Eksploracja danych dla telekomunikacji

Mieczysław Muraszkiewicz
(htttp://www.icie.com.pl/mrm.htm ; mietek@mimuw.edu.pl)
Instytut Informatyki Teoretycznej i Stosowanej PAN,
Instytut Informatyki Politechniki Warszawskiej

Streszczenie. Artykuł wprowadza w problematykę eksploracji danych (*ang. data mining*) i pokazuje jakie są możliwości wykorzystania eksploracji danych w telekomunikacji. Opisano cztery techniki eksploracji danych, a mianowicie: klasyfikację, regresję, grupowanie i kojarzenie. W celu przybliżenia omawianej problematyki podano przykład eksploracji danych.

1. Wstęp

Spójrzmy na pewne dwie równolegle występujące w informatyce tendencje: pierwsza – zachodzi w świecie zastosowań, druga – w świecie badań.

W obszarze zastosowań obserwujemy w ostatnich trzech dekadach nadzwyczaj szybki i powszechny rozwój systemów informacyjnych, a zwłaszcza ogromne przyspieszenie, które w tym względzie spowodował Internet. Właściwa ludziom skłonność do dokumentowania swych działań i gromadzenia informacji oraz długotrwałego ich przechowywania sprawiły, że istniejące zasoby informacyjne zawarte w różnorakich bazach danych są niezwykle duże i stale rosną. Danych tych jest tyle, że ich pełna i pogłębiona analiza jest niezwykle trudnym, czasochłonnym i kosztownym przedsięwzięciem. A jednocześnie doświadczenie i intuicja podpowiadają, że w tym oceanie informacji może być ukryta nieznana nam, acz prawdopodobnie cenna i pożyteczna wiedza o świecie, z którego te informacje pochodzą.

Nie dziwi zatem pytanie właścicieli bardzo dużych baz danych, w rodzaju operatorów telekomunikacyjnych, globalnych sieci handlowych, czy banków, o to czy istnieją – a jeśli tak, to jakie – metody odkrywania ukrytej w tych bazach wiedzy. Pytanie takie nie jest zapewne motywowane ciekawością poznawczą potentatów gospodarczych, chodzi raczej o opanowanie i włączenie do swych rutynowych prac techniki, która zapewni przewagę konkurencyjną na rynku i pozwoli zwiększyć zyski. Tą techniką ma być *odkrywanie wiedzy w bazach danych*.

Co do obszaru badań informatycznych, to wśród informatyków uprawiających refleksję nad stanem i rozwojem ich dziedziny coraz częściej i wyraźniej artykułowane są opinie, że po skutecznym wyposażeniu komputerów w środki operowania na liczbach i przetwarzania tekstu nadszedł czas, aby wykorzystać je do zrozumienia zasad rządzących światem, w którym żyjemy. Richard Hamming powiada wprost: „celem i przedmiotem przetwarzania komputerowego jest wgląd w nasz świat, a nie liczby". Chodzi więc o to, aby komputery stały się narzędziami do badań o charakterze epistemologicznym.

Bez ryzyka pomyłki można powiedzieć, że odkrywanie wiedzy i pomoc w rozumieniu otaczającego nas środowiska niebawem nabiorą większego znaczenia niż klasyczne zastosowania komputerów takie, jak automatyzacja magazynów, optymalizacja produkcji, projektowanie wspomagane komputerowo itd. Gio Widerhold ze Stanford University twierdzi, że „odkrywanie wiedzy staje się najbardziej pożądanym produktem końcowym przetwarzania komputerowego, i że znaczenie wiedzy uzyskiwanej w ten sposób jest tak duże, iż tylko zabiegi mające na celu ochronę środowiska naturalnego mają większą wagę". Opinia ta znajduje potwierdzenie w stwierdzeniu Johna Naisbetta, który powiedział, że „choć toniemy w informacji, to najbardziej potrzebujemy wiedzy".

Terminy *dane, informacja, wiedza* nie poddają się łatwo definiowaniu i od dawna, jeśli nie od początku ich istnienia, są przedmiotem kontrowersji; w artykule tym zakładamy, że intuicja Czytelnika w tym względzie jest w zgodzie z najczęstszym rozumieniem tych terminów.

Artykuł ten ma następującą budowę. W rozdziale drugim wyjaśnimy termin *eksploracja danych,* po czym spróbujemy uzasadnić dlaczego warto korzystać z eksploracji danych (rozdział trzeci), następnie w rozdziale czwartym omówimy ważniejsze techniki eksploracji takie, jak klasyfikacja, regresja, grupowanie i kojarzenia. Kolejny, piąty rozdział, jest poświecony dyskusji na temat tego czym eksploracja danych nie jest. Dalej, w rozdziale szóstym, w celu lepszego przybliżenia problematyki przeanalizujemy wyimaginowany przykład, który posłuży do przeprowadzenia eksploracji danych. Rozdział siódmy zarysuje strukturę procesu eksploracji danych, po czym - w rozdziale ósmym - wyjaśnimy termin *odkrywanie wiedzy* i relację tego terminu z eksploracją danych. Rozdział dziewiąty w całości poświęcimy eksploracji danych w telekomunikacji.

2. Eksploracja danych

Rozważania rozpoczniemy od terminu węższego niż odkrywanie wiedzy, a mianowicie od terminu *eksploracja danych* (*ang. data mining*). W największym skrócie rozumie się przez nią odkrywanie z dostępnych zasobów danych różnego rodzaju uogólnień, regularności, prawidłowości, reguł, a zatem czegoś, co stanowi pewną wiedzę zawartą *implicite* w tych zasobach.

Eksploracja danych jest obecnie jednym z najżywiej rozwijanych tematów w informatyce. Jest przedmiotem rozległych badań, dyskusji, także sporów. Powstają czasopisma poświęcone tej dziedzinie, odbywają się liczne konferencje oraz doskonale funkcjonują ośrodki internetowe zajmujące się tą tematyką (np. *www.kdnuggets.com*). Jest to zatem dziedzina młoda, w trakcie poszukiwania i tworzenia własnej tożsamości, metodologii i narzędzi. Nie dziwi więc, że środowisko nie dopracowało się uznanych przez wszystkich szczegółowych definicji używanej terminologii, a w tym tak podstawowych terminów jak, eksploracja danych, czy odkrywanie wiedzy w bazach danych (*knowledge discovery in databases*). O wzajemnej relacji tych dwóch terminów powiemy rozdziale ósmym.

Eksploracja danych i odkrywanie wiedzy przyciągają wiele uwagi i wywołują emocje zarówno w środowiskach badawczych, jak i wśród grup przemysłowych, w biznesie, bankowości, handlu, ubezpieczeniach itp. Prowadzi się sporo projektów z tego zakresu, wciąż jednak nie do końca wiadomo jakie są możliwości eksploracji i odkrywania wiedzy, w jakich obszarach można je stosować najskuteczniej i jakimi do tego celu posługiwać się metodami. Ważne więc jest w takim nieustalonym stanie umieć oddzielić nadzieje i obietnice od istniejących realnie możliwości.

Sama idea eksploracji danych i odkrywania wiedzy jest niezwykle prosta i bez przeszkód odwołuje się do ludzkiej wyobraźni. Trzeba jednak od razu mocno podkreślić, że praktyczna realizacja tej łatwej w zrozumieniu idei jest przedsięwzięciem technologicznie i organizacyjnie złożonym, niekiedy bardzo trudnym. Potrzebne tu są zaawansowane środki programistyczne, nietypowa organizacja pracy oraz – bardzo często – sięgnięcie po kosztowne konsultacje specjalistyczne.

W tym artykule przez eksplorację danych rozumiemy proces automatycznego odkrywania znaczących, pożytecznych, dotychczas nieznanych i wyczerpujących informacji z dużych baz danych, informacji ujawniających ukrytą wiedzę o badanym przedmiocie; wiedza ta przyjmuje postać reguł, prawidłowości, tendencji i korelacji, i jest następnie przedstawiana przygotowanemu do jej spożytkowania użytkownikowi w celu rozwiązania stojących przed nią/nim problemów i podjęcia istotnych decyzji.

Po tej nieco zawiłej definicji spójrzmy na eksplorację przez pryzmat jej dowcipnego określenia: „eksploracja danych polega na torturowaniu danych tak długo, aż zaczną zeznawać". Inne, równie opisowe spojrzenie na eksplorację zawiera się w poleceniu, które chciałoby się skierować do bazy danych: „pokaż mi nie tylko to, co widzę gołym okiem (twoje zasoby), pokaż także to, czego nie widzę".

Tak więc zasadniczym celem eksploracji danych jest sięgnąć możliwie najgłębiej do dostępnych zasobów informacyjnych, po to aby odpowiedzieć na pytania użytkownika o regularności i prawidłowości istniejące w świecie reprezentowanym przez te zasoby, aby móc zweryfikować hipotezy statystyczne dotyczące tego świata czy po to, aby skutecznie prognozować.

3. W jakim celu prowadzić eksplorację danych ?

Praktyczne pożytki ekstrahowania danych ujawniają się w dwóch dziedzinach, którymi są:

– prognozowanie (ang. prediction, forecasting),

– *opis (*ang. description*)*.

Prognozowanie polega na wykorzystaniu znanych w chwili obecnej wartości interesujących nas zmiennych (lub pól w bazie danych) w celu przewidywania wartości tych lub innych zmiennych w przyszłości. Na przykład, model prognostyczny opracowany dla banku dotyczący pożyczek korzysta z historii kont osób zabiegających o pożyczki, pomagając wskazać tych, którzy prawdopodobnie będą mieli trudności ze spłaceniem pożyczek.

Opis polega na tworzeniu czytelnej i zrozumiałej dla człowieka reprezentacji wiedzy wydobytej z danych w postaci wykresów, wzorów, reguł, tabel. Opisy takie, w postaci modeli deskrypcyjnych, są często używane do wspomagania procesów decyzyjnych.

Firma IBM wymienia m.in. następujące różnego rodzaju powody, które zachęcają do prowadzenia eksploracji danych:

– w dużych bazach danych zawarta jest cenna, ukryta wiedza, która może okazać się przydatna w prowadzeniu różnorakich prac i rozumieniu otoczenia,

– istnieje potrzeba konsolidacji rekordów bazy danych w celu zapewnienia spójnego, jednolitego jej obrazu w oczach użytkownika (może to miedzy innymi prowadzić do budowy hurtowni danych),

– należy zmniejszać koszty przechowywania i przetwarzania danych,

– konkurencja na rynku wzmaga się i wymusza większą produktywność,

– nasila się tendencja do indywidualizowania produkcji oraz wyszukiwania i zajmowania niewielkich nisz rynkowych.

Oto trzy przykładów skutecznego zastosowania eksploracji danych: (i) firma American Express podała, że wykorzystanie technik eksploracji na bazie danych klientów pozwoliło zwiększyć o 10 – 15 % użycie jej kart kredytowych; (ii) inna duża firma oferująca karty kredytowe dzięki eksploracji potrafiła określić 5-cio procentowy segment wszystkich swych klientów, którzy charakteryzują się tym, że regularnie udzielają odpowiedzi na różne zapytania firmy. Klienci ci dostarczali 60 % wszystkich odpowiedzi. Dzięki ustaleniu tego faktu firma zwiększyła 12-krotnie stopę odpowiedzi i zmniejszyła koszty opłat pocztowych o 95 %; (iii) poważna firma telekomunikacyjna za sprawą przeprowadzonej analizy danych drogą eksploracji odkryła, że istnieje podgrupa użytkowników, którzy przez 3 miesiące w roku nie korzystają z usług. Informacja ta spowodowała opracowanie specjalnego programu zachęt dla tych użytkowników, co przyniosło doskonałe rezultaty komercyjne.

4. Techniki eksploracji

Najczęściej eksplorację danych wiąże się z następującymi typami działań:

– *klasyfikowanie* (ang. classification*)*,

– *regresja* (ang. regression),

– *grupowanie (*ang. clustering*)* ,

– *kojarzenie* (ang. association).

Dla porządku odnotujmy, że pełniejsza lista rodzajów działań, które mogą być wykorzystane do eksploracji byłaby znacznie dłuższa. Poniżej pokrótce omówimy poszczególne typy działań.

Klasyfikacja

Jest ona prawdopodobnie najczęściej stosowaną techniką eksploracji danych. Klasyfikacja jest procesem uczenia się, którego celem jest określenie reguły, która – kiedy już została zaakceptowana – służy do przyporządkowania (zaklasyfikowania) branego pod uwagę elementu do jednej lub więcej wcześniej zdefiniowanych klas (zbiorów). Proces ten korzysta ze zbioru wcześniej poklasyfikowanych przykładów, po to aby określić sposób (model) klasyfikowania całej dostępnej populacji elementów. Ten typ analizy daje szczególnie dobre wyniki przy wykrywaniu nadużyć oraz przy identyfikowaniu tych próśb o zasoby, gdzie istnieje duże ryzyko ich zmarnowania.

Klasyfikacja często korzysta z algorytmów opartych na drzewach decyzyjnych lub sieciach neuronowych. Użycie tych algorytmów rozpoczyna się od podania im w ramach uczenia się (treningu) zbioru przykładów już sklasyfikowanych. W wypadku wykrywania nadużyć, zbiór taki zawierałby przypadki (przykłady) gdzie wystąpiło nadużycie oraz przypadki „uczciwe".

Regresja

Regresja również korzysta z procesu uczenia się, z tą różnicą w stosunku do klasyfikacji, że powstaje tu funkcja (a nie odwzorowanie), która danemu elementowi przyporządkowuje konkretną wartość. Przykładem jej zastosowania jest przewidywanie popytu na nowy produkt w zależności od wydatków na reklamę. Jeśli zmienne wykorzystywane w modelach opartych na regresji mają złożoną naturę (np. wielkość sprzedaży, wskaźniki giełdowe), to zwykle do zaimplementowania regresji korzysta się z sieci neuronowych, a to z uwagi na ich przydatność w „sytuacjach nieliniowych".

Grupowanie

Grupowanie polega na przyporządkowaniu branego pod uwagę elementu do jednej lub wielu grup (klas, zbiorów), przy czym grupy te są wyznaczana przez sam proces grupowania na podstawie analizy danych o wszystkich dostępnych elementach, a nie jak w przypadku klasyfikacji, gdzie klasy zostały zdefiniowane wcześniej, niejako poza procesem klasyfikacji. Grupy wyznaczane są na podstawie pewnych czynników albo wskazujących na podobieństwa elementów albo opartych na przyjętych rozkładach prawdopodobieństwa, albo korzystających z jeszcze innych przesłanek.

Grupowanie jest szczególnie przydatne w rozwiązywaniu problemów segmentowania. Algorytm grupowania wyznacza czynnik dywersyfikujący elementy rozważanej populacji, definiuje grupy (segmenty) i przyporządkowuje do nich poszczególne elementy. Grupowanie jest często pierwszym etapem w eksploracji danych: po wyznaczeniu segmentów można do nich zastosować inne techniki w zależności od oczekiwanych rezultatów.

Kojarzenie

Kojarzenie polega na odszukiwaniu tych elementów, które wiążą się z zadanym zdarzeniem lub innym elementem. Algorytmy tu wykorzystywane pozwalają odkrywać reguły, które przyjmują postać:

> *jeśli* element A jest składnikiem danego zdarzenia, *to* w X % przypadków element B jest także składnikiem tego zdarzenia

na przykład:

> *jeśli* klient kupuje płatki owsiane, *to* w 65 % przypadków klient ten kupi mleko „Łaciate"

Jest rzeczą ciekawą, że zainteresowanie kojarzeniem niezwykle wzrosło wraz z upowszechnieniem się w handlu detalicznym czytników kodów paskowych, co pozwala zbierać ogromne ilości danych już „skojarzonych" w koszyku kupującego. Z tego powodu zapewne ten rodzaj analizy jest nazywany niekiedy *market-basket analysis*. Kojarzenie jest także stosowane do opracowywania kampanii marketingowych czy analizy portfeli inwestycyjnych.

Pewną odmianą kojarzenia jest uwzględnienie czynnika czasu. Na przykład,

> j*eśli* w czasie operacji wykonana zostanie procedura X, *to* w 45 % przypadków zakażenie Y pojawi się w ciągu 5 dni

Zakończmy ten rozdział następującym podsumowaniem: klasyfikacja i regresja są szczególnie pożyteczne i skuteczne do tworzenia prognoz, czyli do przewidywania zdarzeń, grupowanie i kojarzenie natomiast doskonale nadają się do opisu procesów (zachowań) jakie mają miejsce w świecie, o którym dane znajdują się w bazie.

5. Czym eksploracja danych nie jest ?

W uzupełnieniu do definicji eksploracji danych warto podkreślić czym eksploracja nie jest. A to dlatego, że nieporozumienia i nadmierne, niekiedy nawet fałszywe oczekiwania w kontekście eksploracji danych zdarzają się stosunkowo często. A zatem eksploracja danych nie jest:

- odkrywaniem wiedzy; jest ona tylko częścią procesu odkrywania wiedzy, o czym powiemy więcej w rozdziale ósmym,
- nieodzownie związana z hurtowniami danych; eksploracja może być prowadzona na dowolnej bazie, choć naturalnie hurtownie są szczególnie dobrymi miejscami do jej uprawiania,
- typowym narzędziem analitycznym i środkiem do tworzenia sprawozdań. Zasadnicza różnica pomiędzy eksploracją a typowymi narzędziami analitycznymi polega na podejściu do eksploracji danych i badaniu występujących pomiędzy nimi relacji. Otóż narzędzia analityczne, w tym OLAP (*ang. On-Line Analytical Processin*) stosuje się głównie do weryfikowania hipotez wysuniętych przez analityka; nie mogą one natomiast same tworzyć hipotez, odkrywać zasad i reguł – a to jest właśnie możliwe za pomocą technik eksploracji danych,
- uczeniem się maszyn (*ang. machine learning/discovery*), które dotyczy odkrywania praw empirycznych na podstawie obserwacji i eksperymentów,
- całkowicie zautomatyzowanym procesem; eksploracja danych jest w ogromnym stopniu uzależniona od prowadzącego ją człowieka, który określa warunki początkowe, dobiera metody eksploracji i ocenia otrzymane rezultaty i wreszcie to on decyduje czy uzyskane zależności są interesujące, czyli czy mają jakąkolwiek wartość praktyczną lub poznawczą dla organizacji, na której zlecenie eksploracja jest prowadzona,
- łatwym, tanim i szybkim do wdrożenia w organizacji procesem. Włączenie eksploracji danych do rutynowych operacji organizacji wymaga starannych prac przygotowawczych, eksperymentowania i współpracy ekspertów w zakresie eksploracji danych i specjalistów w dziedzinie, której dane dotyczą. Typowy projekt trwa wiele miesięcy, a nawet lat, jest miejscem gdzie uczą się wszystkie zaangażowane strony; oprogramowanie narzędziowe jest raczej kosztowne (od kilku tysięcy do kilkuset tysięcy dolarów), a eksploatacja i pielęgnacja systemu wymagają znakomicie wyszkolonego i godnego zaufania personelu,
- przysłowiowym, wielozadaniowym scyzorykiem armii szwajcarskiej dobrym na wszelkie okazje (ta opinia bierze się albo z nadmiernego entuzjazmu w odniesieniu do potencjału tkwiącego w technikach eksploracji danych albo jest wynikiem nieuczciwego prezentowania ich możliwości przez sprzedawców oprogramowania i konsultantów)

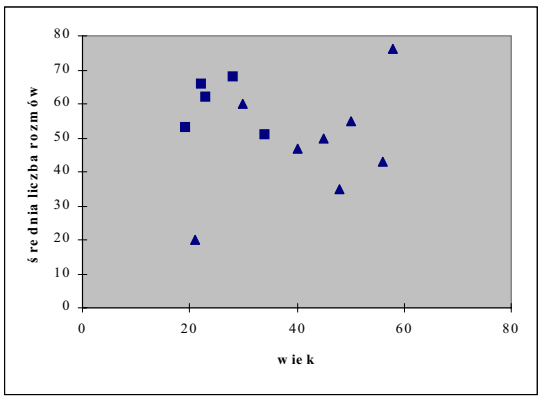
6. Przykład

W celu lepszego wyjaśnienia na czym polega eksploracja danych rozważmy wyimaginowaną sytuację w firmie telekomunikacyjnej. Kierownictwo tej firmy zostało poinformowane, że nasila się zjawisko przechodzenia jej klientów do firmy konkurencyjnej. Zarząd podjął decyzje o zbadaniu sprawy i ustaleniu przyczyn tego zjawiska. W tym celu rozpoczęto projekt eksploracji danych, którego zadanie brzmiało: podać charakterystykę (profil) klienta, który ma skłonność do zmiany firmy.

ID osoby	wiek	Średnia liczba rozmów zamiejscowych /tydzień	Zmiana operatora
1	23	62	Tak
2	40	47	Nie
3	21	20	Nie
4	56	43	Nie
5	45	50	Nie
6	34	51	Tak
7	22	66	Tak
8	19	53	Tak
9	28	68	Tak
10	30	60	Nie
11	58	76	Nie
12	50	69	Nie
13	48	35	Nie

Rozpoczniemy od wyboru grupy klientów firmy. Dla uproszczenia rozważymy skromny zbiór złożony z 13 osób. Musimy także zdecydować, które atrybuty charakteryzujące klientów zostaną wzięte pod uwagę w prowadzonej analizie. Odnotujmy przy tym, że decyzja ta jest już w jakimś stopniu naszą hipotezą o przyczynach przechodzenia do innych operatorów. W przykładzie weźmiemy pod uwagę następujące atrybuty: identyfikator klienta (ID), wiek, średnią liczbę rozmów zamiejscowych na tydzień, które przeprowadził klient i atrybut zawierający informację o tym czy osoba nadal jest naszym klientem, czy przeszła do innego operatora. Załączona tabelka jest częścią pewnej hipotetycznej bazy danych i zawiera dane historyczne o analizowanych osobach. Wydzielenie danych z bazy w postaci tabelki kończy krok gromadzenia danych, które będą przedmiotem eksploracji. Zauważmy przy tym, że krok ten zawierał zapewne kilka pod-zadań, na przykład obliczenie średniej liczby rozmów zamiejscowych w tygodniu. Innymi zadaniami, które mogły mieć miejsce są wyeliminowanie szumu i nadmiarowości danych (w bazie hipotetycznej jest pole „data urodzenia", z którego wyeliminowano dzień i miesiąc i obliczono wiek osoby), konsolidacja danych itp.

Patrząc na tabelkę można zapytać czy odnalezienie powodu zmiany operatora jest możliwe natychmiast, bez prowadzenia żadnych operacji. Być może dla tych, którzy lubią reprezentację danych w postaci tabel jest to zadanie do wykonania, dla większości wszak łatwiejsza do analizy byłaby reprezentacja danych w dwu-wymiarowej przestrzeni (na płaszczyźnie). Załączony rysunek tak właśnie przedstawia dane z tabelki. Każdy punkt reprezentuje klienta. Dane zostały sklasyfikowane w dwóch zbiorach w zależności od wartości atrybutu „zmiana operatora". Kwadraty oznaczają tych, którzy zmienili operatora, trójkąty – tych, którzy pozostali. Oś pozioma pokazuje wiek osoby, zaś oś pionowa – średnią liczbę rozmów zamiejscowych w tygodniu.



W przykładzie techniką eksploracji jest klasyfikacja polegająca tu na znalezieniu funkcji, która pozwoli przypisać osobę do jednej z dwóch klas: „klient, który zamierza zmienić operatora" i

„klient, który raczej nie zmieni operatora". Poszukiwanie tej funkcji wykona program komputerowy. Może to być program uczący się na danych treningowych z tabeli.

Jako punkt wyjścia dla tego programu przyjmiemy pewien *model eksploracji danych*; będzie nim funkcja liniowa $f(x) = \alpha x + \beta$. Teraz jesteśmy już gotowi (a dokładniej program komputerowy jest gotowy) do rozpoczęcia iteracyjnego *wyznaczania wartości parametrów* modelu, tzn. współczynników α oraz β. Po zakończeniu tego procesu dokonujemy *oceny modelu* w ten sposób, że dla wyznaczonych parametrów sprawdzamy jak uzyskana konkretna funkcja liniowa spełnia przyjęte *kryteria eksploracji danych*. Jako kryteria możemy przyjąć dokładność klasyfikacji i zrozumiałość dla człowieka (można też przyjąć jeszcze inne kryteria). Podsumujmy: model eksploracji danych, wyznaczanie parametrów modelu, ocena wyników na podstawie kryteriów tworzą razem to, co nazywa się *algorytmem eksploracji danych*. Zauważmy, że jeśli przyjęty model nie jest zadawalający, to trzeba poszukać innego modelu – czynność ta również należy do algorytmu.

W wyniku pracy programu poszukującego współczynniki uzyskaliśmy następującą liniową funkcję decyzyjną:

$f(x) = 1{,}3\ x$

która została pokazana na załączonym rysunku. Od razu widzimy, że nie możemy za jej pomocą (ani za pomocą żadnej innej funkcji liniowej) całkowicie rozdzielić dwóch założonych klas. Innymi słowy dokładność klasyfikacji nie jest doskonała.

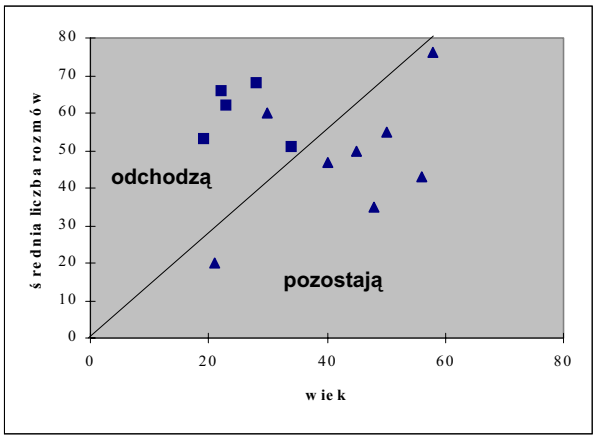
Ostatnim krokiem procesu eksploracji jest *interpretacja wyników*, co ma duże znaczenie gdyż mogą one mieć wpływ na decyzje dotyczące zarządzania firmą. W przykładzie okazało się, że większość młodszych klientów, którzy mają na swoim koncie dużą liczbę rozmów zamiejscowych skłonnych jest zmienić operatora (nie wszyscy jednak odeszli; pozostał na przykład klient ID = 10). A zatem wynik eksploracji można zawrzeć w następującym zdaniu: skłonność do zmiany operatora mają młodzi klienci, poniżej 35 lat, którzy mają na swoim koncie średnią lub więcej niż średnią liczbę rozmów zamiejscowych.

Nasuwa się tu natychmiast pytanie dlaczego liczba rozmów młodszych klientów jest czynnikiem krytycznym w decyzji o zmianie operatora ? Odpowiedzi należy szukać w następnej sesji eksploracji danych.

7. Zarys procesu eksploracji danych

Eksploracja danych, jak wspomnieliśmy, nie jest łatwym procesem. Poniżej podajemy sześć podstawowych kroków, które pozwolą uczynić ten proces skutecznym.

1. Zrozumieć i starannie zdefiniować problem/zadanie, który jest przedmiotem eksploracji. Ponadto, należy zanalizować i zrozumieć otoczenie, w którym ten problem występuje.
2. Wybrać zbiór danych, w których przeprowadzimy eksplorację. Zbiór ten musi być znaczącą próbką całego zasobu danych. Wybór dotyczy obiektów, ich atrybutów (zmiennych), przedziału czasu, zakresu geograficznego, wielkości próbki itd.
3. Zdecydować jak przygotować dane do przetwarzania. Na przykład: czy chleb i ciastka tortowe należą do grupy pieczywo? Czy wiek reprezentować jako przedział (np. 40-45 lat), czy jako liczbę (np. 40 lat).
4. Wybrać algorytm (lub ich kombinacje) eksploracji danych i wykonać program realizujący ten algorytm na przygotowanych danych. Odnotujmy, że często w



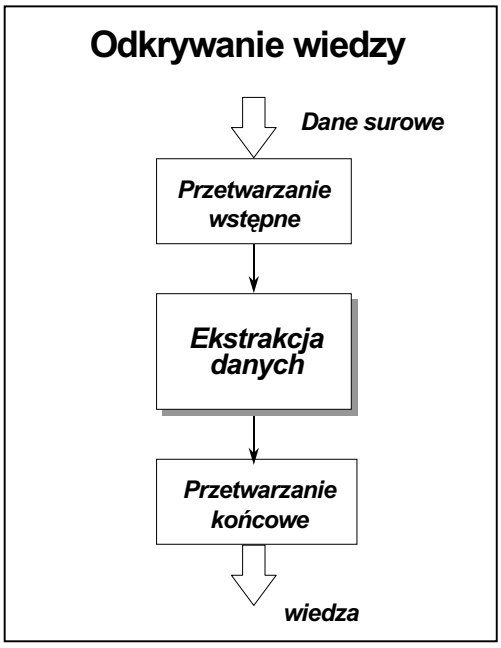
sposób iteracyjny musimy wrócić do kroku 3., a nawet 2. jeśli rezultaty nie są zadawalające.

5. Zanalizować wyniki wykonania programu i wybrać te, które stanowią rezultat pracy. W tym miejscu potrzebna jest ścisła współpraca analityka i specjalisty w dziedzinie, którą poddajemy badaniu. Wyniki należy przedstawić w formie przyjętej w organizacji, gdzie proces eksploracji jest prowadzony.
6. Przedłożyć wyniki kierownictwu organizacji i zasugerować sposób ich wykorzystania.

8. Odkrywanie wiedzy w bazach danych

W literaturze przedmiotu, zwłaszcza w pracach o charakterze teoretycznym, odróżnia się termin *eksploracja danych* od terminu *odkrywanie wiedzy*, a dokładniej *odkrywanie wiedzy w bazach danych* (*ang. knowledge doscovery in databases – KDD*). Zazwyczaj odkrywanie wiedzy odnosi się do całego procesu odkrywania przydatnych i pożytecznych informacji i wiedzy drogą eksplorowania baz danych, podczas gdy eksploracja danych ma węższe znaczenie, gdyż dotyczy tylko wyboru i zastosowania algorytmów i programów służących do wydobycia z baz reguł, zależności, schematów.

Odkrywanie wiedzy jest wielostopniowym procesem, który ma na celu uzyskanie nowych, wiarygodnych, potencjalnie pożytecznych i zrozumiałych dla człowieka informacji o prawidłowościach występujących w świecie reprezentowanym w bazie danych. W najogólniejszym zarysie proces ten składa się z trzech kroków (patrz rysunek), a mianowicie: (i) *przetwarzania wstępnego*, które obejmuje m.in. przygotowanie danych, wybór próbki danych, „czyszczenie" danych; (ii) *eksploracji danych*; (iii) *przetwarzania końcowego*, w ramach którego dokonuje się m.in. wieloaspektowej oceny, filtrowania, wariantowej wizualizacji i interpretacji uzyskanych wyników.



Trzeba mocno podkreślić, że w procesie odkrywania wiedzy niezwykle istotną rolę odgrywa człowiek, analityk problemu, którego umiejętności, doświadczenie i praca mają kluczowe znaczenie w otrzymaniu znaczących rezultatów. Jego rola polega na stałej krytycznej ocenie każdego kroku w procesie odkrywania, swoistym „cenzurowaniu" otrzymywanych rezultatów cząstkowych i sterowaniu całym procesem.

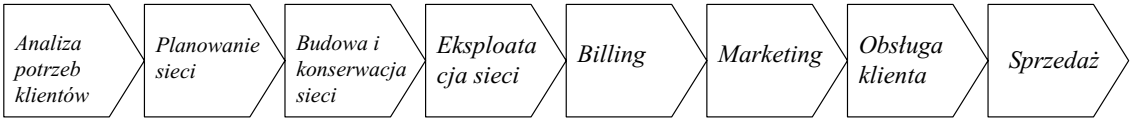
Historycznie rzecz ujmując termin „odkrywanie wiedzy w bazach danych" został utworzony w 1989 roku na określenie szeroko i ogólnie rozumianej koncepcji poszukiwania wiedzy zawartej w bazach danych. Pojęcie „eksploracja danych" natomiast zostało utworzone jako odnoszące się do technik i narzędzi używanych do wydobycia, analizy i prezentacji danych wydobytych z baz. Zdarza się wszak, zwłaszcza w dyskursie kolokwialnym, że oba terminy używane są wymiennie, jako synonimiczne. „Eksploracja danych" jest określeniem szczególnie chętnie używanym w środowiskach statystyków, analityków danych i grupach zajmujących się bazami danych i systemami informacyjnymi, podczas gdy termin „odkrywanie wiedzy" pojawia się przede wszystkim wśród badaczy pracujących w obszarze sztucznej inteligencji. Nie jesteśmy tu rygorystami językowymi, i tak długo jak nie prowadzi to do nieporozumień, akceptujemy wymienialność tych terminów. W literaturze anglosaskiej można natknąć się na spokrewnione określenia, takie jak: *knowledge extraction, data archaeology* lub *information harvesting*.

Na marginesie odnotujmy, że dotychczas najwięcej zastosowań technik odkrywania wiedzy miało miejsce w tzw. marketingu bazodanowym, który polega na analizie baz danych o klientach w celu ustalenia ich preferencji i wykorzystaniu otrzymanych rezultatów w akcjach marketingowych.

9. Eksploracja danych dla telekomunikacji

Firmy telekomunikacyjne generują, zbierają i przechowują każdego dnia ogromne ilości informacji, poczynając od danych o funkcjonowaniu sieci, przez dane bilingowe, aż po informacje na temat klientów. Rzadko jednak udaje się tym firmom w pełni skorzystać z zebranych danych, zwłaszcza że duża część wartościowych informacji jest na „pierwszy rzut oka" niewidoczna. Mając tego świadomość firmy telekomunikacyjne coraz chętniej sięgają po techniki eksploracji danych. Towarzyszy temu nadzieja, że dodatkowym efektem eksploracji będzie szansa na przeprowadzenie lepszej strukturyzacji i skonsolidowanie posiadanych zasobów, co jest jednym z warunków udanych prac nad pojawiającymi się coraz częściej hurtowniami danych.

Spoglądając na typowy łańcuch wartości firmy telekomunikacyjnej, który przedstawia się następująco:



dostrzegamy, że eksploracja danych może być przydatna w każdym ogniwie tego łańcucha, a w szczególności może:

w ramach zarządzania i eksploatacji sieci

- usprawnić zarządzanie procesami biznesowymi firmy,
- usprawnić zarządzanie funkcjonowaniem sieci i wykorzystanie infrastruktury technicznej,
- ułatwić planowanie budowy, rozwoju i konserwacji sieci,
- usprawnić zarządzanie alarmami generowanymi przez sieć i ułatwić obsługę usterek /awarii sieci,
- lepiej alokować posiadane zasoby,
- umożliwić nawiązanie dialogu i wymianę doświadczeń z współpracującymi firmami telekomunikacyjnymi w zakresie stosowania eksploracji danych.

w dziedzinie zarządzania kosztami

- zwiększyć współczynnik „lojalności" klientów,
- zmniejszyć nadużycia ze strony klientów,
- uczynić procesy finansowe przejrzystszymi i prostszymi, usprawnić księgowość i zarządzanie kredytami.

w zakresie marketingu i obsługi klienta

- lepiej rozpoznawać i zaspokajać potrzeby klientów,
- opracowywać i analizować możliwości nowych usług i produktów,
- zwiększyć efektywność sprzedaży i obsługi dotychczasowych klientów,
- znajdować nowe możliwości rozwoju firmy.

Oto kilka konkretnych zagadnień, zdefiniowanych za pomocą pytań stawianych przez operatorów telekomunikacyjnych, gdzie eksploracja danych może okazać się przydatna:

(a) W jaki sposób planować i optymalizować inwestycje na budowę i rozwój sieci utrzymując wysoki poziom usług ale bez nadmiernej rozbudowy infrastruktury ?
(b) Jaka jest struktura i regularności ruch w sieci ?
(c) Jak optymalizować topografię sieci ?

(d) Jak minimalizować koszty i nakłady czasowe na pomiary ruchu i parametrów eksploatacyjnych sieci ?
(e) Jak rozpoznawać i klasyfikować alarmy generowane przez sieć ?
(f) Jak rozpoznawać i klasyfikować problemy techniczne (anomalie, awarie), także problemy chronicznie powtarzające się, oraz ujawniać przyczyny anomalii ?
(g) Czy istnieją regularności i powtarzające się schematy dotyczące inicjowania połączeń w sieci?
(h) Jakie są wzorce zachowań użytkowników i jak rozpoznawać połączenia stanowiące nadużycie w stosunku do operatora sieci ?
(i) Jaki jest profil użytkownika i motywacja, które mogą skłonić go do zmiany operatora sieci ?
(j) Jaki jest profil użytkowników, którzy płacą wysokie rachunki ?
(k) Jakiej reakcja użytkowników można się spodziewać na wprowadzenie nowych rodzajów usług czy taryf, uwzględniając różnorodność profili użytkowników ?

W połowie 1999 roku Polska Telefonia Cyfrowa (ERA GSM) rozpoczęła projekt „Data Mining", którego celem było rozszerzenie stosowanych w tej firmie metod analizy danych przez wprowadzenie technik eksploracji danych, zwłaszcza w odniesieniu do zagadnień planowania, budowy i eksploatacji sieci, a więc zagadnień natury technicznej. Projekt ten realizowany jest z udziałem zespołu Instytutu Informatyki Politechniki Warszawskiej. Oto przykłady kilku zadań, które przeanalizowano za pomocą metod eksploracji danych.

Zadanie	**Zastosowane Metody**	**Efekty**
Wyszukiwanie anomalii działania sieci na podstawie logów routerów w sieci korporacyjnej	reguły asocjacyjne, grupowanie	Zbiór reguł (które potwierdziły wiedzę ekspertów)
Przewidywanie ruchu w sieci komórkowej	Grupowanie, drzewa decyzyjne, regresja	Model predykcyjny ruchu w sieci z akceptowalnym przez ekspertów błędem
Przewidywanie anomalii w działaniu sieci komórkowej; analiza w pojedynczych komórkach	reguły asocjacyjne, drzewa decyzyjne, wizualizacje	Zbiór reguł :95% reguł znanych ekspertom – oczywistych, 4% potwierdzających ich intuicje, 1% interesujących
Przewidywanie anomalii w działaniu sieci komórkowej, z uwzględnieniem wpływu komórek sąsiednich	reguły asocjacyjne, drzewa decyzyjne, wizualizacje	Zbiór reguł :90% reguł znanych ekspertom – oczywistych, 7% potwierdzających ich intuicje, 3% interesujących
Wykrywanie sekwencji czasowych alarmów w sieci komórkowej	reguły asocjacyjne, własne metody badania sekwencji czasowych	Eksperyment w toku

Do najważniejszych wniosków ogólniejszej natury, które wyciągnięto z dotychczasowych prac należą:

– zasadniczym warunkiem powodzenia eksperymentów jest udział specjalistów zlecających zadania, zwłaszcza w fazie definiowania zadania i ewaluacji wyników cząstkowych,

– przetwarzanie wstępne i końcowe danych stanowią około 85 % czasu przeznaczonego na rozwiązywanie zadania,

- to samo zadanie warto rozwiązywać stosując rożne metody eksploracji danych (wyniki mogą być zaskakująco różne),
- jeśli wybrano już metodę rozwiązania zadania, to należy zabiegać o możliwość prowadzenia eksperymentów na rożnych zbiorach danych dotyczących tego zadania,
- komercyjne oprogramowanie do prowadzenia eksperymentów eksploracji danych nie zawsze jest skuteczne do rozwiązywania zadań stawianych przez operatorów telekomunikacyjnych; dotyczy to zwłaszcza analizy zadań gdzie występują bardziej złożone struktury danych oraz zależności temporalne (sekwencje zdarzeń),
- transfer wiedzy w zakresie eksploracji danych dla telekomunikacji praktycznie nie istnieje; operatorzy bowiem nie są zainteresowani udostępnianiem swoich doświadczeń, gdyż traktują wiedzę pozyskaną za pomocą eksploracji danych jako element swej przewagi nad konkurentami.

10. Podziękowania

Autor składa podziękowania wszystkim kolegom z zespołu eksploracji danych, który działa w Instytucie Informatyki Politechniki Warszawskiej za współpracę w zakresie metod eksploracji oraz za informacje i ocenę komercyjnego oprogramowania do prowadzenie eksploracji danych. Podziękowania kieruję także do p. Tomasza Gerszberga, Dyrektora Departamentu Analiz i Budżetu Środków Trwałych w Polskiej Telefonii Cyfrowej (PTC), który zainicjował projekt „Data Mining" oraz do p. Roberta Parzydło, Kierownika projektu „Data Mining" w PTC, za stworzenie efektywnej platformy współpracy nad problemami eksploracji danych dla telekomunikacji oraz umożliwienie przeprowadzenia szeregu eksperymentów na danych rzeczywistych i wszechstronnego przedyskutowania uzyskanych wyników ze specjalistami PTC.

Literatura

W tym rozdziale podajemy kilka pozycji, które mogą pomóc Czytelnikowi w poszerzeniu informacji o eksploracji danych i odkrywaniu wiedzy w bazach danych, także w telekomunikacji.

[1] Berry, M. J. A., Linoff G., Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, 1997.

[2] Cox K. C., Eick S.G/, Wills G. J., Brachman R. J.: Visual Data Mining: Recognizing Telephone Calling Fraud, Data Mining and Knowledge Discovery, vol. 1, issue 2, 1997.

[3] Daszczuk W., Muraszkiewicz M. et al., Data Mining for Technical Operation of Telecommunications Companies: a Case Study, Proc. of Int. Conf. SCI/ISAS, USA, 2000.

[4] Data Mining Special Issue, Communications of the ACM, vol. 39, no 11, Nov. 1996.

[5] Dhar V., Stein R., Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall Computer Books, 1997.

[6] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.

[7] Mattison R.: Data Warehousing and Data Mining for Telecommunications, Artech House, 1997.

[8] Muraszkiewicz M., „ Data Mining at a Glance", Proc. of Int. Conf. TEMPUS PHARE JEP-12165-97,10-12 June, 1999, Gdansk.

[9] Weiss S., Predictive Data Mining: A Practical Guide, Morgan Kaufman Publishers, 1997.

[10] Sasisekharan R., Seshardi V.: Data Mining and Forecasting in Large-Scale Telecommunication Networks, IEEE Expert Intelligent Systems and their Applications, Feb. 1996.

Wybrane źródła w Internecie

[1] Data Warehousing Information Center, pwp.starnetinc.com/larryg/index.html

[2] Data Mining and Knowledge Discovery Resource Center
(także znany jako *Knowledge Discovery Mine*), www.kdnuggets.com

[3] DBMS Buyer's Guide, www.dbmsmag.com

[4] Knowledge Discovery Mine web site, info.gte.com/~kdd/index.html
Zawiera często zadawane pytania dotyczące eksploracji danych, odkrywania wiedzy i tematów pokrewnych

[5] Two Crows Corp., www.twocrows.com

[6] Two Crows opublikowało tu studium na temat narzędzi i użytkowników technik eksploracji danych