Zastosowanie drzew decyzyjnych w systemach wspomagających pracę działów IT

Tomasz Wyrozumski
Biuro Matematyki Stosowanej S.C.
e-mail: tw@bms.krakow.pl

Abstrakt. Działy IT dużych korporacji muszą kontrolować pokaźne, nierzadko rozległe systemy informatyczne, a także służyć pomocą ich użytkownikom. W tym celu wykorzystują one różne narzędzia wspomagające zarządzanie pracą serwerów, stacji roboczych, aktywnych urządzeń sieciowych itp. Systemy zarządzające zbierają dane o funkcjonowaniu wspomnianych komponentów i przechowują je w odpowiednich bazach. Pokazujemy, że przy pomocy algorytmów drzew decyzyjnych dane te mogą być z powodzeniem analizowane w celu usprawnienia pracy działów IT, a tym samym również całych korporacji. Koncentrujemy się przy tym głównie na problemach interesujących kadrę kierowniczą. Do eksploracji danych wykorzystujemy oprogramowanie Oracle Darwin.

1. Wstęp

W dzisiejszych czasach trudno już sobie wyobrazić dużą korporację lub nawet przedsiębiorstwo średniej wielkości funkcjonujące bez komputerów. Z jednej strony niewyobrażalnie ułatwiają nam one pracę, z drugiej jednak - jesteśmy równie niewyobrażalnie od nich uzależnieni. Warto choć na chwilę uzmysłowić sobie, że drobny defekt półprzewodnika, brak styku we wtyczce, czy wadliwe łożysko wentylatora może spowodować poważne perturbacje w funkcjonowaniu światowego giganta. A co powiedzieć o niedopatrzeniu projektanta systemu lub wynikłym ze zwykłego zmęczenia błędzie programisty, którego nie wychwycono podczas testów? Oczywiście, w zależności od stopnia ryzyka podejmuje się różne środki zabezpieczające, aż po zwielokrotnianie systemów informatycznych, jak to ma miejsce np. w projektach kosmicznych. Wróćmy jednak na ziemię i zajmijmy się tymi zastosowaniami informatyki, które nie mają bezpośredniego związku ze zdrowiem i życiem ludzkim. Tu awarie były są i będą, a im bardziej dana organizacja jest zinformatyzowana, tym bardziej awarie dezorganizują jej pracę i w efekcie - tym więcej kosztują.

Oczywiście, zadaniem kierownictwa jest minimalizowanie kosztów działalności, nie zaś redukowanie prawdopodobieństwa awarii za wszelką cenę do jakiegoś założonego minimum. Niektóre elementy systemu można więc i trzeba zwielokrotnić (w praktyce - najczęściej zdublować), jednak trudno sobie wyobrazić, by np. każdy pracownik miał na swoim biurku drugi komputer - ot tak, na wszelki wypadek. W praktyce działy IT tworzą komórki zajmujące się pomocą użytkownikom (tzw. Helpdesk), których pracownicy przybywają na wezwanie, usuwają uszkodzenie i przywracają stan pierwotny. Byłoby przy tym dziwne, gdyby informatyzacja nie dotknęła również informatyków. Ich praca nie jest przecież tania, a oni sami, jak mało kto wiedzą, w czym człowiekowi mogą pomóc maszyny. Nie obawiają się przy tym specjalnie utraty swoich miejsc pracy, bo doświadczenia pokazują, że komputery generalnie nie odbierają zatrudnienia takim jak oni.

W efekcie więc działy IT - i to nie tylko sami szefowie - dostrzegają potrzebę wdrażania systemów wspomagających ich pracę, tzn. szeroko rozumiane zarządzanie strukturą informatyczną (lub nawet teleinformatyczną) przedsiębiorstw.

Firma BMS wykonywała ostatnio analizę zapotrzebowania na taki właśnie zintegrowany system zarządzający dla jednej z polskich spółek. Nie wdając się w szczegóły możemy stwierdzić, że chodziło o przedsiębiorstwo bardzo duże (choć raczej pod względem kapitalizacji niż ilości sprzętu informatycznego). W jego posiadaniu znalazło się kilkaset komputerów osobistych, ok. dziesięciu serwerów (Sun Solaris oraz Intel Windows NT) i odpowiednia liczba różnych drukarek, aktywnych urządzeń sieciowych itp. Dalsza część tych rozważań opiera się między innymi na doświadczeniach zebranych podczas realizacji wspomnianego projektu, jak również innych podobnych,

prowadzonych wcześniej, zaś celem artykułu jest zwrócenie uwagi osobom odpowiedzialnym za funkcjonowanie działów informatyki na nowe możliwości, jakie oferują im techniki eksploracji danych, ze szczególnym zwróceniem uwagi na algorytmy drzew decyzyjnych. Zostały one, nota bene, zaimplementowane w nowym narzędziu firmy Oracle - oprogramowaniu o nazwie Darwin, oryginalnie stworzonym przez firmę Thinking Machines Corporation.

2. Potrzeby informatyków

Zainteresowania pracowników działu IT w odniesieniu do systemów wspomagających zarządzanie infrastrukturą koncentrują się zasadniczo wokół trzech grup zagadnień:

- inwentaryzacji sprzętu, oprogramowania, licencji oraz umów serwisowych,
- monitorowania pracy urządzeń, wykrywania awarii i zagrożeń, analizowania ich skutków oraz rejestrowania związanych z tym kontaktów z użytkownikami,
- tworzenia różnego rodzaju raportów dotyczących funkcjonowania infrastruktury informatycznej.

Cele te są w jakimś stopniu realizowane przez różne „duże" systemy dostępne na rynku, takie jak np. HP OpenView, Tivoli NetView, czy CA Unicenter TNG, a także przez rozwiązania wycinkowe, jak Microsoft SMS. Aby posiąść niezbędne informacje, na poszczególnych komputerach, bądź aktywnych urządzeniach sieciowych systemy te instalują agentów, którzy na bieżąco kontrolują konfigurację maszyn oraz ich pracę. Zasadniczo możliwe są dwa schematy komunikacji z centralą nadzorującą. W pierwszym, typowym dla zagadnień inwentaryzacyjnych, agenci gromadzą informacje w lokalnych bazach danych, które synchronicznie lub na żądanie są odpytywane przez oprogramowanie centrali. Spotyka się też jednak inne rozwiązanie, polegające na tym, że to sami agenci wysyłają komunikaty, zazwyczaj posługując się w tym celu protokołem SNMP (Simple Network Management Protocol), a centrala nasłuchuje i analizuje odebrane informacje. Ten model jest idealny w przypadku monitorowania pracy urządzeń, gdyż nie powoduje nadmiernego obciążenia sieci. Korzystając z niego można wykryć nie tylko rzeczywiste awarie, lecz również potencjalne zagrożenia. Na przykład, jeżeli kontrolowana jest temperatura procesora, to jej wzrost ponad ustaloną wartość spowoduje wygenerowanie odpowiedniego powiadomienia. W efekcie wskazani pracownicy IT mają szansę dowiedzieć się o zagrożeniu jeszcze zanim dojdzie do takiego przegrzania układu, że komputer przestanie pracować.

Wszystkie informacje spływające z poszczególnych elementów systemu do centrali, niezależnie od sposobu ich przekazywania, mogą być wykorzystywane zarówno w bieżących działaniach operacyjnych (np. wymiana wadliwego wentylatorka na procesorze), jak i w analizach o charakterze strategicznym. Te ostatnie, stanowiące oczywiście domenę kierownictwa, bazują na różnego rodzaju raportach tworzonych na podstawie zebranych informacji. Jak wiadomo, kierownicy na ogół lubią być o wszystkim dobrze poinformowani i to w możliwie syntetyczny sposób. Można zresztą śmiało powiedzieć, że im wyższy szczebel zarządzania, tym bardziej syntetyczna powinna być informacja, a na szczeblu najwyższym oczekuje się zwykle głównie pewnych wskaźników o charakterze finansowym. Jeżeli raportowania nie traktujemy jedynie jak sztuki dla sztuki, albo generowania dokumentów, które obowiązkowo muszą być przekazywane określonym osobom, bądź instytucjom, to musimy stanąć nie tylko przed technicznym problemem tworzenia raportów, ale też przed problemem ich czytania. Dla przykładu, szef IT, który zażyczy sobie zestawienia wszystkich nieudanych prób logowania w systemie w ostatnim miesiącu, otrzyma wielostronicowy wydruk i w najlepszym razie spróbuje przerzucić pracę na któregoś ze swych podwładnych mówiąc: „Sprawdź, co z tego wynika."

Oczywiście, w miarę łatwo jest znaleźć, jeśli wiadomo, czego się szuka. Można wtedy tworzyć skomplikowane, przekrojowe raporty w oparciu o bazę relacyjną lub wielowymiarową. Jeśli jednak nie mamy jasności co do przedmiotu poszukiwań, a jedynie przeświadczenie, że dane zawierają jakieś cenne informacje, wtedy znacznie lepiej odwołać się do technik eksploracyjnych (data mining).

3. Drzewa czy sieci?

Jak wiadomo, istnieje wiele metod eksploracji danych i nie jest obecnie naszym zamiarem szczegółowe ich omawianie. Chcielibyśmy natomiast zwrócić uwagę, że jedna z nich wyróżnia się w sposób wyjątkowy. Chodzi o tzw. drzewa decyzyjne, które w przeciwieństwie do metod detekcji klastrów, czy sieci neuronowych, pozwalają generować zdania analityczne opisujące dane. Jest to niezwykle cenna właściwość z punktu widzenia istot ludzkich, gdyż dla ludzi umiejętność formułowania tego rodzaju wypowiedzi o rzeczywistości jest koniecznym (choć nie wystarczającym) warunkiem dla stwierdzenia, że się tę rzeczywistość „rozumie". Cokolwiek by to miało znaczyć i niezależnie, ile poziomów takiego rozumienia można wyróżnić, nie podlega dyskusji, że wszelkie wypowiedzi naukowe są formułowane właśnie w postaci takich zdań.

Sztuczne sieci neuronowe działają, jak wiadomo, zupełnie inaczej. Z punktu widzenia człowieka stanowią one czarne skrzynki, produkujące np. całkiem trafne prognozy rzeczywistości, jednakże w sobie tylko wiadomy sposób. Mogą one być niekiedy bardzo użyteczne w praktyce, zwłaszcza dla tych, którzy rozumieją ich ograniczenia. Mogą też dostarczać cennych wskazówek ułatwiających dalszą analizę danych, jednak bez wątpienia nie produkują bezpośrednio wiedzy. Aby to unaocznić, rozważmy następujący, nieco akademicki przykład.

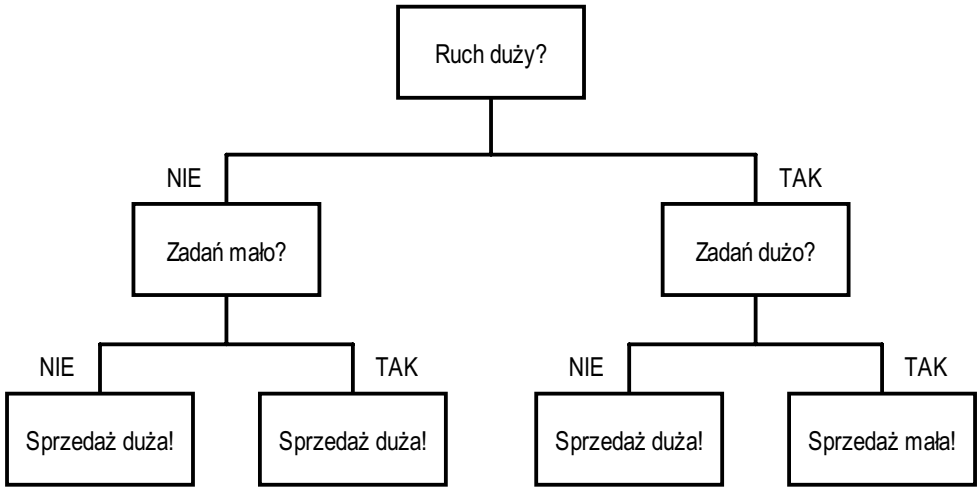
W pewnym rozległym systemie, wspomagającym sprzedaż, co jakiś czas pojawiają się zatory. Wskutek tego spadają obroty i w efekcie - wynik finansowy, a to bardzo niepokoi udziałowców. Użytkownicy skarżą się, ale - jak łatwo zgadnąć - ich skargi mają dość niekonkretny charakter. Trudno zresztą wymagać, aby zajęci obsługą podenerwowanych klientów, na bieżąco telefonowali do działu IT i relacjonowali mu sytuację. Chcąc więc zdiagnozować problem możemy posłużyć się jedynie raportami ze sprzedaży i badać, w jakim stopniu są one skorelowane z różnymi zdarzeniami w systemie, bowiem zanim zaproponujemy zarządowi zakup klastra serwerów za niebagatelną sumę, powinniśmy się zastanowić, czy nie wystarczy zwykłe dostrojenie baz danych, nie mówiąc już o zadbaniu, by niekorzystne zdarzenia nie zbiegały się ze sobą w czasie. Kto wie zresztą, czy klaster w ogóle coś da, jeżeli przyczyną jest, powiedzmy, wąskie gardło w sieci?

Do analizy sytuacji możemy oczywiście użyć sieci neuronowych, które np. dostarczą nam dość szybko prognoz na przyszłość. O ile tylko różne niekorzystne zdarzenia w systemie mają charakter regularny, prognozy te będą raczej trafne. Generalnie nie chodzi jednak o to, aby przewidzieć, kiedy będą kłopoty, ale o to, by zrozumieć ich przyczyny i podjąć odpowiednie środki zaradcze. Do tego właśnie doskonale nadają się drzewa decyzyjne. W najprostszym podejściu możemy spróbować obliczyć różne średnie dzienne poszczególnych dostępnych charakterystyk, a więc sprzedaży, obciążeń procesorów, urządzeń sieciowych, baz danych, ilości zapytań, zadań drukowania na poszczególnych drukarkach itp. oraz wprowadzić dwie kategorie „dużo" (znacznie powyżej średniej) i „mało" (znacznie poniżej średniej). Słowo „znacznie" wymaga tu oczywiście pewnego dookreślenia, np. przy pomocy odchylenia standardowego. Drzewa decyzyjne mogą teraz wyprodukować coś bardzo cennego, tzn. zdania w języku naturalnym, takie jak choćby: „Jeżeli ruch w pewnej lokalnej podsieci duży i na określonej drukarce zadań dużo, to sprzedaż na stanowiskach odległych mała". Oczywiście nie jest to wynik podany „na talerzu", ale też znacznie więcej niż np. gołe liczby. Mamy przed sobą otwartą drogę do rozwiązania: otóż w podsieci znajduje się drukarka, która jest wykorzystywana do drukowania miesięcznych raportów, poważnie obciążających bazę danych. To właśnie spowalnia pracę na stanowiskach operacyjnych. Oczywiście, korelacje mogą mieć bardziej złożony charakter. Oprócz raportów miesięcznych tworzy się przecież kwartalne i roczne, co jakiś czas aktualizuje się lokalne cenniki, wysyła klientom spore partie listów elektronicznych itp. A priori nie wiadomo, co z czym i jak jest powiązane, więc jeśli nie zechcemy użyć drzew decyzyjnych, będziemy zmuszeni poruszać się na oślep, i to w wielowymiarowej przestrzeni!

4. Jak działają drzewa decyzyjne?

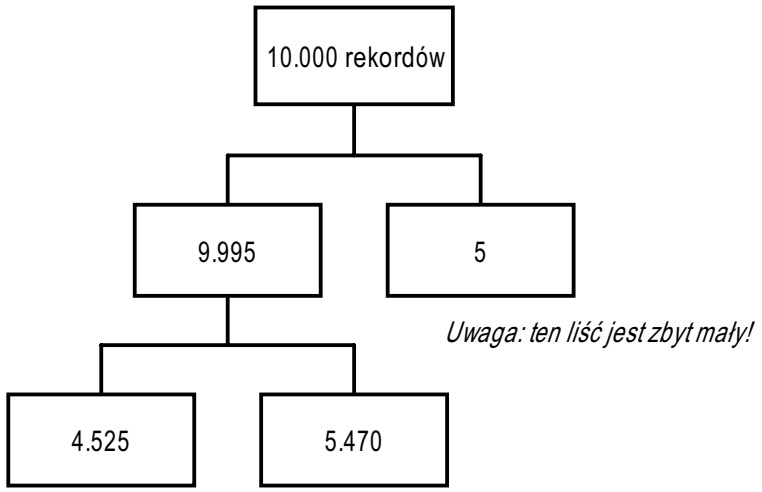
Zdania takie, jak przedstawione w poprzednim ustępie („Jeżeli ruch w pewnej lokalnej podsieci

duży i na określonej drukarce zadań dużo, to sprzedaż na stanowiskach odległych mała.") w istocie klasyfikują dane. Widać to doskonale na א, na którym wspomniane zdanie zostało przedstawione graficznie w postaci drzewa.



rys. 1

Mając zestaw tego typu zdań możemy bez problemu utworzyć drzewo, a następnie podzielić dane na określone podzbiory, umieszczone w liściach (drzewo rośnie z góry na dół!). Ciekawszy jest jednak problem odwrotny, tzn. mając do dyspozycji dane, zbudować klasyfikujące je drzewo. Aby to osiągnąć, należy dysponować jakimś sensownym algorytmem, który potrafi dzielić dane. Oczywiście nie chodzi o to, aby algorytm „rozumiał" głęboki sens podziału, gdyż jest to całkowicie niewykonalne. Algorytm może jedynie zadbać o systematyczne malenie liczebności klas, czyli mówiąc obrazowo o to, by drzewo rozrastało się harmonijnie, by pień dzielił się na coraz mniejsze gałęzie i aby nie wyrastały wprost z niego drobniutkie listki (ב).



rys. 2

Aby zrealizować ten cel w praktyce, definiuje się pewną wielkość *R*, zwaną indeksem różnorodności danych (*index of diversity*). Duża wartość indeksu oznacza jednorodny rozkład danych względem poszczególnych cech, natomiast indeks maleje, jeżeli któraś cecha jest wyraźnie liczniej reprezentowana niż pozostałe. Podział przeprowadza się teraz w taki sposób, aby możliwie jak najbardziej obniżyć różnorodność, a mówiąc ściślej, zmaksymalizować funkcję

$$R_0 - \left(P_1 R_1 + P_2 R_2\right),$$

gdzie R_0 oznacza indeks różnorodności przed podziałem, R_1 i R_2 - indeksy obu podzbiorów powstałych wskutek podziału, a P_1 i P_2 - odpowiednio prawdopodobieństwa, że wybrany rekord znajdzie się w określonym podzbiorze. W omawianym przez nas przypadku rekordy danych opisujące wartość sprzedaży najbardziej różnicuje atrybut wielkości towarzyszącego jej ruchu w „pewnej lokalnej podsieci". Jeśli więc rozdzielimy dane względem niego, to suma indeksów poszczególnych podzbiorów, ważona odpowiednimi prawdopodobieństwami będzie najmniejsza.

Oczywiście nie ma jakiegoś jednoznacznego przepisu na wyliczenie indeksu różnorodności. Jeżeli wyobrazimy sobie, że pewien atrybut danych może przyjmować dwie wartości, a P_+ i P_- są prawdopodobieństwami odpowiadającymi tym wartościom, to dobrym kandydatem na nasz indeks będzie każda funkcja przyjmująca maksimum, gdy prawdopodobieństwa te są równe, a malejąca, gdy któreś z nich zaczyna dominować. Może to więc być chociażby

$$\min\left(P_+, P_-\right).$$

Oprogramowanie Oracle Darwin umożliwia nam wybór między formułą

$$2P_+P_-,$$

zwaną funkcją Giniego, a miarą entropową

$$-\left(P_+ \log P_+ + P_- \log P_-\right)$$

.

Nie jest w tym momencie naszym zamiarem szczegółowe objaśnianie sensu powyższych wyrażeń, więc ograniczymy się jedynie do stwierdzenia, że każde posiada pewną statystyczną interpretację i że nie można a priori powiedzieć, które z nich jest lepsze - w praktyce wszystko zależy od specyfiki rozwiązywanego problemu.

Jak już zauważyliśmy, w celu dokonania podziału należy przeprowadzić odpowiednie obliczenia dla każdego atrybutu, a następnie wybrać ten, który zapewni największy spadek indeksu i względem niego rozdzielić dane. Cały proces powtarza się tak długo, aż obniżenie indeksu przestaje być możliwe. Oznacza to, że odpowiednia gałąź drzewa nie może już dalej rosnąć, wobec czego uzyskany zbiór rekordów określa się mianem liścia. Oczywiście, droga od pnia do liścia daje się w prosty sposób opisać zdaniem w języku SQL, a więc przedstawiony algorytm w istocie generuje takie właśnie zdania.

Niestety, trudno oczekiwać, że otrzymana klasyfikacja okaże się stuprocentowo trafna. W rzeczywistości w liściach znajduje się zwykle pewna domieszka niepożądanych rekordów, posiadających inną niż oczekiwana cechę. Oznacza to, że na końcu odkrytej reguły powinniśmy dodać sformułowanie: „z prawdopodobieństwem M / N", gdzie *N* jest liczbą wszystkich rekordów w liściu, a *M* - liczbą rekordów pożądanych. $(N - M) / N$ można zatem interpretować jako błąd dokonanej klasyfikacji. Dla przykładu, w liściu opisanym wyrażeniem „...ruch w pewnej lokalnej podsieci duży i na określonej drukarce zadań dużo..." znajdzie się *M* przypadków odpowiadających małej sprzedaży na stanowiskach odległych, przy czym jeśli *M* jest znacznie większe od 0,5, to klasyfikację można uznać za dobrą.

Suma błędów poszczególnych liści ważona prawdopodobieństwem, że rekord znajdzie się w tym liściu stanowi rozsądną miarę błędu całego drzewa. Tak więc np. liście obarczone dużym błędem, ale zawierające mało rekordów mają pomimo to stosunkowo mały wkład do błędu drzewa.

Pewna trudność, polega na tym, że drzewa zazwyczaj zbytnio się rozrastają. W efekcie dochodzimy do bardzo drobnych listków (czyli zbiorów zawierających niewiele rekordów), a odkryte reguły mają dość przypadkowy charakter. Dla przykładu, w wybranej próbie danych mogło się zdarzyć, że małej sprzedaży na stanowiskach odległych towarzyszyły awarie monitora przy serwerze. Ta bezsensowna korelacja nie była oczywiście specjalnie silna i w ogóle nie zostałaby wykryta, gdyby konstrukcję drzewa przerwać na pewnym etapie. Jeżeli postrzegać drzewa jako

struktury samoorganizujące się poprzez proces nabywania wiedzy, to można tu mówić o swoistym przetrenowaniu, efekcie analogicznym jak w przypadku sieci neuronowych. Podobne jest również rozwiązanie tego problemu - należy zbadać działanie drzewa na zbiorze danych testowych, różnym od wykorzystanego do jego konstrukcji, a następnie przyciąć nieco gałęzie. Wtedy wyeliminujemy różne, całkowicie przypadkowe „reguły". Rzecz jasna, trudno z góry powiedzieć, które gałęzie i w jakim stopniu przycinać, lecz tu właśnie można posłużyć się miarą błędu drzewa. Tworzymy zatem rodzinę poddrzew, powstałych poprzez przycięcie wyjściowego obiektu, a następnie wybieramy to, dla którego błąd jest najmniejszy.

Widać doskonale, że budowanie drzewa nie jest procedurą jednoznaczną, a to z uwagi na pewną dowolność w wyborze indeksu różnorodności, tudzież sposobu przycinania gałęzi. W przeciwieństwie do sieci neuronowych, nie ma tu jednak żadnego elementu losowości i całe postępowanie ma w pełni deterministyczny charakter. Dodajmy jeszcze, że budowanie drzew decyzyjnych jest procesem żmudnym i czasochłonnym, jako że za każdym razem przed podjęciem decyzji o rozdziale danych należy przebadać wszystkie możliwości i wybrać najkorzystniejszą, a wraz z rozrostem drzewa rośnie też ilość miejsc, w których dokonuje się rozdziału. Aby zapewnić odpowiednią moc obliczeniową, wspomniane oprogramowanie Oracle Darwin zostało wykonane w technologii klient-serwer, przy czym o ile klient, zawierający przyjazne graficzne narzędzia analiz danych, może być uruchamiany na komputerach klasy PC, pracujących pod kontrolą dowolnego systemu WIN 32, o tyle wykonujący obliczenia serwer wymaga maszyny SUN lub HP i - co najważniejsze - potrafi rozdzielać zadania pomiędzy dostępne procesory. Nota bene, cecha ta jest również bardzo ważna dla innych zaimplementowanych w Darwinie metod eksploracji, tzn. sieci neuronowych i tzw. algorytmów MBR, służących do detekcji klastrów danych.

5. Obszary zastosowań

Podczas wspomnianych prac analitycznych prowadzonych ostatnio przez firmę BMS, udało się zidentyfikować cztery główne grupy zagadnień interesujących kierownictwo działu informatyki (ta klasyfikacja odnosi się de facto do ostatniego punktu poprzedniej). Były to:

a) kwestie dotyczące inwentaryzacji mienia,

b) zagadnienia wydajności i wykorzystania infrastruktury teleinformatycznej,

c) szeroko rozumiane problemy bezpieczeństwa,

d) różne aspekty funkcjonowania samego działu IT.

O ile pierwsza z wymienionych grup jest nieciekawa z naszego punktu widzenia, o tyle w obrębie drugiej bez wątpienia można już mówić o zastosowaniu metod eksploracji danych, np. drzew decyzyjnych. Omawiany w poprzednich ustępach przykład zaliczał się w zasadzie właśnie do tej grupy i choć rzeczywiste przypadki są na ogół znacznie bardziej złożone, to jednak generalnie mają podobny charakter.

Znaczenia trzeciej grupy po prostu nie sposób przecenić. Oczywiście, podstawowe minimum zabezpieczeń systemów komputerowych przed włamaniami stanowią elementy pasywne, tzn. chroniony hasłami dostęp do danych oraz usług, serwery proxy, ściany ogniowe („firewall") itp., jednak w przypadku dużych i narażonych na rozmaite ataki firm jest to stanowczo za mało. Systemy muszą bronić się aktywnie, a więc na bieżąco monitorować i diagnozować sytuację, a także reagować alarmami oraz odłączać poszczególne elementy w wypadku dostrzeżenia prób ataku. Warto zaznaczyć, że istnieją na rynku różne gotowe rozwiązania, jak np. ISS, które w czasie rzeczywistym badają ruch w sieci pod kątem ewentualnych włamań. Ponieważ jednak każdy system informatyczny i każda instytucja mają swoją specyfikę, nie od rzeczy wydaje się wnikliwa analiza raportów zawierających informacje np. o aktywności zainstalowanych modemów, czy zdublowanych kart sieciowych w powiązaniu choćby z nietypowymi próbami dostępu do baz danych, tudzież innymi, z pozoru niewinnie wyglądającymi zdarzeniami. Doświadczenie uczy, że jeśli tylko ktoś wymyśli jakiś zamek, zawsze znajdzie się ktoś inny, kto nauczy się go otwierać bez

klucza. Nigdy nie przewidzi się wszystkiego, więc warto zaprząc drzewa decyzyjne do przeglądania wspomnianych raportów, bo właśnie one mają szansę odnaleźć to, co nigdy nie przyszłoby nam do głowy, a więc wyśledzić działania potencjalnego włamywacza, i to jeszcze zanim mu się powiedzie.

Ostatnia, nie mniej ważna grupa zagadnień dotyczy funkcjonowania samego działu IT, a więc jakby wewnętrznej kontroli. Chodzi o informacje o reklamacjach użytkowników i reakcjach na nie, o awaryjności systemów i usuwaniu awarii, a także o usługach serwisowych świadczonych przez firmy zewnętrzne. W czym może tu pomóc eksploracja danych? Kierownictwo działu IT wyznacza sobie zazwyczaj trzy strategiczne cele. Dwa pierwsze wynikają z ogólnych założeń funkcjonowania każdej korporacji. Są to więc: poprawa jakości świadczonych usług i obniżenie ich kosztów. Trzeci, niemniej istotny, choć pachnący partykularyzmem, polega na uzasadnieniu własnego istnienia i potrzeb rozwoju. Ponieważ informatycy na ogół nie zarabiają bezpośrednio pieniędzy dla swojej instytucji, bywają niekiedy postrzegani wyłącznie jako jednostka generująca koszty, co ma oczywiście wiadomy efekt, gdy zarząd zaczyna szukać oszczędności. Stąd też w bardzo wielu przypadkach kierownictwo IT musi ciągle udowadniać, jak potrzebna jest praca jego ludzi. Udaje się to w mniejszym lub większym stopniu, a autorowi znane są przykłady firm, w których pod presją informatyków wprowadzono nawet wewnętrzne fakturowanie usług IT świadczonych na rzecz innych działów, powiązane oczywiście z odpowiednimi rozliczeniami pieniężnymi. Zazwyczaj przyjmuje się jednak mniej radykalne rozwiązania, polegające na formalnym obciążaniu określonych centrów kosztowych, z których budżetu finansuje się odpowiednie projekty informatyczne, czy nawet bardziej przyziemne działania, nie zasługujące na nazwę projektu.

Widać więc wyraźnie, że kierownictwo IT zmuszone jest przeprowadzać różne analizy o charakterze ekonomicznym i odpowiadać na pytania w rodzaju: „Co należy zrobić, aby zmniejszyć koszty, poprawić efektywność, udowodnić zasadność wydatków?”. Tu właśnie drzewa decyzyjne mogą okazać się bardzo pomocne, pokazując np., jakie elementy systemów informatycznych są najdroższe w ogólnym rozrachunku, tzn. po uwzględnieniu ceny początkowej, kosztów awarii, serwisu itp., a jakie tylko wydają się drogie. Mogą też uwidocznić, jak różne koszty rozkładają się w czasie, jak obciążają poszczególne centra i kiedy powstają. Dysponując wypowiedziami w rodzaju: „Jeżeli cena zakupu niska, awaryjność duża, serwis tani, to komputer firmy A.”, „Jeżeli cena zakupu wysoka, awaryjność mała, serwis drogi, to komputer firmy B.” itp., możemy łatwiej podejmować decyzje dotyczące nowych zakupów. Niekiedy dla jednego działu lepiej będzie wybrać sprzęt B, dla innego zaś - A. Oczywiście, każdy chciałby mieć rozwiązania z najwyższej półki, lecz na ogół fundusze bywają ograniczone, a jeśli już chce się przekonać do czegoś zarząd, to powinno się mówić nie o bitach czy bajtach, lecz raczej o złotych lub o dolarach.

6. Podsumowanie

W ramach krótkiego podsumowania chcielibyśmy zamieścić dwie uwagi. Pierwsza z nich ma charakter bardziej techniczny i dotyczy samych drzew decyzyjnych, a konkretnie tego, że klasyfikacja uzyskana przy ich użyciu rzadko bywa idealna, tzn. drzewo obarczone jest na ogół pewnym, niezerowym błędem. Można postrzegać to zarówno jako wadę, jak i zaletę klasyfikacji. Potocznie mówimy bowiem: „Od każdej reguły są wyjątki.” i to zdanie bardzo dobrze oddaje nasz, ludzki sposób myślenia. W istocie mniej interesują nas klasyfikacje idealne, gdyż na ogół mają one dość trywialny charakter i nawet w naukach ścisłych, gdzie spotyka się je najczęściej, funkcjonują jedynie w obrębie pewnych modeli matematycznych o ograniczonym zakresie stosowania. Znacznie ciekawsze natomiast są klasyfikacje przybliżone. Z takimi mamy do czynienia choćby w ekonomii, ale również i we wspomnianych naukach ścisłych, czy technice, jeśli chodzi np. o ocenę użyteczności modeli. Sam fakt istnienia wyjątków od reguł nie jest dla nas ludzi czymś zaskakującym, ani też niezadowalającym. Przyzwyczailiśmy się do ułomności świata w którym żyjemy (albo też do ułomności własnego umysłu) i nauczyliśmy się z tą ułomnością żyć. Drzewa decyzyjne są więc takie same jak i my, przemawiają do nas naszym własnym, nie do końca precyzyjnym, lecz jakże nośnym informacyjnie językiem.

Uwaga druga, a zarazem swoista konkluzja tych rozważań, dotyczy użyteczności eksploracji

danych w ogóle. Otóż obecnie eksploracja jest jeszcze ciągle kosztownym luksusem. Prezesi pytają: „Czy to potrzebne?”, mówią: „Przecież można się bez tego obejść...”, albo „Może kiedyś wdrożymy takie rozwiązania, lecz na razie mamy ważniejsze problemy...”. Oczywiście, wszystko można robić ręcznie. Przecież - przypomnijmy to raz jeszcze - nie tak dawno w ogóle nie było komputerów, a istniał przemysł, telekomunikacja, banki. Prawdopodobnie za jakiś czas eksploracja danych stanie się absolutnym standardem i ludziom trudno będzie uwierzyć, że ongiś brało się do ręki pełne liczb raporty i przeglądało je, rozważając, co też z nich wynika.

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.