interakcji są:

- Analiza różnego rodzaju danych z baz. Każdy użytkownik może być zainteresowany innym rodzajem wiedzy, dlatego data mining musi pokrywać szerokie spektrum analizy danych i odkrywania wiedzy (różne sposoby analizy - wymienione prędzej).
- Interaktywne drążenie wiedzy o wielu poziomach abstrakcji. Jest to proces niezwykle ważny wówczas, gdy nie wiemy, co dokładnie może zostać odkryte w bazie; cały proces powinien być interaktywny. Pozwala to użytkownikowi skupić się na szukaniu wzorców.
- Dołączanie tła wiedzy. Wiedza i informacje dotyczące dziedziny, którą drążymy, może zostać wykorzystana podczas przeprowadzania procesu odkrywania wiedzy. Te dane pomogą skupić system na naszych celach poprzez podniesienie jego efektywności (wydobywanie ciekawszych danych) i zwiększenie szybkości.
- Języki zapytań data mining i ad hoc data mining. Relacyjne języki zapytań (jak SQL) pozwalają użytkownikom zadawać pytania wprost, w celu wyszukania danych. Podobny język musi zostać rozwinięty w systemach data mining i zoptymizowany dla podniesienia efektywności i elastyczności procesu eksploracji danych.
- Prezentacja i wizualizacja wiedzy. Odkryte informacje powinny zostać wyrażone poprzez języki wysokiego poziomu (język naturalny) przedstawione graficznie lub w inny sposób

czytelny dla użytkownika. Dzięki odkrytej wiedzy będą one zrozumiałe dla każdego (bez potrzeby zatrudniania sztabu specjalistów), co jest tym ważniejsze wówczas, gdy system jest interaktywny. Wiedza ta powinna być reprezentowana przez drzewa decyzyjne, reguły, wykresy, mapy, tablice, matryce (macierze) i krzywe.

- Odporność na szumy i braki w danych. Informacje zawarte w bazie danych mogą zawierać złe lub wyjątkowe dane. Należy zapobiegać poddawaniu analizie niewłaściwych danych, gdyż uzyskane rezultaty nie będą wówczas zadawalające. Ten fakt powoduje, że procesy czyszczenia danych i analizy wyjątków są potrzebne.

- Ocena wzorców. Systemy data mining są w stanie odkryć tysiące wzorców, z których wiele może przedstawiać powszechnie znaną wiedzę.

6.2 Wydajność

Z problemem wydajności związana jest skalowalność, efektywność i równoległość algorytmów analizujących dane.

- Efektywność i skalowalność - czas drążenia ogromnej ilości danych musi być przewidywalny i możliwy do zaakceptowania.

- Algorytmy równoległe i algorytmy wzrostu. Istniejące duże ilości baz danych zwracają uwagę na potrzebę równoległego ich przetwarzania. Algorytmy wzrostu analizują uaktualniane bazy danych bez potrzeby wykonywania ich drążenia od początku. Algorytmy równoległe dzielą dane na części i poddają je analizie równoległej; na końcu rezultaty są łączone.

6.3 Różnorodność baz danych

Od momentu, kiedy relacyjne bazy danych i hurtownie znalazły szerokie zastosowanie, zaistniał problem ich ekploracji.

- Relacyjne i kompleksowe bazy danych. Każdy rodzaj baz danych wymaga innego podejścia, innych algorytmów i innego rodzaju celów. Nie jest możliwe więc skonstruowanie jednego systemu przydatnego w procesie drążenia różnego rodzaju danych. Dlatego musi powstawać wiele systemów, których algorytmy są optymizowane w celu zastosowania ich na konkretnych danych.

- Drążenie informacji z heterogenicznych baz i globalnych systemów informacyjnych. Sieci lokalne i globalne (jak internet) łączą wiele źródeł danych i w ten sposób formują ogromną, heterogeniczną bazę danych. Odkrywanie wiedzy z różnych źródeł strukturalnych, pół strukturalnych lub niestrukturalnych jest ogromnym wyzwaniem dla data mining (Web mining).

Rozdział 7

Oprogramowanie komercyjne

Aktualny trend w dziedzinie obróbki informacji spowodował wzrost zainteresowania firm robiących oprogramowanie do obsługi baz danych i nastawieniem się na pisanie programów odkrywających ukryte zależności z zawartych w składnicach informacji. Opiszę teraz przykładowe programy komercyjne, zaś porównanie zawarłem w aneksie formie tabeli.

7.1 Accure Hit List

Program Hit List M/Audit pozwala na prezentacje efektywności istnienia w sieci. Intuicyjny interfejs graficzny powoduje, że informacje są łatwe do zrozumienia. Pełna analizy dostarcza podstawowe informacje w pojedynczym raporcie. Raport taki zawiera liczbę wizyt na stronie, popularność poszczególnych stron, adresy z których pochodzi największa liczba przekierowań i inne kluczowe wartości dla efektywności Web site. Posiada spis rzeczy, który pozwala na łatwą nawigację.

Raport “Marketing Overview” dostarcza podstawowych informacji biznesowych - zarówno o aktywności na stronach internetowych jak i trendach. Ten rodzaj raportu może być przeprowadzony w każdym okresie działalności firmy upewniając użytkownika, że zarówno taktyka dotyczące kampanii jak i strategia długo okresowa jest wybrana optymalnie. Taki rodzaj raportu może być wykorzystany, by ulokować odpowiednie fundusze na reklamę czy inną działalność, która przyniesie najlepsze efekty.

Raport “Advertisement versus Public Relations” porównuje “jakość”, odwiedzających odkrywając zależności miedzy konkretnym rodzajem (czy umiejscowieniem) reklamy i poźniejszą aktywnością użytkownika. Ten raport ocenia wynikające korzyści opierając się na wynikach pomiarów biznesowych zawartych w bazie danych, nie zaś na pomiarach samego ruchu.

Raport “Hit List Demographics” analizuje zachowanie się, preferencje i odwiedzających. Identyfikuje różnice w zachowaniu wynikające z zależności położenia geograficznego klienta (np. w zależności od państwa). Wskazuje w ten sposób najbardziej odpowiednich czy potencjalnych klientów, na których zwrócić należy szczególną uwagę.

Inne rodzaje raportów potrafią dostarczyć informacji w jaki sposób odwiedzający znajdują strony internetowe. Umożliwiają także porównanie ważności (tzn. stosunek zawartych transakcji od ilości wizyt) odwiedzających, co pomaga ocenić wpływ reklam w sieci i korzyści z ich umieszczania. Za pomocą Hit List możliwy jest również automatyczny billing. Dane są obrabiane w czasie rzeczywistym, więc raporty tworzone przez program są dokładne a wyniki analizy dostępne są w różnych formach.

System znalazł zastosowanie w:

- publicystyce
- handlu elektronicznym
- technologii
- ubezpieczeniach
- rozrywce

7.2 KnowledgeSTUDIO

KnowledgeSTUDIO został zaprojektowany zarówno dla analityków profesjonalnych jaki i dla ludzi zajmujących się biznesem. Każdy, kto zna pakiet Office firmy Microsoft nie powinien mieć problemów z obsługą programu. System jest łatwy do użycia i posiada intuicyjny interfejs. Profesjonalni analitycy powinni docenić, iż KnowldgeSTUDIO jest obiektowo zorientowany i pozwala na pełny proces data mining oraz na sprawdzanie rezultatów.

Dane do systemu mogą być importowane z różnego rodzaju formatów takich jak: ASCII, dBase, Excel, ODBC, SAS oraz z innych rodzajów pakietów statystycznych. KnowledgeSTUDIO do eksploracji i wizualizacji relacji w danych używa drzew decyzyjnych. Udostępnia także inne narzędzia wraz z drzewami (np. ich eksploracyjna natura może zostać użyta, by ocenić i wyjaśnić rezultaty otrzymane za pomocą sieci neuronowych. Możliwe jest także generowanie reguł z drzew decyzyjnych. KnowledgeSTUDIO pozwala również w łatwy sposób na podzielenie danych na część testową i treningową. Tak podzielony rodzaj danych może być użyty do trenowania sieci neuronowych.

Klient KnowledgeSTUDIO działa na komputerach klasy PC wyposażonych w Windows 9x lub NT. KnowledgeSERVER działa na komputerach wyposażonych w Windows NT Server i Windows NT Workstation. Planowana jest także wersja dla systemu Solaris. Do wspieranych przez system formatów baz danych należą: Access, dBase(II, III i IV), ODBC, SAS oraz SPSS. Algorytmy zawarte w systemie to CHAID, XAID, K-Means, Entropy decision tree, oraz MLP, RBF, probabilistyczne sieci neuronowe i regresje (liniowa i logistyczna). System KnowledgeSTUDIO znalazł zastosowanie w:

- telekomunikacji
- finansach i ubezpieczeniach
- produkcji

7.3 XpertRule® Miner

XpertRule® Miner wyprodukowany przez Attar Software jest skalowalnym produktem data mining o architekturze klient/server. Używa technologii ActiveX może być używany na wiele sposobów. Może tworzyć rozwiązania jako samodzielny system data mining na pojedynczym komputerze lub jako jeden z komputerów sieci intranet czy internet. Klient The ActiveX Miner współpracuje również z wysoko wydajnymi serwerami data mining w celu dostarczenia informacji z bardzo dużych baz danych. Miner zawiera rozszerzone rodzaje transformacje danych, wizualizację oraz możliwość tworzenia raportów. Dane mogą być manipulowane za pomocą mechanizmu “drag and drop”. Użytkownik może samodzielnie projektować graficznie spersonalizowane przez siebie procesy drążenia i raportowania. Możliwa jest pośrednia kontrola aplikacji przez wybieranie właściwości obiektów oraz stosowanych metod. Pozwala to na budowanie równolegle działających aplikacji (np. za pomocą języków VB, Delphi etc.) takich jak systemy zarządzania relacjami z klientem. Wszystko to może zostać osiągnięte bez straty skalowalności czy wydajności.

Wizualizacja i eksploracja danych mogą być uważane za krok w procesie drążenia. Dane są filtrowane przez profile drzew decyzyjnych a następnie generowany zostanie raport. Wszystkie rodzaje graficznego przedstawiania informacji (wykresy, mapy, tabele) są połączone z graficznym drzewem pokazującym dane.

Tymczasowe raporty mogą być wykorzystane do monitorowania postępów. Dane wejściowe i wyjściowe mogą pochodzić ze źródeł kompatybilnych z ODBC. Program pozwala na manipulację tablicami danych (użytkownik może łączyć i dzielić tabele z danymi, filtrowanie tablic - istnieje możliwość wydobycia kolumny czy rzędu według określonego kryterium; możliwe jest porządkowanie tablic według określonego porządku). Dodatkowymi właściwościami systemu są: stosowanie logiki rozmytej oraz optymalizacja za pomocą algorytmów genetycznych.

System XpertRule® jest używany w:

- badaniach naukowych (Rockwell Aerospace i NASA)
- energetyce
- finansach
- produkcji

Rozdział 8

Zakończenie

Technologia baz danych przeszła długą drogę od prymitywnej obróbki plików do baz z obsługą zapytań. Dalszy postęp prowadzi nas do stworzenia sprawniejszych i bardziej efektywnych narzędzi potrzebnych do zrozumienia danych. Użytkownik chce wydajnej analizy danych oraz jej zrozumienia, nie chce już widzieć tylko liczb, ale pragnie poznać zależności obecne w dużych ilościach danych, co wcześniej było niemalże niemożliwe. Ta potrzeba wyniknęła z eksplozji zbieranych danych i ich zastosowań w różnych dziedzinach, np. w zarządzaniu i biznesie, administracji, nauce i kontroli środowiska.

- *Data mining* jest procesem odkrywania zależności, które są ukryte w dużych ilościach danych mieszczących się w bazach lub hurtowniach. Data Mining to młoda, interdyscyplinarna dziedzina, obejmująca zagadnienia dotyczące systemów baz danych, hurtowni danych, statystyki, uczenia maszynowego, wyszukiwania i wizualizacji danych oraz symulacji. Drążenie zasilają także sieci neuronowe, rozpoznawanie wzorców, analiza danych przestrzennych, bazy obrazów, obróbka sygnałów, a także inne dziedziny, w których jest ono stosowane (biznes, ekonomia, bioinformatyka).

- *Proces odkrywania wiedzy* zawiera czyszczenie i integrację danych, selekcję, transformację, drążenie danych, a następnie rozpoznawanie wzorców i ich prezentację.

- *Wzorce* można “wydobyć” z różnego rodzaju systemów przechowywania danych. Możliwe jest także znalezienie ciekawych informacji w innych źródłach (np. multimedia i oczywiście WWW).

- *Hurtownie danych* są składnicami dla danych zbieranych przez długi okres czasu z wielu źródeł. Składnica tego rodzaju posiada także pewnego rodzaju możliwości analizy.

- *Cele data mining* obejmują odkrywanie opisów klas, asocjację, klasyfikację, klastering, przewidywania, analizę trendów, odchyleń i podobieństw.

- *Klasyfikacji systemów data mining* można dokonć według różnego rodzajów kryteriów np. typu drążonych baz, wiedzy, używanych technik lub zastosowań.

- *Wydajność i efektywność* systemów jest bardzo ważna i stanowi wielkie wyzwanie dla projektantów. “Problemy” związane z metodologią, interakcją z użytkownikiem, a także zastosowanie systemów D-M oraz ich wpływ na środowisko są tu nie bez znaczenia.

Bibliografia

1. Cezary Głowiński: Sztuka wysokiego składowania, PC Kurier 12, 2000
2. Jiawei Han, Micheline Kamber: Data Mining concepts and techniques, 2000
3. Systemy uczące się, Paweł Cichosz, Warszawa 2000,
4. Biocybernetyka 2000, Tom 6: Sieci neuronowe, pod red. W. Ducha, J. Korbicza, L. Rutkowskiego i R. Tadeusiewicza), Warszawa 2000,

Źródła z internetu:

1. Firma Accure - producent programu *Accrue Hit List* - http://www.accrue.com/
2. Firma Attar - producent programu *XpertRule Miner* - http://www.attar.com/
3. Firma Angoss - producent programu *KnowledgeSudio* - http://www.angoss.com/
4. Firma ThinkingMachines - producent programu *Darwin* - http://www.think.com/
5. Firma SPSS - producent systemu *SPSS* - http://www.spss.com/
6. Polskie strony firmy SPSS dotyczące drążenia danych w internecie - http://www.webmining.pl/
7. Firma SAS - producent programu *EnterpriseMiner* - http://www.sas.com/
8. Firam SGI - producent programu - http://www.sgi.com/solutions/
9. Firma IBM - producent programu *IBM Intelligent Miner* - http://www-4.ibm.com/software/data/
10. Strona poświęcona wydobywaniu wiedzy i zastosowaniom D-M - http://www.kdnuggets.com/

Aneks

PRODUCENT	PRODUKT	Używane modele sieci neuronowych	Uzywane rodzaje drzew decyzyjnych	Klasyfikacja	Regresja	Przewidywanie szeregów czasowych	Clastering	Odkrywanie asocjacji	Detekcja powtórzeń	Dane nieznane
AbTech	ModelQuest	BP	C4.5	regresja logistyczna, K-nn, sieci statystyczne, C4.5	sieci statystyczne, K-nn, regresja liniowa	?	-	-	-	?
Angoss	KnowledgeSeeker	N/A	ID3, CART, CHAID	+	+	-	-	-	-	?
Angoss	KnowledgeSTUDIO	MLP, BP, probabilistyczne sieci neuronowe, RBF (K-średnich)	CHAID, Gini, entropia, wariancja, ANOV A/F; dyskretyzacja warotści ciągłych	sieci neuronowe, drzewa decyzyjne	sieci neuronowe, drzewa decyzyjne	-	K-średnich	-	-	ignoruje, średnia, min, max
Attar	XpertRule Miner	brak	ProfilerX (oparty na ID3), wartści dyskretne dzieli między dwie gałęzie; F-test, Chi-kwadrat	+	+	+	-	+	-	?
Buisiness Objects	BuisnessMiner	?	Gini, automatyczna rozmyta dyskretyzacja	drzewa decyzyjne, Gini	drzewa decyzyjne	?	?	?	?	?
Cognos	4Thought	feed-forward z BP; wybór archit. sieci neuronowych	--	-	+	+	-	-	-	?
Cognos	Scenario	--	Bonferroni	CHAID	drzewo regresyjne, CHAID	4Thought	?	?	?	domyślnie wykluczane
Group1	Model 1	MLP BP	CHAID, CART, Chi-kwadrat	regresja liniowa i logistyczna, CHAID, CART, Naive Bayes, sieci neuronowe	sieci neuronowe, regresja liniowa	makra	K-średnich, CHAID	?	?	interpretowane jako zero przez sieć neuronową

HNC Software	DataBase Mining Marksman	MBPN,	N/A	+	+	+	+',neural gas, Batch Neural Gas (Neural Gas + K-średnich)	-	-	zamieniane (automatycznie lub ręcznie)
ISL	Clementine	RBF, Kohonen,	C5.0	MLP BP, RBF, C5.0	MLP, RBF, rozszerzony ID3, regresja liniowa	sieci neuronowe, indukcja reguł	Kohonen, K-średnich	GRI, wizualizacja sieciowa	GRI	C5.0
IBM	IBM Intelligent Miner for Data	MLP, RBF	binarne drzewa decyzyjne, Gini,	CART, sieci neuronowe, BP	regresja logistyczna, sieci neuronowe, (BP i regresja nieliniowa) RBF, dopasowywanie krzywych	Univariate Curve Fitting	NN (kohonen feature maps),	+	+	średnia
Magnify	PATTERN	brak	binarne drzewa decyzyjne, Gini,	ACT, CART,	ART, CART	ART.	K-średnich	+	-	?
MathSoft	S-PLUS	N/A	binarne drzewa decyzyjne, CART,	GAM, classification trees, logistic regression	ANOVA, różne typy regresji, ACE, AVAS, LOESS (lokalna regresja liniowa)	AR, MA, ARIMA	K-średnich, analiza rozmyta, klasteryzacja oparta na moidelu	możliwość zaprogramowania w języku S	możliwość zaprogramowania w języku S	?
NCR	NCR TeraMiner	MLP BP, RBF, Kohonen,	C5.0	MLP, RBF, C5.0	MLP, RBF, rozszerzony ID3, regresja liniowa	sieci neuronowe, indukcja reguł	Kohonen, K-średnich	GRI, wizualizacja sieciowa	GRI	?
NEOVISTA SOFTWARE	Decision Series	?	pruned using pessimistic assumption of error rate	drzewa decyzyjne, sieci Bayesowskie	Decission Net, Bayes, Cubist	-	K-średnich	Decision AR	Decision AR	?
Quadstone		N/A	ID3, drzewa klasyfikacyjne i regresyjne	drzewa decyzyjne oraz scorecards	drzewa decyzyjne oraz scorecards	-	-	-	-	?

Salford	CART	-	CART, Gini, twoing,	CART	CART		opiera się na drzewach	opiera się na drzewach	-	?
Salford	MARS	N/A	N/A	+	+	?	?	?	?	?
SAS	EnterpriseMiner	MLP, RBF, Levenberg-Marquardt, Quasi-Newton, wiele funkcji aktywacji	CART, CHAID	sieci neuronowe, drzewa decyzyjne, regresje	OLS, ogólne modele liniowe, regresja logistyczna	?	+	+	+	?
Silicon Graphics	MineSet	-	C4.5	drzewa decyzji (C4.5), prosty i naiwny klsasyfikator Bayesowski, drzewa opcji, tablice decyzyjne	CART	+	K-średnich, ACPro (AutoCl)	+	-	?
SPSS	SPSS	MLP, RBF, Kohonen, Bayesian	CHAID, QUEST	CHAID, drzewa klasyfikacyjne i regresyjne, QUEST, cross tabulacja, regresja logistyczna, probit regresion	siecu neuronowe, drzewa klasyfikacyjne i regresyjne, liniowe dopasowanie krzywych, dopasowywanie wag	autoregresja, ARIMA, sieci Bayesian , MLP, RBF, Kohonen	K-średnich, analiza klastrowa, analiza dyskryminacyjna, analiza czynnikowa	?	?	?
ThinkingMachines	Darwin	trening i test, walidacja skrośna, BP, algorytmy genetyczne	C&RT, gini, entropia	C&RT, drzewa decyzyjne, sieci neuronowe (BP, algorytmy genetyczne), K-nn, MBR	C&RT, drzewa decyzyjne, sieci neuronowe (BP, algorytmy genetyczne), K-nn, MBR	+	K-średnich	-	-	?

Torren	Orchestrate Analytics	BP, conjugate gradient, quick prop, simulated annealing, adaptive boise	?	drzewa decyzyjne, sieci neuronowe, drzewa neuronowe, probabilistyczne sieci neuronowe	drzewa decyzyjne, sieci neuronowe, drzewa neuronowe, probabilistyczne sieci neuronowe, RBF	algorytmy regresji	Kohonen, K-śrenich	+	+	?
Trajecta	dBProphet	BP	-	+	+	+	K-średnich, kohonen,	-	-	?
Unica Tech.	PRW	MLP (BP, algorytm genetyczny)	-	regresja logistyczna i liniowa, K-knn, MLP BP, RBF	regresja logistyczna i liniowa, K-knn, MLP BP, RBF	+	K-średnich	-	-	?
Urban Science	GainSmarts	-	CAHID, algorytm genetyczny	CHAID, standardowe drzewo (AID), algorytm genetyczny	Dicsrete Choice, Multinominal Choice, Ordered Choice, Two Stage, Continuous Mulitivariate and Tobit Regression	<-----	-	-	-	średnia, median