



DATA MINING – AUTOMAT CZY METODA NAUKOWA?

*Andrzej Sokółowski, StatSoft Polska Sp. z o.o.;
Akademia Ekonomiczna w Krakowie, Katedra Statystyki*

Wielu statystyków, głównie tych zajmujących się teorią statystyki wydaje się nie dostrzegać bujnego rozwoju gałęzi zwanej *data mining*. Niektórzy z nich odnoszą się do *data mining* z lekceważeniem lub wręcz wrogością, uznając to podejście za „nienaukowe” i podejrzane. Umożliwiające realizację złośliwych żartów o statystyce, że można dotąd poszukiwać „właściwej” metody i tak długo „męczyć” dane, aż uzyska się wyniki pożądane przez badacza lub zleceniodawcę. Wydaje się, że przynajmniej w części te opinie mają źródło w lękach spowodowanych przez nadzieje twórców pierwszych metod *data mining* zafascynowanych rozwojem komputerów. Mieli oni nadzieję, że uda się stworzyć takie programy, które będą uczyć się z danych, a następnie wykonają analizę i wyciągną wnioski bez udziału człowieka (czyli bez „nas” – tak rozumieli to niektórzy statystycy, matematycy i analitycy). Typową reakcją obronną było zamknięcie się w wysokiej wieży z kości słoniowej i strzelanie do „przeciwnika” amunicją epitetową w rodzaju: *nienaukowe, bezzalożeniowe, bez teorii, nieeleganckie, bez dowodów formalnych, prymitywne, nastawione tylko na zastosowania (sic!) itp.*

Zwykle podaje się trzy źródła, dzięki którym rozwinęło się *data mining*:

- ◆ gromadzenie bardzo dużych zbiorów danych i rosnąca świadomość, że nie są one wykorzystywane do poszukiwania prawidłowości niemożliwych do stwierdzenia metodami stosowanymi dotychczas,
- ◆ rozwój metodologii statystycznej lub ogólniej metod analizy danych,
- ◆ rozwój technologii informatycznej umożliwiający z jednej strony gromadzenie danych, ich porządkowanie, ułatwienie dostępności, a z drugiej realizację numeryczną skomplikowanych (lub „operacjochłonnych”) algorytmów w rozsądnym czasie.

Poszukiwawcza analiza danych

Od strony statystyków od wielu lat rozwijał się nurt badawczy kwestionujący omnipotencję i wyłączność klasycznych metod, eleganckich i udokumentowanych teoretycznie. Już w 1957 roku Roy [1957] zwracał uwagę na konieczność rozwoju metod nieparametrycznych, wolnych od kłopotliwych założeń. Tukey [1962] ostrzegał przed przesadną koncentracją na matematycznych aspektach metod statystycznych, ze szkodą dla analizy rzeczywistych danych. To właśnie John Tukey (magister chemii i doktor matematyki) wolał, żeby określano go mianem analityka (*data analyst*), a nie statystyka. To on w 1977 roku opublikował rewolucyjną książkę *Exploratory Data Analysis*. Tę poszukiwawczą ana-



lizę danych określił on jako pracę detektywistyczną, wykorzystującą proste metody obliczeniowe i graficzne. Jej charakterystyczne cechy to: elastyczność, praktyczność, innowacyjność, uniwersalność, prostota. Na EDA składają trzy elementy charakteryzujące proces badawczy: aktywne podejście do problemu, elastyczność i bystrość w interpretacji wyników.

Ratner [2003] określa klasyczne badanie statystyczne za pomocą następującego schematu:

Problem → Model → Dane → Analiza → Wyniki/Interpretacja

Z kolei schemat badania właściwy dla poszukiwawczej analizy danych to

Problem ↔ Dane ↔ Analiza ↔ Model → Wyniki/Interpretacja

Zauważmy, że EDA rozwijała się jeszcze przed upowszechnieniem tanich komputerów osobistych. Rozsądne więc były obawy badaczy, że poszukując prawidłowości bez szczegółowych wskazań w postaci hipotez do zweryfikowania, nie będą w stanie przegłądać wszystkich kombinacji czynników, zastosować wszystkich metod i w ten sposób mogą stracić możliwość odkrycia poszukiwanych prawidłowości. Bez wykorzystania komputerów metody propagowane przez EDA nadawały się do analizy niewielkich zbiorów danych.

Widać, że podejście EDA przygotowało grunt do rozwoju data mining wtedy, gdy pojawiły się odpowiednie maszyny cyfrowe. W literaturze można znaleźć wiele definicji data mining. Przytoczmy dwie z nich:

Hand i in. [2001]: „Data mining to analiza zazwyczaj dużych, wcześniej zgromadzonych zbiorów danych w celu odkrycia nowych prawidłowości i opisanie danych w nowy sposób, który jest zrozumiały i użyteczny dla właściciela danych”

Cabena i in. [1998]: „Data mining to interdyscyplinarne podejście wykorzystujące techniki z nauczania maszyn, rozpoznawania obrazów, statystyki, baz danych oraz wizualizacji w celu wydobycia informacji z dużych baz danych”.

Za prekursora data mining uważa się niekiedy sławnego astronoma i matematyka Johannesesa Keplera (1571–1630), który był asystentem innego sławnego astronoma - Tycho Brache (1546–1601). Brache dzięki wsparciu króla Fryderyka II zdołał zgromadzić dużą ilość wyjątkowo dokładnych danych dotyczących pozycji ciał niebieskich, obserwowalnych gołym okiem. W oparciu o te dane próbował on znaleźć jakiś schemat umożliwiający przewidywanie położenia ciał niebieskich. Próby astronoma skończyły się niepowodzeniem. Kepler postawił sobie za cel znalezienie matematycznej formuły opisującej ruchy ciał niebieskich. Studiując katalogi przez wiele lat, po wielu nieudanych próbach, ostatecznie opisał dane Tycho Brache w postaci trzech prostych praw, znanych obecnie jako prawa Keplera. Można powiedzieć, że Tycho Brache był pierwszym, który stanął przed problemem data mining, a Kepler pierwszym – który go w zadowalający sposób rozwiązał.



CRISP-DM

Pod tym tajemniczym skrótem kryje się opis typowych faz stosowania data mining opracowany w 1996 roku przez przedstawicieli trzech firm: produkcyjnej (z Niemiec), statystycznej (z Wielkiej Brytanii) i telekomunikacyjnej (z Danii). Jest to pewien standard postępowania - *The Cross-Industry Standard Process for Data Mining*. Obejmuje on sześć faz:

- ◆ *Zrozumienie problemu biznesowego* – określenie celów projektu, wyrażenie ich w języku problemów data mining, określenie wstępnej strategii osiągnięcia tych celów.
- ◆ *Poznanie danych* – zbieranie danych, wykorzystanie prostych metod analizy danych do zapoznania się z danymi, ocena jakości danych, ewentualnie wstępne określenie podzbiorów danych, które mogą zawierać informacje prowadzące do ważnych prawidłowości.
- ◆ *Przygotowanie danych* – przygotowanie wstępnego oraz ostatecznego zbioru danych, wybór zmiennych i obiektów do analizy, ewentualna analiza niektórych zmiennych, „czyszczenie danych”.
- ◆ *Modelowanie* – wybór technik modelowania, budowa modelu.
- ◆ *Ocena* – ocena zbudowanych modeli pod względem dobroci dopasowania, efektywności, interpretowalności, użyteczności w realizacji celów projektu; określenie elementów, których znaczenie nie zostało uwzględnione, wstępne określenie możliwości wdrożenia wyników w praktyce.
- ◆ *Wdrożenie* – przygotowanie raportu, wykorzystanie modeli, zastosowanie modelu do podobnego zagadnienia lub innych obiektów, ocena efektów biznesowych.

Moim zdaniem niestety te punkty są typowym produktem nauk o zarządzaniu, które wielokrotnie zajmują się zbędnym uogólnianiem rzeczy oczywistych lub zdroworoządkowych. Gdyby istniał dobry statystyk-praktyk, który nie słyszał o data mining i gdybyśmy go zapytali, jak rozwiązuje zadania zlecone mu przez klientów z wielu dziedzin, od ekonomii do medycyny, to w swojej odpowiedzi z pewnością zawarłby wszystkie elementy zawarte w CRISP-DM. Przecież zawsze zaczynamy od sformułowania i zrozumienia problemu (pierwsze należy do zlecającego, drugie do analityka), potem kompletujemy dane, kontrolujemy je, oczyszczamy z błędów, potem dokonujemy wstępnej analizy danych (w wielu dziedzinach ten etap pracy jest obecny w publikacjach w części zwanej *Przegląd materiału*), stosujemy różne metody, oceniamy ich wyniki, interpretujemy wyniki i wreszcie wykorzystujemy je w praktyce. Dobry badacz robi tak bez względu na wielkość próby, liczbę cech czy zakres przestrzenny lub czasowy analiz.

Data mining w praktyce

KD Nuggets to firma badawcza i konsultingowa. Organizuje ona w internecie szereg „głosowań” na różne tematy związane z data mining. W dniach od 18 lipca do 7 sierpnia



2005 roku zbierano zgłaszane, skuteczne zastosowania metod data mining w różnych dziedzinach, w ostatnich trzech latach. Czołówka tej listy to:

- ◆ bankowość (12% głosów),
- ◆ zarządzanie relacjami z klientami (12%),
- ◆ ocena wiarygodności kredytobiorców (8%),
- ◆ marketing bezpośredni (8%),
- ◆ wykrywanie oszustw (7%),
- ◆ ubezpieczenia (6%),
- ◆ sprzedaż detaliczna (6%),
- ◆ produkcja (5%),
- ◆ telekomunikacja (5%),
- ◆ ochrona zdrowia (4%),
- ◆ nauka (4%).

Zwraca uwagę stosunkowo duża różnorodność obszarów zastosowań. Mniejszy procent zgłoszeń odnotowały m.in.: biotechnologia, handel elektroniczny, rozrywka i muzyka, administracja publiczna, giełda, internet (głównie na polu walki z niechcianą pocztą), medycyna i farmacja, podróże oraz bezpieczeństwo i działania antyterrorystyczne. Jak widać, stosowanie data mining do czystych badań naukowych to ułamek zastosowań.

Przytoczmy wyniki jeszcze jednej ankiety, tym razem dotyczącej używanych metod i technik analitycznych, przeprowadzonej w lutym 2005 roku. Tutaj mamy następującą pierwszą dziesiątkę grup metod:

- ◆ drzewa klasyfikacyjne (14%),
- ◆ analiza skupień (13%),
- ◆ regresja (11%),
- ◆ statystyka (10%),
- ◆ wizualizacja (8%),
- ◆ sieci neuronowe (8%),
- ◆ reguły asocjacji (7%),
- ◆ najbliższe sąsiedztwo (4%),
- ◆ SVM (4%),
- ◆ wnioskowanie Bayesowskie (4%).

Powyższa lista stanowi częściowe wytłumaczenie irytacji niektórych statystyków ekspansją data mining (po części marketingowo-medialną). Za nowe propozycje, jakich dostarczyło data mining, można uznać reguły asocjacji i w pewnym sensie drzewa klasyfikacyjne. Klasyczna statystyka, regresja, analiza skupień, wnioskowanie Bayesowskie to olbrzymie grupy metod od dawna obecne w analizach; wizualizacja została spopularyzowana przez



EDA, a sieci neuronowe – to koncepcja sama w sobie wystarczająco bogata. Tak więc to wrogie przejście przez nową (?) dyscyplinę pól uprawianych przez innych i sukcesy finansowe pod nowym szyldem nie mogły pozostać bez reakcji.

Automat czy metoda naukowa?

Na pierwszą część tego pytania odpowiedź jest raczej oczywista. Berry i Linoff [1997] stwierdzili, że data mining to proces rozpoznawania i analizy, sposobami automatycznymi lub półautomatycznymi, dużych zbiorów danych w celu odkrycia znaczących prawidłowości i wzorców. W swej kolejnej książce wydanej w 2000 roku, rozważając ponownie definicje data mining przyznali, że użycie konotacji *automatyczne lub półautomatyczne* było błędem.

Larose [2005] streszcza wystąpienie J.Q. Louie, prezydenta firmy Nautilus Systems na forum Podkomitetu ds. technologii, polityki informacyjnej, stosunków międzyrządowych oraz spisów powszechnych Izby Reprezentantów z marca 2003 roku, jednocześnie konfrontując te tezy z czymś co nazywa „rzeczywistością” (*reality*). Louie wymienił sześć mitów data miningu. Przytaczamy je tu wraz z komentarzami Larose’a.

Mit 1 – *Są takie metody data mining, które można zapuścić w nasze zbiory danych i one znajdą odpowiedzi na nasze problemy.*

Rz.: Nie ma automatycznych narzędzi data mining, które na poczekaniu, mechanicznie rozwiązują problemy. Data mining to pewien proces, którego typowym opisem może być standard CRISP-DM.

Mit 2 – *Data mining to proces autonomiczny, który wymaga niewielkiego lub żadnego nadzoru człowieka.*

Rz.: Data mining na każdym etapie wymaga istotnego udziału człowieka. Nawet jeżeli model jest już wdrożony, to często napływ nowych danych wymusza korektę modelu. Analityk musi stale kontrolować jakość modelu.

Mit 3 – *Koszty zastosowania data mining szybko się zwracają.*

Rz.: Jest tu bardzo różnie. Zależy to między innymi od kosztów uruchomienia projektu, kosztów osobowych, kosztów dopasowania danych do wymogów projektu itd.

Mit 4 – *Programy komputerowe realizujące data mining są intuicyjne i łatwe w użyciu.*

Rz.: Znowu – to zależy. Analityk musi łączyć wiedzę metodologiczną ze znajomością merytorycznej strony problemu i – jak sama nazwa wskazuje – mieć zdolności analityczne.

Mit 5 – *Data mining znajdzie przyczyny twoich problemów badawczych lub biznesowych.*

Rz.: Proces odkrywania wiedzy pozwala znaleźć pewne wzorce zjawisk, ale to tylko badacz może identyfikować przyczyny.



Mit 6 – Data mining automatycznie wyczyści błędy w bazie danych.

Rz.: Tak naprawdę to nie automatycznie. W ramach pierwszego etapu data mining mamy często do czynienia z danymi, które były zbierane bardzo długo, ale nikt nie kontrolował ich poprawności.

W świetle powyższych uwag trzeba zgodzić się z opinią, iż mimo wstępnych nadziei nie udało się zbudować metodologii, która analizowałaby dane w sposób automatyczny, bez udziału człowieka. Po wczesnych fascynacjach możliwościami obliczeniowymi komputerów zauważono, że jednak sztuczna inteligencja nie może jeszcze zastąpić inteligencji badacza.

Z drugiej strony wiemy, że komputery potrafią coś, czego nie potrafi człowiek. Wykonywanie obliczeń i porównań w niewyobrażalnie krótkim czasie pozwala na „prze-trząsanie” takiej ilości informacji, która jest nie do ogarnięcia dla badacza. Co oferuje nam data mining, jakie problemy pozwala rozwiązać, wobec których bezsilne są statystyka klasyczna i rozpoznawcza analiza danych?

W każdym porządnym podręczniku statystyki znajdziemy informację o tym, że prawdziwym obiektem zainteresowania badacza jest populacja, zwana też zbiorowością generalną. Jednostki statystyczne, charakteryzowane cechami statystycznymi pozostają pod wpływem działania przyczyn głównych (które są źródłem prawidłowości) oraz przyczyn ubocznych (których efekt reprezentuje składnik losowy). Esencją statystyki jest analiza próby losowej i uogólnianie wyników na populację. Niewiele miejsca poświęca się na uzasadnienie praktycznej niemożliwości (no może poza spisem powszechnym) badań całkowitych, obejmujących całą populację, uznając rzecz za oczywistą – bo populacja jest zazwyczaj nieskończona, bo badanie może być niszczące, a w ogóle to by za dużo kosztowało. Wydaje się, że ten schemat należy zrewidować. Oprócz populacji i próby proponuję wprowadzić trzeci element – nazwijmy go *próba maksymalna*. W zasadzie pojęcie *populacji* warto by zastąpić pojęciem *zjawiska* – jako bardziej ogólnym. Zjawisko zachodzi w czasie i przestrzeni. Przy ustaleniu jednego z tych elementów lub obydwu – w wielu przypadkach mamy do czynienia ze skończoną liczbą realizacji zjawiska. Jeżeli interesujemy się pewnymi zwyczajami konsumentów, znajdującymi odzwierciedlenie w tym co i jednocześnie z czym kupują – to jest to nasze „zjawisko”. Ale w konkretnym miesiącu, w konkretnym hipermarkecie była konkretna liczba klientów (duża liczba) i to jest nasza próba maksymalna. Przy okazji trzeba zauważyć przewartościowanie pojęcia liczebności próby. To już nie magiczne 30, wynikające ze zbieżności rozkładów statystyk do rozkładu normalnego jest granicą pomiędzy małą a dużą próbą. Próba maksymalna może być bardzo duża, często wielomilionowa i wtedy statystyk pyta: „po co mam analizować całą próbę maksymalną?, wystarczy wylosować powiedzmy 50000 jednostek – a prawo wielkich liczb załatwi resztę”. Są jednak sytuacje wymagające analizy całej próby maksymalnej:

- ◆ poszukiwanie bardzo słabych wzorców i prawidłowości,
- ◆ wyszukiwanie wszystkich (!) nietypowych obiektów w próbie maksymalnej,
- ◆ klasyfikacja wszystkich obiektów z próby maksymalnej.



Są też problemy, które podejście data mining pozwala rozwiązać inaczej. Zamiast elegancji matematycznej, analitycznego poszukiwania ekstremum funkcji celu, stosuje się przeszukiwanie sieciowe, czy wręcz rozważenie wszystkich możliwości. Łatwo można sobie wyobrazić takie poszukiwanie najlepszych wartości parametrów funkcji regresji nie poprzez metodę najmniejszych kwadratów. Tutaj tak naprawdę „kopiemy” nie w danych, tylko w potencjalnych rozwiązaniach problemu.

Czy data mining to *metoda naukowa*? Sądzę, że jest to zbiór metod w większości zaczerpniętych z innych metodyk analiz, zawierający też oryginalne propozycje i to wszystko poparte techniką obliczeniową (rozumianą szeroko) pozwala pod nowym szyldem przeszukiwać duże zbiory danych. I to, co się znajduje, okazuje się bardzo przydatne od strony praktycznej. Rosnąca popularność data mining musi wynikać z korzyści (finansowych i poznawczych), jakie przynosi. Może więc nie warto pytać, czy jest to *metoda naukowa*, a raczej trzeba spytać czy jest to *metoda skuteczna*.

Literatura

1. Berry M, Linoff G. [1997], *Data Mining Techniques for Marketing, Sales and Customer Support*, Wiley, Hoboken, NJ.
2. Berry M., Linoff G. [2000], *Mastering Data Mining*, Wiley, Hoboken, NJ.
3. Cabena P., Hadjinian P., Stadler R., Verhees J., Zanasi A. [1998], *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NY.
4. Hand D., Mannila H., Smyth P., *Principles of Data Mining*, MIT Press, Cambridge.
5. Larose D.T. [2005], *Discovering Knowledge in Data – An Introduction to Data Mining*, Wiley-Interscience, Hoboken, NJ.
6. Ratner B., *Statistical Modelling and Analysis for Database Marketing – Effective Techniques for Mining Big Data*, Chapman & Hall / CRC.
7. Roy S.N. [1957], *Some Aspects of Multivariate Analysis*, Wiley, New York.
8. Tukey J. [1962], *The Future of Statistics*, Annals of Mathematical Statistics, 33, 1-67.