

# Data mining

## – inteligencja biznesowa

Adam WALANUS,  
Tomasz DEMSKI

Dysponując dziś często obszernymi i wszechstronnymi danymi o procesie produkcyjnym i mając przy tym świadomość wielkiej wartości tych danych, ociągamy się z zastosowaniem właściwych w takiej sytuacji analitycznych metod statystycznych. Przyczyna jest prosta: obawa przed trudnymi zagadnieniami statystyki matematycznej i skomplikowaniem nowoczesnych, zaawansowanych metod analiz. Tymczasem jest do dyspozycji *data mining* – technologia praktycznej realizacji projektów wydobywania wiedzy z danych i stosowania ich *on line* w procesie produkcji.

**W**spółczesne przedsiębiorstwa gromadzą coraz więcej danych o procesach produkcyjnych, dostawcach, klientach i ich wymaganiach, o awaryjności produktów i opiniach klientów. W danych tych ukryta jest cenna i potencjalnie użyteczna wiedza, która powinna posłużyć do podejmowania trafnych decyzji, zarówno na etapie projektowania produktu, w procesie wytwórczym, jak i w marketingu.

Tradycyjne, elementarne metody analizy danych, bynajmniej nie specjalnie łatwe w zastosowaniu, okazują się dziś często niewystarczające. A to dlatego, że:

- ▶ zmuszeni jesteśmy interesować się bardzo wieloma cechami produktu i procesu produkcyjnego, liczącymi nawet w tysiącach,
- ▶ mamy bardzo wiele pokrewnych, aczkolwiek różnych produktów,
- ▶ między ich własnościami występują skomplikowane zależności,
- ▶ produkty nasze często modyfikujemy, reagując na sytuację na rynku, musimy więc mieć elastyczne narzędzie analityczne,
- ▶ dane o rzeczywistych procesach by-

wają niestety niskiej jakości, mogą być niekompletne, zawierać przekłamanie itp.,

▶ tradycyjne metody statystyczne, a właściwie tradycyjny sposób ich użycia wymaga dość sporo „uniwersyteckich” umiejętności.

Celem tradycyjnej analizy danych (statystyki) jest najczęściej weryfikacja pewnej hipotezy lub teorii naukowej. Tymczasem w praktyce przemysłowej zazwyczaj chodzi o jak najszybsze uzyskanie prostej podpowiedzi decyzji. Metody zgłębiania danych są właśnie tak zaprojektowane, aby uwzględnić ten oraz wypunktowane wyżej fakty. Starając się uchwycić istotę tych metod, mówi się czasem o „inteligentnych obliczeniach”, mając na myśli uczenie się maszyny (komputera) z praktyki i wskazywanie kierunku działania po pojawieniu się nowych danych. Oczywiście powinniśmy też mieć możliwość wykonania, w sposób maksymalnie ułatwiony, podstawowych analiz statystycznych oraz otrzymywania różnorodnych, zawsze użytecznych i przemawiających do wyobraźni wykresów. Bezpośrednim celem *data mining* jest sprawne i szybkie wydobywanie z danych całej zawartej w nich wiedzy. Cel ekonomiczny jest już wtedy bliski.

### Six Sigma

Metodyka Sześć Sigma (Six Sigma) to dobrze zorganizowana, bazująca na danych strategia zapewnienia jakości dotycząca wszystkich rodzajów produkcji i usług, zarządzania i innej działalności biznesowej. Coraz bardziej popularna w USA (ze względu na wiele udanych wdrożeń), ale znana też w Polsce. Sześć Sigma to, jak niektórzy mówią, filozofia realizowana w etapach: Definiuj – Mierz – Analizuj – Poprawiaj – Sprawdź (ang. DMAIC). Chodzi w niej o redukcję zmienności do poziomu obniżającego frakcję braków do 3,4 ppm (przypadków na milion możliwości). Więcej informacji o Sześć Sigma i SPC oraz ich porównanie można znaleźć w artykule „Czym się różni Sześć sigma od Trzy sigma” (dostępnym w dziale „Jakość” na wi-

trynie internetowej [www.statsoft.pl/czytelnia/czytelnia.html](http://www.statsoft.pl/czytelnia/czytelnia.html))

Nowe metody nie obalają starych i wypróbowanych, wymagają jednak efektywniejszego ich wykorzystania. *Data mining* jest w pełni zgodny z najnowszymi ideami Sześć Sigma.

### Trzy etapy

Są trzy naturalne etapy realizacji projektu *data mining*: eksploracja danych, analiza (budowanie modeli i ich ocena) i wdrożenie wyników.

Wstępnym elementem etapu eksploracji jest przygotowanie danych: czyszczenie i przekształcanie, wybór podzbiorów rekordów (przypadków), ewentualny wstępny wybór zmiennych (cech), którego celem jest inteligentne zredukowanie wielkości danych. Eksploracja obejmuje bardzo różne metody: od elementarnej regresji liniowej do wyrafinowanego badania danych metodami graficznymi i statystycznymi. Celem jej jest zbadanie struktury danych, wybranie najważniejszych cech i wskazanie ogólnej natury i koniecznego stopnia złożoności modelu, który ma dobrze opisywać rzeczywistość.

Etap drugi projektu *data mining* to budowanie modelu i ocena jego adekwatności. Rozważane są tu różne modele, po czym wybierany jest najlepszy z nich. Stosowane są różne techniki oceny i ewentualnego łączenia modeli, na przykład przez agregację, czyli głosowanie i uśrednianie (*bagging*), wzmacnianie (losowanie adaptacyjne, *boosting*), kontaminację modeli (*stacking, stacked generalizations*) i metauczenie (*meta-learning*).

W ostatnim etapie otrzymane modele podlegają wdrożeniu do życia, czyli zastosowaniu do wydobywania wiedzy z nowych, napływających na bieżąco danych. Wiedza ta wyraża się w postaci liczbowych wartości lub klasyfikacji.

Kompletny system *data mining* powinien zawierać narzędzia do wszystkich trzech etapów projektu (por. np. opis systemu *STATISTICA Data Miner* na stronie [www.statsoft.pl/dataminer.html](http://www.statsoft.pl/dataminer.html)).

## Typy zadań

W realnej działalności produkcyjnej występuje duża różnorodność zadań, do których stosuje się metody *data mining*. Z grubsza wydzielić można dwie ich grupy: bardziej bezpośrednie zagadnienia predykcyjne i ogólniejsze zadania odkrywania wiedzy. Do predykcji, czyli określania krytycznej własności na podstawie cząstkowych danych, należą: klasyfikacja, regresja, predykcja szeregów czasowych i segmentacja danych. Do metod odkrywania wiedzy należą: wykrywanie odchyleń, analiza skupień, analiza asocjacji (koszykowa), wizualizacja, zestawienia podsumowujące oraz *text mining*.

**Predykcja** dotyczy sytuacji, gdy dysponujemy zestawem kompletnych danych z przeszłości, a na bieżąco przewidywać chcemy krytyczne wartości na podstawie niepełnych danych, jakimi dysponujemy. Na przykład, mierząc cechy surowca chcemy przewidzieć cechy końcowego wyrobu. Dysponując starymi danymi o surowcu i wyrobie możemy „nauczyć” model istniejącego tu związku. Model ten powie nam później, jaki będzie (zapewne) wyrób, gdy dostarczymy mu dane o aktualnym surowcu. W banku, dysponując danymi o nowym kredytobiorcy, użyjemy modelu do zaklasyfikowania go do klasy wiarygodnych lub nie (o ile wcześniej, do budowy modelu użyliśmy danych o wielu klientach, łącznie z informacją *post factum* o spłacie kredytu).

W pierwszym przykładzie, o ile przewidujemy wartość ilościowej cechy produktu, np. twardość opony, zastosowanie miałaby **regresja**, drugi jest typowym przykładem elementarnej **klasyfikacji** typu TAK/NIE.

Szczególnym przypadkiem regresji jest przewidywanie przyszłych wartości **szeregu czasowego**. Choć wymagania w dziedzinie przewidywania zachowania się procesu produkcyjnego (albo giełdy) często mają zakres wymagający metod „magicznych”, to jednak „zwykły” **data mining** jest tu przydatny, gdyż po prostu powie nam o przyszłości to, co da się powiedzieć (na podstawie danych).

**Odkrywanie wiedzy** to zwykle wnikliwy opis danych, wskazujący na istotne struktury, zależności i prawidłowości ukryte w danych. Model *data mining* będzie tu odwzorowywał wszelkie zależności pomiędzy poszczególnymi wielkościami. W przypadku analizy polegającej na **wykrywaniu odchyleń** (w odniesieniu do określonej normy), występują pewne analogie do zwykłej analizy statystycznej. W klasycznej statystyce, w celu wykrycia odchyleń testujemy istotności różnic. Jednak wykrycie dużych różnic może nie być wystarczające. Na przykład, wykrywając defraudacje finansowe nie wystarczy weryfikować tych klientów, którzy jednorazowo złożyli na koncie duże sumy, gdyż ten sam klient może mieć kilka różnych kont bankowych, a stan każdego z nich tylko nieznacznie będzie prze-

kraczać standardowe normy. Konieczne jest szersze, ogólniejsze podejście.

W **analizie asocjacji**, zwanej też **analizą koszykową** poszukiwane są logiczne reguły wiążące, dotyczące zawartości „koszyka zakupów”. Wynikiem analizy może tu być np. stwierdzenie, że pewne typy wad produktu często występują razem.

## Co niesie przyszłość?

*Data mining* to terazniejszość nowoczesnego podejścia do danych, i na pewno przyszłość, przynajmniej ta wyobrażalna. Konieczność zgłębiania danych w ogóle jest koniecznością natury ekonomicznej. Natomiast istnienie takiej właśnie metody, jak *data mining*, wynika z aktualnego stanu technologii. Zarówno technologii produkcji jak i informatyki. Złożoność procesów przemysłowych z jednej strony i potężne oprogramowanie z drugiej tworzą harmonię pozwalającą na dalszy rozwój technologii. **MM**

## Literatura:

- [1] Berry M. J. A., Linoff G. 1997, *Data mining techniques: for marketing, sales and customer support*, Wiley & Sons.
- [2] STATISTICA Data Miner, 2002, StatSoft Inc.
- [3] Weiss S. M., Indurkha N., 1998, *Predictive data mining. A practical guide*, Morgan Kaufman Publishers.
- [4] *Data Mining – metody i przykłady*, 2002, StatSoft Polska.
- [5] Artykuły z działu *Data Mining* Czytelnia StatSoft ([www.statsoft.pl/czytelnia/czytelnia.html](http://www.statsoft.pl/czytelnia/czytelnia.html)), w szczególności „Data mining w sterowaniu procesem (QC Data Mining)”.