VIII Konferencja PLOUG
Kościelisko
Październik 2002

Eksploracja danych – dlaczego nie w przemyśle?

Tomasz Wyrozumski

Biuro Matematyki Stosowanej S.C.

e-mail: tw@bms.krakow.pl

Abstrakt

Metody eksploracji danych znalazły jak dotąd stosunkowo niewiele zastosowań w przemyśle, mimo iż potencjalne, choć nieuświadomione zapotrzebowanie na nie jest bardzo duże. Aparatura kontrolująca procesy produkcji dostarcza olbrzymich ilości danych pomiarowych, które na ogół wykorzystywane są jedynie do elementarnego sterowania i tworzenia bardzo prostych wykresów. Mogą one jednak być z powodzeniem użyte podczas złożonych badań, mających na celu poprawę jakości finalnego produktu, zmniejszenie ryzyka awarii itp. Analizujemy trzy rzeczywiste przypadki, w których zastosowanie data mining może przynieść konkretne korzyści. Przy okazji dość szczegółowo omawiamy analizę skupień przy użyciu tzw. metody k-uśrednień, jednej z zaimplementowanych w oprogramowaniu Oracle Darwin.

1. Wstęp

Gdy przegląda się literaturę poświęconą tematyce eksploracji danych (*data mining*), włączając w to materiały z konferencji, bądź też komercyjne teksty reklamujące poszczególne narzędzia programistyczne, można odnieść wrażenie, że zastosowania są tu praktycznie ograniczone do problematyki marketingu, a nieliczne wyjątki zdają się co najwyżej potwierdzać regułę. W istocie w ogóle mało które (a zwłaszcza polskie) przedsiębiorstwo naprawdę korzysta z gromadzonych przez siebie danych, poszukując w nich głęboko ukrytej wiedzy. Jeśli już jednak ktoś się na to zdecyduje, to na ogół chodzi o dane o sprzedaży i o wykorzystanie pozyskanej informacji w celach marketingowych. Dobre i to, lecz czy na prawdę cała korporacyjna wiedza sprowadza się do kwestii czysto handlowych? Czy np. w sferze produkcji przemysłowej nie pojawiają się żadne problemy, które dałoby się z powodzeniem rozwiązać przy pomocy metod data mining? Na te pytania nasuwa się tylko jedna odpowiedź: o ile w przedsiębiorstwach generalnie brak świadomości, jak bardzo przydatna może okazać się eksploracja, o tyle świadomości, jak przydatna może się ona okazać w przemyśle brak nawet jej propagatorom – firmom informatycznym świadczącym odpowiednie usługi.

Niniejszy artykuł wybiega więc w przyszłość nie o jeden, lecz o dwa kroki, zaś autor ma nadzieję że przynajmniej w jakimś stopniu nie jest to falstart. Optymizmem napawać może wyrywkowa analiza rynku prowadzona przez firmę BMS w okresie ostatnich kilkunastu miesięcy. Z rozmów naszych pracowników z kadrą zarządzającą szeregu polskich przedsiębiorstw produkcyjnych jasno wynika, że choć świadomość istnienia metod eksploracji danych oraz możliwości ich stosowania jest tu praktycznie żadna, to jednak potencjalne zapotrzebowanie na data mining wydaje się w tym sektorze ogromne. Odrębną kwestią jest duża ostrożność i konserwatyzm, jeśli chodzi o informatykę w ogóle, a o jej nowoczesne aspekty w szczególności. W jakimś stopniu wynikają one z faktu, że dużo łatwiej zdecydować się na eksperymenty w sferze marketingu, niż produkcji, gdzie liczy się konkret, a wszelkie nowe pomysły łatwo i nieuchronnie poddają się weryfikacji. Przy tych wszystkich niesprzyjających okolicznościach udało się nam jednak w kilku przypadkach stworzyć podwaliny pod przyszłą współpracę, a w szczególności doprowadzić do wyspecyfikowania zagadnień, w których użycie eksploracji danych mogłoby być pomocne.

Niniejszy artykuł prezentuje trzy takie rzeczywiste problemy, które w jakimś stopniu mogą być uznane za generyczne dla branży produkcji przemysłowej. Mając jednak na uwadze wspomniany już fakt, iż w środowisku inżynierów tematyka eksploracji danych jest mało znana, próbujemy najpierw przybliżyć nieco to zagadnienie, odwołując się do prostego, ale instruktywnego przykładu. Aby nie obracać się w sferze czystej abstrakcji, przedstawiamy w miarę szczegółowo jedną z metod data mining, pomocną w analizie skupień, tzw. metodę k-uśrednień. Omawiamy też pobieżnie inne metody, ich zastosowanie i charakterystyczne cechy, zaś kończymy dwiema uwagami natury ogólnej.

Artykuł nawiązuje wprawdzie do konkretnego oprogramowania (Oracle Darwin), jednak w istocie przedstawia zagadnienia, które są od oprogramowania niezależne. Powinno to usatysfakcjonować nie tylko tych czytelników, którzy korzystają lub zamierzają korzystać z systemów innych dostawców, lecz również osoby wierne Oracle'owi, a to dlatego, że firma konsekwentnie zmienia swoją ofertę w zakresie narzędzi eksploracyjnych, sukcesywnie włączając je do bazy danych 9i.

2. Co to jest eksploracja danych?

Zacznijmy więc ab ovo, czyli od krótkiego wyjaśnienia, czym jest eksploracja danych. Nie chcemy przy tym silić się na formułowanie jakiejś precyzyjnej definicji, bo w nie-formalnym języ-

ku i tak się to nie uda, ale raczej spróbujemy w kilku zdaniach przybliżyć istotę rzeczy osobom, które z eksploracją jeszcze się nie spotkały. Wyobraźmy więc sobie ogromną bazę danych, naszpikowaną informacją. Można z niej taką informację wydobyć zadając konkretne pytanie, sformułowane w języku SQL, można ją następnie przetworzyć przy użyciu mniej lub bardziej skomplikowanej matematyki. I tak na przykład, mając bazę wszystkich osób zamieszkujących nasze miasto możemy bez problemu dowiedzieć się, ilu urodzonym w tym roku chłopcom dano na imię Staś. Wyobraźmy sobie jednak inne pytanie: „Czy w różnych okresach czasu i w różnych grupach społecznych bywają modne różne imiona?" Jest ono ogólnie zrozumiałe dla ludzi (odpowiedź oczywiście brzmi „Tak"!), jednak nie da się go zadać wprost bazie danych. Trzeba je najpierw odpowiednio przetworzyć i to w bardzo nietrywialny sposób. Można więc „ręcznie" raportować imiona nadawane dzieciom w poszczególnych okresach czasu i grupach społecznych, porównywać owe raporty, zmieniać okresy i definicje grup, znów raportować i porównywać, aż wreszcie zauważy się pewne zależności. Można też zautomatyzować ten proces i to w dodatku na bardzo ogólnym poziomie, używając do tego celu zaawansowanych metod matematycznych. I to właśnie jest eksploracja danych – **poszukiwanie odpowiedzi na nieprecyzyjnie postawione pytania**, oczywiście nie „na piechotę", ale przy pomocy komputerów i specjalnie w tym celu opracowanych algorytmów.

Metod eksploracji jest wiele, począwszy od klasycznej statystyki aż po sieci neuronowe i wyrafinowane algorytmy genetyczne. Pewnie trudno byłoby też powiedzieć, co jest jeszcze bardzo zaawansowanym raportem, a co już zasługuje na miano data mining, zawsze bowiem chodzi o jedno – o ekstrakcję z bazy danych wiedzy, która nie jest bezpośrednio dostępna.

3. Informatyka w fabryce i źródła informacji

Systemy informatyczne w firmie produkcyjnej wspomagają niejako dwie platformy jej działalności. Jedna wiąże się z ogólną obsługą, a więc administracją, zarządzaniem, księgowością, zaopatrzeniem, zbytem itp., druga zaś bezpośrednio z procesami technologicznymi i to ona jest oczywiście istotna z punktu widzenia naszych obecnych rozważań. Przechodząc więc do sedna sprawy, zacząć musimy od tego, że każdy proces technologiczny wymaga sterowania, zaś samo sterowanie – stałego kontrolowania parametrów procesu. Służą do tego różne urządzenia pomiarowe, np. termometry, manometry, amperomierze itp. Niekiedy mamy do czynienia ze sterowaniem ręcznym, jeśli pracownik obserwując przyrządy kręci gałką w jedną bądź w drugą stronę, lecz zazwyczaj w nowoczesnych fabrykach sterownie odbywa się automatycznie, a nawet cały proces technologiczny kontrolowany jest przez komputery. Odpowiedni program symuluje wtedy przebieg procesu i na bieżąco porównuje ów proces wirtualny z rzeczywistym, dokonując niezbędnych korekt parametrów tego ostatniego. Jeżeli mamy do czynienia z taką sytuacją, zbieranie danych nie stanowi oczywiście żadnego problemu, jednak w rodzimych warunkach częściej spotyka się sterowanie półautomatyczne lub automatyczne, lecz nie scentralizowane. Mimo wszystko niektóre zakłady przemysłowe odczuwają jednak potrzebę gromadzenia danych pomiarowych i nabywają dedykowane do tego celu oprogramowanie. Przykładem dość często spotykanego rozwiązania jest Industrial SQL Server firmy Wonderware (poza nazwą nie ma nic wspólnego z serwerem Microsoft), który pozwala stworzyć centralne repozytorium danych pomiarowych, a także umożliwia ich wizualizację w postaci różnych wykresów. Analizując je technologowie znajdują odpowiedzi na pytania związane z przebiegiem procesu produkcyjnego – znajdują lub nie, bo niekiedy problemy są na tyle trudne i złożone, że zwykłe oglądanie wykresów ich nie rozwiąże. Oczywiście, w takich właśnie sytuacjach zalecalibyśmy stosowanie eksploracji danych.

Ważne jednak, aby w ogóle było co eksplorować. Jeżeli już fabryka zdecydowała się na wdrożenie oprogramowania takiego jak Industrial SQL Server, to znaczy, że jakimiś danymi dysponuje i co najwyżej trzeba będzie je poddać odpowiedniej obróbce wstępnej. Niestety, bardzo często danych po prostu nie ma, albo są tak rozproszone po różnego rodzaju rejestratorach, że ich zebranie

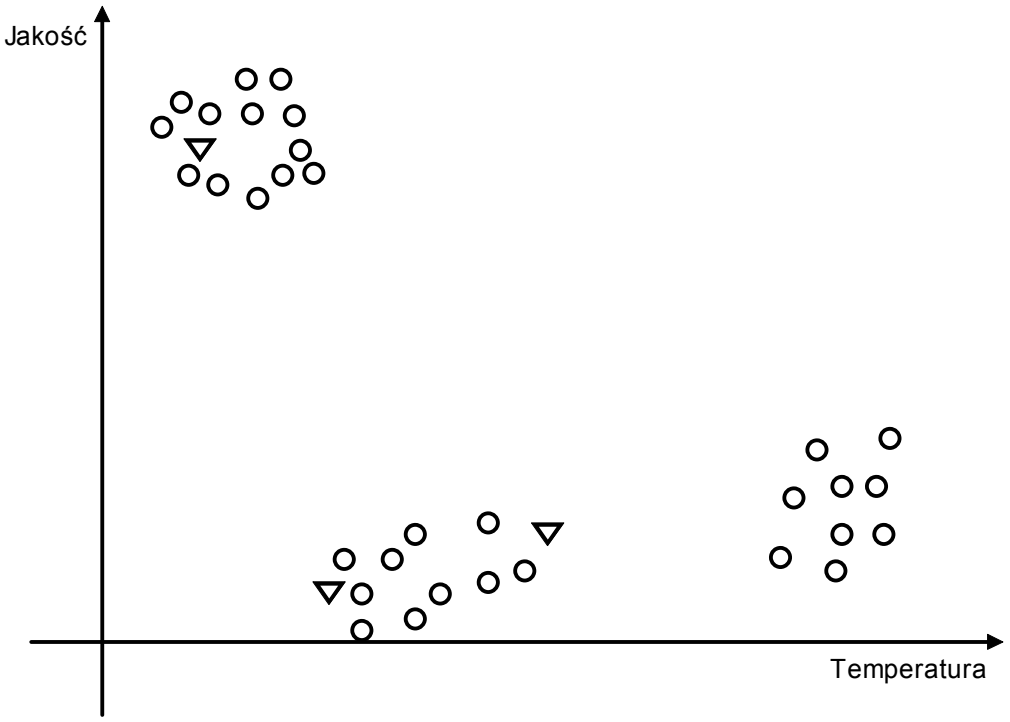
stanowi nie lada problem. Niekiedy zwykło się mówić – w kontekście systemów informatycznych związanych z obsługą biznesową – że najpierw buduje się hurtownie danych, a potem na ich bazie wdraża systemy eksploracji. Może nie do końca jest to prawdą, lecz w jakimś stopniu oddaje istotę rzeczy. W przypadku samej produkcji Industrial SQL Server (lub jego odpowiedniki) stanowi taką właśnie hurtownię, a jej posiadanie świadczy o pewnym stopniu świadomości informatycznej w przedsiębiorstwie i – niestety – w polskich warunkach wcale nie jest regułą. W każdym bądź razie, jeśli mamy już dane, możemy brać się do eksploracji.

4. Poprawa jakości produktu i analiza skupień

Jeden z klientów naszej firmy przedstawił nam do rozwiązania następujący problem. Wytwarzany produkt powstaje w wyniku skomplikowanego procesu technologicznego. Na linii produkcyjnej zainstalowano kilkaset czujników, które stale mierzą różne wielkości i przekazują dane do centralnego repozytorium. Zgodnie z opisem procesu, poszczególne parametry powinny być utrzymywane w odpowiednich reżimach. Okazuje się jednak, że nawet jeśli tak się dzieje, jakość produktu nie zawsze jest zadowalająca i nikt dokładnie nie wie od czego ona zależy. Technologowie „rozumieją" zaledwie kilka spośród kilkuset mierzonych parametrów, a to oznacza, że w większości przypadków nie mają intuicji, co do ich znaczenia i wpływu niewielkich zmian na jakość finalnego produktu. Widać więc wyraźnie, że ręczne strojenie raczej nie wchodzi tu w rachubę. Z pomocą może natomiast przyjść eksploracja danych. Istota rozwiązania polegałaby na odnalezieniu w danych historycznych korelacji pomiędzy jakością produktu, a mierzonymi wielkościami. Rzecz jasna, należy najpierw odrzucić zależności trywialne, tzn. sytuacje, w których jakość produktu była zła, ale ze znanych powodów. To niby oczywiste, jednak powinno uświadomić czytelnikom fakt, że narzędzia data mining nie są jakimiś uniwersalnymi maszynkami, do których wrzuca się wszystko „jak leci", a otrzymuje gotowy rezultat. W istocie dane zawsze trzeba odpowiednio wyselekcjonować, przygotować, przetworzyć itp.

Jak uchwycić korelacje? Jednym ze sposobów może być tzw. **analiza skupień** *(cluster detection)*, tzn. poszukiwanie zgrupowań w wielowymiarowej przestrzeni danych. W dwóch wymiarach, z których jeden obrazuje jakość mierzoną w pewnej skali, a drugi np. temperaturę w stopniach Celsjusza w pewnym miejscu linii produkcyjnej, sytuacja może wyglądać tak jak na Rys. 1.

Jak widać, dane występują w trzech zgrupowaniach, z których tylko jedno odpowiada zadowalającej jakości. Sens wykresu jest zatem oczywisty. Zaznaczmy jednak, że dostrzeżenie zgrupowań (klastrów) na dwuwymiarowym rysunku nie nastręcza żadnych problemów – po prostu widać je „gołym okiem" – jednak gdyby wymiarów było więcej niż 3, wizualizacja nie wchodziłaby już w rachubę. W istocie rekordy danych można zawsze postrzegać jako punkty w pewnych wielowymiarowych przestrzeniach atrybutów. I tak na przykład, jeśli atrybutami (kolumnami tabeli) są temperatura i jakość, mamy przestrzeń dwuwymiarową, lecz w ogólności atrybutów i, co za tym idzie, wymiarów może być znacznie więcej. Gdybyśmy mierzyli szereg temperatur i ciśnień na różnych etapach procesu produkcyjnego, to wykrycie prawidłowości takiej jak ta z Rys. 1 nie byłoby już wcale proste. Na szczęście z pomocą przychodzi matematyka i oczywiście maszyny cyfrowe. Istnieje szereg sposobów automatycznego wykrywania niejednorodności danych. W dalszej części artykułu przedstawimy bardzo prosty i skuteczny, zwany metodą k-uśrednień, który w 1967 roku zaproponował J.B. MacQueen. Sam przykład zaczerpnęliśmy z bardzo dobrze napisanej książki Berry'ego i Linoffa [1].



Rys. 1. Zależność jakości od temperatury

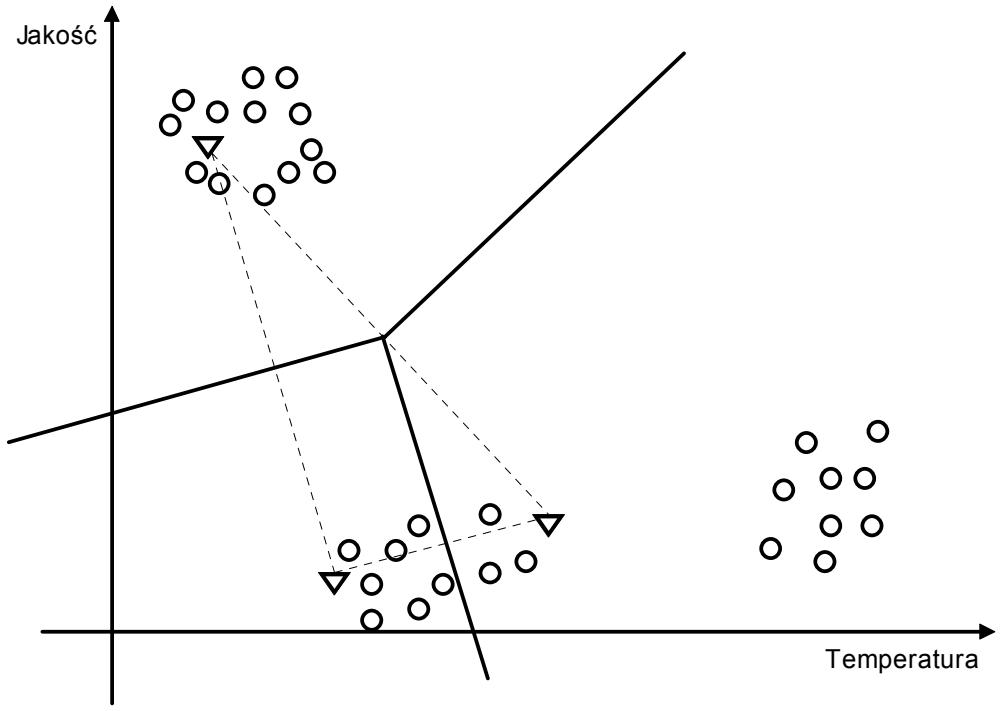
5. Metoda k-uśrednień

Punktem wyjścia jest nasza arbitralna decyzja, ile skupień powinno się znaleźć w zbiorze danych. Skomentujemy ją nieco później. Na razie jednak przyjrzyjmy się jeszcze raz sytuacji z Rys. 1 i załóżmy, że samą ilość ustaliliśmy prawidłowo. Wynosi ona oczywiście 3. Aby określić przynależność rekordów do skupień, wybieramy całkowicie losowo trzy punkty (małe trójkąty), które stanowić będą centra skupień (oczywiście w pierwszym przybliżeniu). Następnie obliczamy odległości pozostałych punktów od każdego z centrów i przyporządkowujemy je do tego centrum, do którego jest im najbliżej. W naszym konkretnym przypadku można to zrobić posługując się elementarną geometrią. Łączymy zatem centra odcinkami i wyznaczamy symetralne boków powstałego trójkąta, które, jak wiadomo, przetną się w jednym punkcie, a zawarte w nich półproste podzielą płaszczyznę na trzy części.

Tą metodą uzyskamy pierwsze przybliżenie naszych skupień. Jak widać gołym okiem, nie jest ono doskonałe, bowiem aż cztery punkty zostały przydzielone do niewłaściwego skupienia. Wykonujemy więc następną iterację, tym razem zaczynając już nie od losowo wybranych punktów, lecz od środków ciężkości poprzednio wyznaczonych skupień. Z elementarnej fizyki wiadomo, że dla N punktów materialnych o identycznych masach, leżących w jednej płaszczyźnie, współrzędne środka ciężkości dane są wyrażeniami:

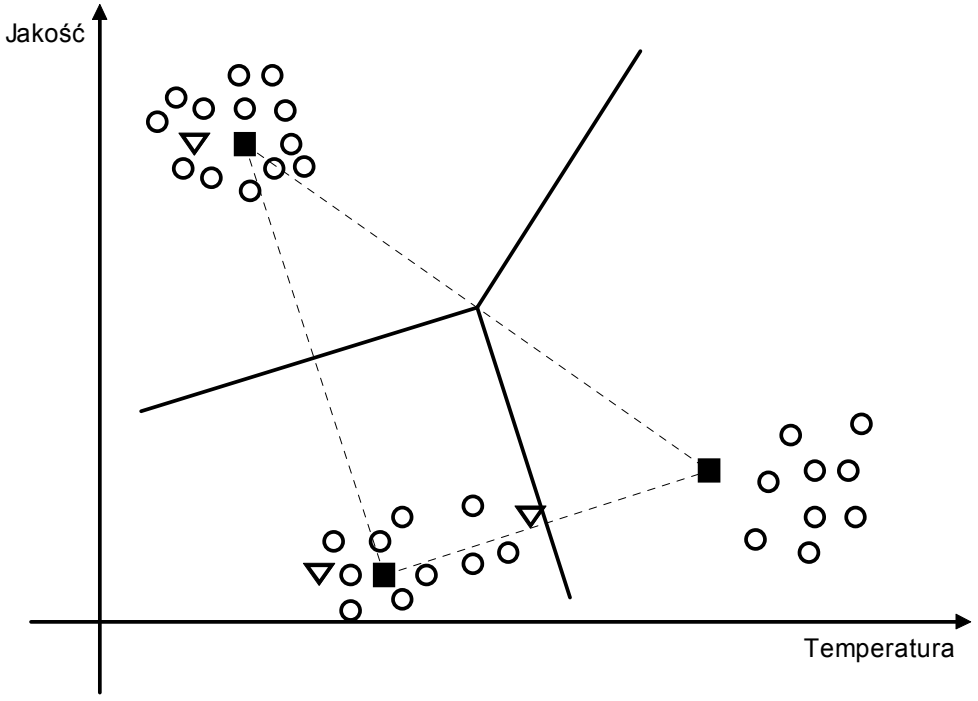
$$x = \frac{1}{N}\left(x_1 + x_2 + \ldots + x_N\right), \quad y = \frac{1}{N}\left(y_1 + y_2 + \ldots + y_N\right),$$

zatem cała operacja nie przedstawia trudności. Na rys. 3 pokazaliśmy sytuację po drugiej iteracji.



Rys. 2. Skupienia po pierwszej iteracji

Widać teraz, że półproste, do których konstrukcji użyto środków ciężkości pierwotnych skupień (czarne kwadraty) prawidłowo podzieliły nasz zbiór danych. Oczywiście, na samym początku, przy losowaniu punktów wyjściowych mogliśmy mieć mniej szczęścia. Wtedy prawidłowe zidentyfikowanie skupień wymagałoby po prostu większej ilości iteracji, ale z uwagi na zbieżność procedury, doprowadziłoby do tego samego rezultatu.



Rys. 3. Skupienia po drugiej iteracji

W tym miejscu chcielibyśmy skomentować dwa ważne aspekty opisanej metody. Zaznaczmy jednak najpierw, że choć doczekała się ona szeregu modyfikacji (między innymi opartych na koncepcji zbiorów rozmytych) zmierzających głównie do poprawy zbieżności ciągu iteracji oraz efektywności numerycznej, to jednak zasadnicza idea pozostała niezmieniona, a poniższe komentarze odnoszą się do wszystkich wariantów metody k-uśrenień.

I tak komentarz pierwszy dotyczy ilości skupień. Na pewno można powiedzieć tylko tyle: jest ich nie mniej niż 1, gdy cały zbiór danych potraktujemy jako jedno wielkie zgrupowanie, i nie więcej niż liczba elementów tego zbioru, kiedy to każdy element stanowi indywidualne skupienie. Przy każdej założonej, pośredniej wartości K znajdziemy jakieś skupienia. Mogą one być przy tym lepsze lub gorsze i, co bardzo istotne, odpowiadać różnym regułom grupowania. Np. w danych opisujących pacjentów szpitala można będzie znaleźć zarówno dwa sensowne skupienia (płeć), jak też i cztery (grupa krwi). Co zatem można zrobić? Z matematycznego punktu widzenia, można wprowadzić pewne miary jakości podziału zbioru rekordów. Można na przykład mierzyć odległości pomiędzy skupieniami (ich środkami ciężkości lub brzegami) i porównywać je z odległościami wewnątrz skupień. Można też zapożyczyć pewne koncepcje z fizyki, postrzegając skupienia, jako układy grawitujących punktów (galaktyki!) i badać energię całego systemu. Analizując takie miary dla różnych ilości odnalezionych klastrów można następnie wypowiadać się o jakości podziału. Warto jednak uzmysłowić sobie praktyczny sens poszukiwania skupień. W rzeczy samej chodzi bowiem o to, aby wynik był jasny dla nas, ludzi, a to z kolei oznacza, że liczba skupień nie może być za duża, bo po prostu przestaniemy rozumieć, co one oznaczają! O ile więc odnalezienie kilku klastrów może dać kapitalny wgląd w dane, kilkunastu – nieco gorszy, choć wciąż jeszcze akceptowalny, o tyle kilkadziesiąt lub kilkaset – najprawdopodobniej pozostawi nas w punkcie wyjścia, oczywiście jeśli chodzi o ekstrahowanie wiedzy z baz danych.

Osobna kwestia związana jest z mierzeniem odległości pomiędzy rekordami. W przypadku dowolnego zbioru, np. danych opisujących proces technologiczny, cała sprawa wydaje się wysoce nietrywialna. Ogólnie rzecz biorąc pomiar odległości nie zawsze jest możliwy. Przestrzenie, w których da się to zrobić, matematycy nazywają przestrzeniami metrycznymi. W rozważanym przez nas przypadku danych opisujących jakość i temperaturę łatwo jest określić odległość w każdym z wymiarów, np. odległość między temperaturami 30°C a 40°C. wynosi 10°C, zaś odległość między jakością 4 a 5 (w jakiejś skali) wynosi 1. Jak jednak określić (jedną liczbą!) odległość między punktem odpowiadającym temperaturze 30°C i jakości 4, a innym, opisującym temperaturę 40°C i jakość 5? Aby przestrzeń danych, którą się posługujemy, stała się przestrzenią metryczną, należy określić w niej funkcję g, zwaną metryką, czyli odległością. Ogólnie wiadomo, że musi ona przyporządkowywać parom punktów nieujemne liczby rzeczywiste oraz spełniać trzy dodatkowe warunki, które wypisaliśmy poniżej:

(i) $g(x, y) = 0 \Leftrightarrow x = y$,

(ii) $g(x, y) = g(y, x)$,

(iii) $g(x, y) + g(y, z) \geq g(x, z)$.

Trzeci z nich, zwany nierównością trójkąta, jest najtrudniejszy do spełnienia (zainteresowanych szczegółami odsyłamy do dowolnego podręcznika geometrii), jednak cały problem w tym, że metryk może być bardzo wiele i sama matematyka żadnej nie daje pierwszeństwa. Wybór jednej z nich jest już wynikiem wstępnej analizy konkretnego problemu i tym samym – elementem eksploracji danych. W naszym przykładzie, nawet jeśli zdecydujemy się na bardzo intuicyjną metrykę typu euklidesowego, to z uwagi na zgodność jednostek będzie ona musiała zawierać pewien dodatkowy parametr w. Jeżeli różnica temperatur dla rozważanych punktów wyniesie ΔT, a różnica jakości ΔJ, to całkowita odległość między nimi będzie dana wyrażeniem

$$\sqrt{w^2(\Delta T)^2 + (\Delta J)^2}\ .$$

Liczba *w* jest jakby wagą określającą znaczenie temperatury w odniesieniu do jakości, oczywiście tylko w sensie pomiaru odległości w naszej przestrzeni. W istocie metryka, nawet bardziej egzotyczna, zazwyczaj zawierać będzie takie wagi, przyporządkowywane atrybutom danych i tym samym informujące o ich znaczeniu. A priori wagi te mogą zmieniać się od punktu do punktu, jeśli np. w pewnym obszarze temperatura odgrywa bardziej istotną rolę niż w innym. Wtedy liczby *w* stają się funkcjami, a nasza przestrzeń danych – przestrzenią zakrzywioną, taką jak np. powierzchnia globusa, a nie płaska kartka papieru. Warto zaznaczyć, że matematycy bardzo dużo wiedzą o przestrzeniach metrycznych (również zakrzywionych) i potrafią dostarczyć nam metod, które pomogą wybrać najlepszy – z punktu widzenia analizy konkretnych danych – sposób mierzenia odległości.

Z powyższych rozważań wyraźnie wynika, że analiza skupień musi zostać poprzedzona skomplikowanymi przygotowaniami. Nota bene, dotyczy to również innych metod eksploracji danych. Z drugiej strony, użycie oprogramowania znajdującego skupienia, czy analizującego dane w inny sposób nie stanowi bynajmniej finału prac. I tu komentarz drugi: każdy, kto uczył się fizyki wie, że wynik pomiaru (a eksploracja to pomiar na danych) powinien zostać uzupełniony o oszacowanie dokładności. Przykładowo, stwierdzenie, że ciśnienie atmosferyczne na poziomie morza wynosi 760 ±1 mm Hg może być błędne (zwłaszcza gdy właśnie rozbudował się wyż), natomiast stwierdzenie, że wynosi ono 700 ±100 mm Hg – całkowicie poprawne. Tym samym, mówienie, że ciśnienie wynosi po prostu 760 mm Hg będzie w każdej sytuacji pozbawione jakiejkolwiek wartości. Podobnie jest z eksploracją danych. Jeśli chcemy podać wynik, musimy oszacować jego dokładność, posługując się w tym celu klasycznymi metodami rachunku błędów.

6. Inne problemy spotykane w przemyśle

Innym problemem, z jakim spotkaliśmy się podczas analizowania zapotrzebowania na usługi eksploracji danych w sektorze przemysłowym, jest badanie przyczyn awarii. Pokazano nam linię produkcyjną, na której co jakiś czas dochodzi do eksplozji. Wprawdzie zdarza się to dość rzadko, a wybuchy nie są potężne i nie wyrządzają nikomu krzywdy, jednak powodują uszkodzenia maszyn, przerwy w produkcji i – co za tym idzie – koszty. Mimo pewnej powtarzalności zjawiska oraz wysiłków inżynierów, nie udało się ustalić jego pierwotnej przyczyny (bezpośrednia jest oczywiście znana). Fabryka wiąże tu spore nadzieje z eksploracją danych, licząc że dzięki analizie korelacji różnych mierzonych parametrów z faktem eksplozji, a także badaniu odpowiednich zależności czasowych, uda się ustalić, co tak na prawdę jest źródłem problemu.

Trzecie zagadnienie, z jakim spotkaliśmy się w praktyce, wydaje się nieco podobne, lecz chodzi w nim o coś trochę innego. Otóż, maszyny pracujące w obrębie linii produkcyjnych bywają szalenie skomplikowane i poza tym, że ulegają mniej lub bardziej poważnym awariom, wymagają zwykłej konserwacji, polegającej nierzadko na wymianie różnych części. Niektóre wymienia się okresowo lub po pewnym przebiegu maszyny i wtedy oczywiście łatwo jest przewidzieć zapotrzebowanie na owe podzespoły. Nierzadko jednak wymiana uzależniona jest od zużycia konkretnej części, a to już znacznie trudniej przewidzieć. Dział serwisu zauważa np. powiązanie między zużyciem niektórych elementów, a porą roku – i to mimo, iż cała linia produkcyjna mieści się wewnątrz klimatyzowanej hali! W tym przypadku przyczyną są prawdopodobnie chwilowe awarie zasilania, powstające w następstwie wyładowań atmosferycznych, które z kolei z różną intensywnością występują o różnych porach roku. Jak widać z tego prostego i bynajmniej nie akademickiego przykładu, złożoność problemu jest duża. W dodatku trudno go ogarnąć, jeśli w grę wchodzi kilka tysięcy części zamiennych, a właśnie tak się w rzeczywistości dzieje. Dział serwisu, którego obowiązkiem jest zapewnienie ciągłości produkcji, jest żywotnie zainteresowany w utrzymywaniu możliwie wy-

sokich zapasów w magazynie części, zwłaszcza, że ich zakupy wiążą się z importem, a to wydłuża termin dostawy nawet do kilku tygodni. Z drugiej strony, dyrektor finansowy przedsiębiorstwa z oczywistych powodów chciałby, aby zapasy były jak najmniejsze. Sposobem rozwiązania owego konfliktu interesów jest dobre prognozowanie zapotrzebowania na części zamienne, czyli po prostu – ich zużycia. I znów można to robić na bazie danych historycznych, analizując korelacje pomiędzy zużyciem poszczególnych elementów, a czasem, względnie innymi zmiennymi niezależnymi.

7. Alternatywne i komplementarne metody eksploracji

Analiza skupień, której świadomie poświęciliśmy wiele uwagi, jest tylko jedną z szeregu znanych technik eksploracji danych. Wydaje się, że może ona znaleźć dużo zastosowań w zagadnieniach, które pojawiają się w przemyśle. Jednak o metodzie tej warto pamiętać również z innego powodu – szczególnie dobrze nadaje się do wstępnych analiz. Krótko mówiąc, jeśli bardzo niewiele wiemy o naszej bazie danych i jest nam trudno choćby trochę sprecyzować pytanie, które chcielibyśmy jej zadać, zawsze możemy pokusić się o poszukanie skupień. Jeśli je znajdziemy – a jakieś znajdziemy niemal zawsze – to wynik może okazać się bardzo interesujący i znaczący. Osobom, które zetknęły się z astronomią, zwracamy w tym miejscu uwagę, że słynny diagram Herzsprunga-Russela, dzielący gwiazdy na ciąg główny, czerwone olbrzymy i białe karły, jest w istocie przykładem zastosowania analizy skupień. Oczywiście, nie zawsze prowadzi ona do naukowych odkryć, ani też do wyników przekładających się natychmiast na konkretne praktyczne korzyści, jednak na ogół pozwala dowiedzieć się czegoś ciekawego o danych, czegoś, co może przydać się w dalszej analizie, do której możemy użyć innych algorytmów. Najważniejsze z nich, sieci neuronowe i drzewa decyzyjne, dostępne nota bene w narzędziu Oracle Darwin, zostały przez nas dość szczegółowo omówione w artykułach zamieszczonych w materiałach z poprzednich konferencji PLOUG [4, 5], wobec czego nie będziemy obecnie wyjaśniać czytelnikowi zasad ich działania, a jedynie odeślemy go do literatury (również [1, 2, 3]), przypominając tu tylko najważniejsze cechy charakterystyczne obydwu wspomnianych metod i porównując je z analizą skupień.

Tak więc (sztuczne) sieci neuronowe są w istocie algorytmami przetwarzania danych, bazującymi na konstruowaniu i rozpoznawaniu wzorców. Wzorce nie są budowane w żadnym języku zrozumiałym dla ludzi, a zatem sieci nie ekstrahują wiedzy. Nadają się natomiast do klasyfikowania danych, a stanowią niemal idealne narzędzie do prognozowania sekwencji czasowych. Można wyobrazić sobie wiele zastosowań sieci neuronowych w przemyśle, jak choćby wspomniane już przewidywanie zużycia części zamiennych, gdzie zamiast poszukiwania (czasem bardzo skomplikowanych) reguł rządzących zjawiskiem, da się alternatywnie zastosować podejście czysto pragmatyczne i bez rozumienia istoty rzeczy, po prostu spróbować stworzyć pewne ilościowe prognozy.

W przeciwieństwie do sieci drzewa decyzyjne dostarczają zrozumiałej wiedzy o danych. W istocie są one pewną formą analizy skupień, choć stosowane algorytmy (np. CART w Oracle Darwin) mają charakter całkowicie odmienny od opisanej metody k-uśrednień. W efekcie stosowania drzew uzyskujemy zdania analityczne w języku SQL (a więc niemal naturalnym) orzekające o danych. Podobne zdania uzyskuje się również de facto stosując metodę k-uśrednień, bo podział zbioru danych na skupienia oznacza przecież umiejętność wypunktowania cech tych ostatnich, czyli właśnie ich opis w języku SQL. Czym więc różnią się oba podejścia poza samą matematyką metody podziału? Różnica naprawdę polega na sposobie parametryzacji. Jak wiadomo, algorytmy drzew decyzyjnych przez samą swoją naturę prowadzą do całkowitego rozdrobnienia danych, produkcji olbrzymiej ilości ich klas i – co za tym idzie – przypadkowych reguł. Aby tego uniknąć przycina się drzewa, stosując odpowiednie techniki, które jednak mają swoje wady, i swobodne parametry, a przez to nie gwarantują całkowitej jednoznaczności tego podejścia. W przypadku analizy skupień metodą k-uśrednień mamy inne swobodne parametry: ilość skupień, współczynniki

metryki. Jeżeli są one bardziej zrozumiałe i naturalne niż te pojawiające się implicite w algorytmach drzew – powinniśmy zastosować metodę k-uśrednień, jeśli nie – może warto pomyśleć o drzewach. Widać zatem, że o wyborze podejścia decydują tu względy bardziej techniczne niż fundamentalne.

8. Uwagi końcowe

Na zakończenie chcielibyśmy zamieścić dwie uwagi. Pierwsza z nich dotyczy zaproponowanej na wstępie „definicji" eksploracji danych, jako poszukiwania w bazach odpowiedzi na nieprecyzyjnie zadane pytanie. Mamy nadzieję, że po lekturze niniejszego artykułu czytelnik będzie miał jasność, jak należy ową nieprecyzyjność rozumieć i nie nabierze (całkowicie mylnego) przekonania, że eksploracja danych realizowana jest przy pomocy jakiejś cudownej maszynki, do której wrzuca się gigabajty i w przybliżeniu mówi, o co chodzi, a w efekcie – otrzymuje gotową odpowiedź, przekładającą się w dodatku natychmiast na konkretne biznesowe korzyści. Naprawdę eksploracja wymaga wiele pracy wysoko kwalifikowanego i doświadczonego konsultanta. Trzeba bowiem przygotować odpowiednio dane (ważenie, skalowanie itp.) wybrać właściwy model, sparametryzować go, a po uzyskaniu wyników – oszacować błędy, czasami zanalizować wrażliwość i – co bardzo ważne – zrozumieć wynik. Oczywiście, jeśli wszystkie te prace zostały wykonane, a następnie chcemy powtórzyć obliczenia na innych danych, o podobnym charakterze, całą procedurę można już zautomatyzować i tym samym oddać klientowi gotową aplikację, którą będzie mógł samodzielnie używać. Jakby jednak na to nie spojrzeć, eksploracja danych nie jest czymś, za co odpowiedzialnie może brać się firma handlowa, sprzedająca gotowe oprogramowanie, ani nawet typowy integrator, nie posiadający odpowiedniego przygotowania matematycznego ani doświadczenia.

Druga uwaga, stanowi nawiązanie do wstępu, niestety – pesymistyczne nawiązanie. Nie chcielibyśmy bowiem również, aby czytelnik nabrał przekonania, że polski przemysł masowo wdraża rozwiązania eksploracji danych, nie ustępując przy tym ani na krok korporacjom zachodnim. Lepsza smutna prawda, lecz prawda: w kraju nad Wisłą wciąż największym powodzeniem cieszą się młotek, gwoździe, sznurek do snopowiązałek... Zakłady przemysłowe otwarte na nowatorskie technologie informatyczne stanowią nieliczne choć chlubne wyjątki, zaś ogólnie komputery postrzega się jako trochę lepsze maszyny do pisania, czy trochę lepsze kalkulatory. Oprogramowanie to przede wszystkim: FK, kadry i płace, sprzedaż i magazyn, może czasem coś jeszcze. Głośne wdrożenia ERP nierzadko sprowadzają się do uruchomienia potężnych systemów, które później realizują najprostsze funkcje, a ich rzeczywisty potencjał wykorzystany jest w niewielkim stopniu. W informatyzacji przodują oczywiście firmy zagraniczne, ale te bardzo często przywożą gotowe rozwiązania „w teczce" i w zasadzie nie chcą nawet rozmawiać z rodzimymi dostawcami. Nawet nie dlatego, że nie traktują ich poważnie, ale dlatego że „taka jest polityka korporacji". Oczywiście, patrząc na to z drugiej strony, rynki zachodnie stoją z kolei przed nami otworem, a eksport myśli technicznej, choć trudny, może stanowić szansę. Może też później ta wyeksportowana myśl wróci do nas z powrotem, a odpowiednio opakowana i odpowiednio droższa zostanie przyjęta jak objawienie. Cóż, podobno nikt nie jest prorokiem we własnym kraju...

Bibliografia

1. Berry M.J.A., Linoff G.: Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley Computer Publishing 1997, ISBN 0-471-17980-9.
2. Mulawka J.J.: Systemy ekspertowe, Wydawnictwa Naukowo-Techniczne 1996, ISBN 83-204-1890-9.
3. Tadeusiewicz R.: Sieci neuronowe, Akademicka Oficyna Wydawnicza RM 1993, ISBN 83-85769-03-X.
4. Wyrozumski T.: Prognozowanie neuronowe w oparciu o dane ekonomiczne z baz Oracle, VI Konferencja PLOUG, Zakopane 2000, ss. 297-304.
5. Wyrozumski T.: Zastosowanie drzew decyzyjnych w systemach wspomagających pracę działów IT, VII Konferencja PLOUG, Zakopane 2001, ss. 141-150.