Zeszyty Naukowe nr 640 2003
Akademii Ekonomicznej w Krakowie

Mariusz Łapczyński
Katedra Analizy Rynku i Badań Marketingowych

Data Mining w badaniach rynkowych i merketingowych

1. Wprowadzenie

Wiele przedsiębiorstw posiada obszerne bazy danych, ale tym, czego rzeczywiście potrzebują, jest informacja. Kim są i jak zachowują się klienci? Jak skutecznie rozmieścić asortyment? Jak minimalizować straty? Pytania takie zadaje sobie zapewne większość analityków rynku pracujących z bazami danych. Naprzeciw tym rozważaniom wychodzi nowy sposób analizy danych – *Data Mining*.

Istnieje kilka definicji tego terminu. *Data Mining* to etap procesu odkrywania wiedzy w bazach danych (*Knowledge Discovery Process in Databases, KDD Process*), w którym analizuje się dane i dostarcza wzorce i modele z tych danych[1]. *Data Mining* poszukuje ukrytych związków, wzorców, zależności i współzależności w dużych bazach danych, które mogą być „niezauważone" przez tradycyjne metody pozyskiwania informacji (np. tworzenie raportów, używanie zapytań)[2]. *Data Mining* znajduje się na pograniczu wielu dziedzin: statystyki, uczenia się maszyn (*machine learning*), teorii rozpoznawania obrazów (*pattern recognition*), technologii baz danych[3] oraz sztucznej inteligencji (*artificial intelligence*) i wizualizacji danych (*data visualisation*)[4]. Pojęcie *Data Mining* nie jest obce statystykom. Uważa się[5], że jest ono synonimem pojęć *data dredging* i *fishing*, które dosłownie oznaczają dragowanie, bagrowanie (pogłębianie, poławianie z dna), poławianie danych w celu ziden-

[1] U. Fayyad, G. Piatesky-Shapiro, P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, Artificial Intelligence Magazine, Fall, 1996, s. 41.

[2] M.L. Gargano, B.G. Raggad, *Data Mining – a powerful information creating tool*, OCLC Systems & Services, vol. 15, 1999, nr 2, s. 81.

[3] D.J. Hand, *Data Mining: Statistics and More?* „The American Statistician" 1998, nr 52, s. 112.

[4] J.H. Friedman, *Data Mining and Statistics: What's the Connection?* Proceedings of the 29th Symposium on the Interface: Computing Science & Statistics, Houston, Texas, May 1997, s. 1.

[5] D.J. Hand, *Data Mining...*, 1998, nr 52, s. 112.

tyfikowania wzorców. Terminy te pojawiły się w połowie lat 60. i oznaczały „nieskrępowaną eksplorację danych”[6]. Choć nie od razu cieszyły się zainteresowaniem statystyków przyzwyczajonych do klasycznych sposobów analizy danych, to jednak już wtedy zaliczone zostały do eksploracyjnych metod analizy danych (*exploratory data analysis – EDA*) przez głównego przedstawiciela tej szkoły J.W. Tukeya.

Data Mining to termin pochodzący z języka angielskiego i tłumaczony na język polski jako:

– dogłębna analiza danych[7],
– drążenie danych[8],
– eksploatacja danych[9],
– eksploracja danych[10],
– głęboka analiza danych[11],
– przekopywanie danych[12],
– wydobywanie danych[13],
– wydobywanie reguł[14],
– wydobywanie wiedzy[15],
– zgłębianie danych[16],
– złożona analiza danych[17].

Tłumacząc dosłownie: *data* to dane, a *mining* to górnictwo, wydobywczy (o przemyśle), górniczy. *Mining* można również zastąpić słowem eksploracja, co z językowego punktu widzenia jest poprawne. Eksploracja to „badanie nieznanych dziedzin albo terenów, dociekanie, poszukiwanie”, eksplorator zaś to „badacz, po-

---

[6] *Statistical Themes and Lessons for Data Mining*, C. Glymour, D. Madigan, D. Pregibon, P. Smyth, Data Mining and Knowledge Discovery 1997, nr 1, s. 16.

[7] www.phys.uni.torun.pl/~duch/ref/ai-med/ai-med2.html – strona Wydziału Fizyki, Astronomii i Informatyki Stosowanej UMK Uniwersytetu Mikołaja Kopernika w Toruniu.

[8] www.spss.pl – strona dystrybutora pakietu SPSS.

[9] M.in.: http://galaxy.uci.agh.edu.pl/~artkraw/strony/cechygis.html – jedna ze stron witryny AGH w Krakowie.

[10] M.in.: www.cgi.btinfo.com.pl – strona Biura Tłumaczeń Informatycznych oraz www.ipipan.waw.pl/~subieta/slownik_obiektowosci/indeks_ang.html – strona Instytutu Podstaw Informatyki Polskiej Akademii Nauk.

[11] http://figaro.ae.katowice.pl/~gatnar – strona E. Gatnara – autora pierwszej w Polsce książki o drzewach klasyfikacyjnych i regresyjnych.

[12] http://galaxy.uci.agh.edu.pl/~artkraw/strony/cechygis.html – jedna ze stron witryny AGH w Krakowie.

[13] www.wiw.pl/informatyka/slownik/Haslo.asp?haslo=d – strona, na której znajduje się Słowniczek komputerowy angielsko-polski.

[14] http://sound.eti.pg.gda.pl/rekonstrukcja/zbiory_przybli_one.html – strona Politechniki Gdańskiej – Wydziału Elektroniki Telekomunikacji i Informatyki – Katedra Inżynierii Dźwięku i Obrazu.

[15] www.ask.eti.pg.gda.pl/~altair/dm – strona Politechniki Gdańskiej – Wydziału Elektroniki Telekomunikacji i Informatyki.

[16] www.statsoft.pl – strona dystrybutora pakietu Statistica.

[17] http://www.teleinfo.com.pl/ti/2000/47/t19.html

szukiwacz”[18]. O eksploracyjnych właściwościach *Data Mining* piszą m.in.: J.H. Friedman ze Stanford University w USA: „ze statystycznego punktu widzenia *Data Mining* może być postrzegane jako komputerowo zautomatyzowana eksploracyjna analiza danych”[19], J. Maindonald z Australian National University: „jednym z celów *Data Mining* jest eksploracyjna analiza obszernych baz danych”[20] czy D.J. Hand z The Open University w Wielkiej Brytanii: „budowa modeli czyni *Data Mining* podobne do konwencjonalnych eksploracyjnych metod statystycznych”[21]. Co prawda, eksploracja to zadanie geologów (*geologists*) – to oni, a nie górnicy poszukują złóż, a nie górników (*miners*) – to oni je eksploatują, ale jeśli przyznamy, że chodzi o metaforę, to tę część wywodu dotyczącego tłumaczenia tego terminu można pominąć. Dlatego właśnie termin *Data Mining* będzie w niniejszym artykule pojawiał się na przemian ze zwrotami: „eksploracja danych” albo „eksploracyjne metody analizy danych”. Pozostając przy tym tłumaczeniu, nie oddaje się ani całej istoty *Data Mining,* ani całej istoty eksploracyjnych metod analizy danych (w tym tych klasycznych statystycznych)[22].

2. Właściwości *Data Mining*

Eksploracyjne metody analizy danych pozwalają odkryć wzorce i zależności między zmiennymi w obszernych zbiorach obserwacji. Te wzorce i zależności są prezentowane w postaci modeli predykcyjnych i deskryptywnych (opisowych). Modele predykcyjne umożliwiają budowę prognoz, natomiast modele deskryptywne przedstawiają wzorce w istniejących zbiorach obserwacji. Zgodnie z definicją[23] predykcja to „proces ekonometrycznego wnioskowania w przyszłość”, natomiast prognoza to „ostateczny wynik tego procesu”. Modele predykcyjne pokazują, jak zmieni się wartość zmiennej zależnej, jeśli zmienią się wartości zmiennych niezależnych. Predykcja może dotyczyć zmiennych ilościowych (mowa wtedy o predykcji ilościowej i zmiennych prognozowanych) lub zmiennych jakościowych (mowa wtedy o predykcji jakościowej i zdarzeniach prognozowanych). Istnieją narzędzia *Data Mining* (np. algorytm drzew klasyfikacyjnych CART), które pozwalają na predykcję zarówno zmiennych mierzalnych, jak i niemierzalnych. Modele takie mają zastosowanie np. w ocenie zdolności kredytowej klientów banku. Kredytodawca ocenia wiarygodność kredytobiorcy, oceniając takie zmienne, jak np. średnie miesięczne wynagro-

[18] W. Kopaliński, *Słownik wyrazów obcych i zwrotów obcojęzycznych z almanachem*, Klub Świat Książki, Warszawa 2000, s. 145.

[19] J.H. Friedman, *op. cit.*, s. 1.

[20] J. Maindonald, *Data Mining from a Statistical Perspective*, artykuł pobrany ze strony Australian National University, wwwmaths.anu.edu.au, s. 5.

[21] D.J. Hand, *Data Mining...*, s. 116.

[22] Jednak jeśli spojrzeć na termin „strach na wróble” – to on również nie wyczerpuje wszystkich sposobów płoszenia wróbli, a tak naprawdę, to dotyczy również innych gatunków ptaków.

[23] Z. Pawłowski, *Zasady predykcji ekonometrycznej*, PWN, Warszawa 1982, s. 26.

dzenie, czas pracy w obecnej firmie, liczbę dzieci, stosunek do służby wojskowej, wysokość zadłużenia w innych bankach. Wartości tych zmiennych pozwalają przewidzieć, czy klient kredyt spłaci, czy nie. Modele deskryptywne przedstawiają natomiast wzorce w istniejących zbiorach obserwacji, albo są wykorzystywane jako pomoc przy konstruowaniu modeli predykcyjnych. Predykcja i opis to dwa nadrzędne cele *Data Mining*. Granica między tymi celami jest dość płynna, gdyż model predykcyjny zbudowany za pomocą drzew klasyfikacyjnych może być na tyle czytelny, że staje się modelem opisowym. I podobnie: model opisowy zbudowany przy użyciu sztucznych sieci neuronowych może być na tyle skomplikowany, że wykorzystany zostanie wyłącznie do predykcji. Różnica między obydwoma typami modeli sprowadza się czasem do dalszego ich wykorzystania przez menedżerów.

Wiele przedsiębiorstw posiada własne wzorce i zależności będące wynikiem wieloletnich obserwacji i analizy rynku. W takiej sytuacji eksploracyjne metody analizy danych, poza potwierdzeniem już znanych wzorców i zależności pomagają znaleźć nowe, nieodkryte jeszcze reguły. Ponadto *Data Mining* pozwala monitorować zmieniającą się rynkową rzeczywistość.

Eksploracyjne metody analizy danych nie eliminują potrzeby zrozumienia branży. *Data Mining* pozwoli znaleźć wzorce i zależności w bazach danych, jednak nie oceni ich przydatności dla firmy. To analityk i specjalista ds. merchandisingu oceniają, czy opłaca się nabyć pikantny sos ośmiornicowo-kalmarowy. Ilu klientów będzie skłonnych go kupić? Ile wydają ci klienci? Jaki jest koszt magazynowania? Czy lokalne hurtownie artykułów spożywczych mają taki produkt? Jeśli nie, to gdzie go można kupić?

Data Mining nie zwalnia też z obowiązku znajomości poziomów pomiaru analizowanych zmiennych i umiejętności kodowania danych. Znając poziom pomiaru zmiennych niezależnych i zależnych, wiadomo jakie narzędzie zastosować. Jeśli np. zmienna zależna mierzona jest na skalach mocnych, to drzewo ma charakter regresyjny; jeśli natomiast zmienna ta mierzona jest na skalach słabych, to drzewo ma charakter klasyfikacyjny. Wiedza nt. kodowania danych ułatwi z kolei pracę z oprogramowaniem różnych producentów (np. CART, STATISTICA, SPSS, See5).

Data Mining nie wyklucza potrzeby gruntownego poznania metodologii. Stosując eksploracyjne metody analizy danych trzeba znać narzędzie, które się stosuje oraz algorytm, na którym to narzędzie jest zbudowane. Nie można stosować drzew klasyfikacyjnych, nie wiedząc, co to węzeł końcowy, macierz błędnych klasyfikacji czy podział rekurencyjny. Znając właściwości drzew klasyfikacyjnych, trzeba z kolei wybrać odpowiedni algorytm (CART, CHAID, C4.5 czy QUEST), regułę podziału (np. Giniego, „dwójkowania") i kryterium stopu.

3. Modele *Data Mining*

Wyróżnia się 6 typów modeli *Data Mining*:
1) dyskryminację (*classification*),
2) regresję (*regression*),

3) szeregi czasowe (*time series*),
4) klasyfikację (*clustering*),
5) skojarzenia (*association*),
6) sekwencje (*sequence*).

Z punktu widzenia teorii rozpoznawania obrazów można podzielić je na modele rozpoznawane z nauczycielem (*supervised learning* 1–3) i modele rozpoznawane bez nauczyciela (*unsupervised learning* 4–6). Rozpoznawanie z nauczycielem polega w skrócie na tym[24], że niesklasyfikowane przypadki trafiają do zbioru obserwacji zwanego zbiorem rozpoznawanym. Następnie przydzielane są do poszczególnych klas, których charakterystyki są określone na podstawie zbioru uczącego (nauczyciela, zbiór przypadków zbadanych wcześniej).

Z punktu widzenia taksonometrii modele te można podzielić na modele taksonomii wzorcowej (1–3) i modele taksonomii bezwzorcowej (4–6). Można również uznać[25], że pierwsze trzy modele to zadania klasyfikacji z nauczycielem, a ostatnie trzy modele to zadania klasyfikacji bez nauczyciela, automatycznej klasyfikacji lub grupowania. Z punktu widzenia celów *Data Mining* modele te można podzielić na predykcyjne (dyskryminacja, regresja i szeregi czasowe) i deskryptywne (klasyfikacja, skojarzenia i sekwencje).

Dyskryminacja. Celem modeli dyskryminacyjnych jest zrozumienie istniejących danych i predykcja zachowań nowych przypadków. Zgodnie z definicją „zagadnienie dyskryminacji polega na przydzieleniu zbioru obserwacji do K klas mających własność jednorodności, przy czym charakterystyki tych klas są przynajmniej częściowo znane"[26]. W literaturze anglojęzycznej zagadnienie to nazywane jest *classification*, jednak autorzy mają na myśli nie klasyfikację, a dyskryminację[27]. Chodzi im o przyporządkowanie nowych przypadków do zdefiniowanych wcześniej klas. Te nowe przypadki pochodzą ze zbioru rozpoznawanego (rozpoznaje się, do jakich klas należą), natomiast charakterystyki klas są wynikiem analizy zbioru uczącego (w wyniku analizy tego zbioru poznano liczbę i charakterystyki klas).

Regresja. Zgodnie z definicją[28] regresja to ważne narzędzie statystyczne służące do badania związków między zjawiskami. Dotyczy zależności zmiennej losowej Y od zmiennych losowych X_1, X_2, ..., X_n i zwykle znajduje zastosowanie w analizach popytu lub jest wykorzystywana do budowy prognoz ekonomicznych. Jeśli nie wystarcza klasyczna liniowa analiza regresji, to można wykorzystać sztuczne sieci neuronowe. Jednak i wtedy istnieje możliwość, że złożoność funkcji zależności utrudni merytoryczną interpretację takiej sieci. Z jednej strony zrealizowany zostaje wyłącznie predyktywny, a nie poznawczy cel analizy – z drugiej zaś, to głównie o to chodzi podczas budowy modeli predykcyjnych w *Data Mining*.

---

[24] K. Jajuga, *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993, s. 134.

[25] J. Kolonko, *Analiza dyskryminacyjna i jej zastosowania w ekonomii*, PWN, Warszawa 1980, s. 15.

[26] K. Jajuga, *op.cit.*, s. 134.

[27] U. Fayyad, G. Piatesky-Shapiro, P. Smyth, *op.cit.*, s. 44.

[28] *Mała encyklopedia statystyki*, pod red. W. Sadowskiego, PWE, Warszawa 1976, s. 522.

Analiza szeregów czasowych. W analizie szeregów czasowych, podobnie jak w analizie regresji, wykorzystuje się istniejące dane do prognozy przyszłych obserwacji. Zgodnie z definicją: „prognozowanie to racjonalne, naukowe przewidywanie przyszłych zdarzeń"[29]. Przewidując zmienną zależną na podstawie analizy szeregu czasowego, należy jednak pamiętać o składowych takiego szeregu, tj. wahaniach cyklicznych, wahaniach sezonowych, trendzie, stałym (średnim) poziomie czy wahaniach przypadkowych. Nagły wzrost sprzedaży kwiatów i alkoholu 29 czerwca wynika z popularnych w tym dniu imienin. Gdyby przypadkiem popularne stało się imię Eudoksjusz, to wzrost sprzedaży w kwiaciarniach i sklepach z alkoholem przypadłby na 2 listopada. Poza składowymi szeregu czasowego należy pamiętać o zdefiniowaniu okresu prognozy, czyli okresu, którego prognoza dotyczy. Przykładowo: tygodniowa prognoza sprzedaży dla hipermarketu Tesco dotyczyć będzie 7 dni (od poniedziałku do niedzieli), tygodniowa prognoza sprzedaży dla zwykłego sklepu spożywczego – 6 dni (od poniedziałku do soboty), a tygodniowa prognoza sprzedaży dla Zamku Królewskiego na Wawelu 6 dni (od wtorku do niedzieli).

Klasyfikacja. Modele klasyfikacyjne nazywane są w literaturze anglojęzycznej *clustering*. W przeciwieństwie do dyskryminacji należącej do metod taksonomii wzorcowej, klasyfikacja należy do metod bezwzorcowych. Oznacza to, że dzieląc zbiór obserwacji na *n* rozłącznych podzbiorów nie wiadomo *a priori*, jakie będą charakterystyki tych podzbiorów. Nie ma rozpoznanych wcześniej wzorców klas. Dlatego uważa się, że jest to automatyczna klasyfikacja albo grupowanie.

Skojarzenia (*asocjacje*). Modele skojarzeniowe przedstawiają współwystępowanie wartości różnych zmiennych w danym wypadku. Modele skojarzeniowe (asocjacyjne) mają postać zdań warunkowych, w których pojawia się spójnik międzyzdaniowy: „jeżeli zdanie Z_1, to zdanie Z_2". Używając tego spójnika w mowie potocznej, przyjmuje się, że między zdaniami składowymi istnieje powiązanie rzeczowe lub formalne, tzn. pierwsze zdanie Z_1 implikuje drugie Z_2. Z punktu widzenia logiki związki między poprzednikiem Z_1 (*antecedent*) a następnikiem Z_2 (*consequent*) mogą mieć różnoraki charakter, jednak w wypadku badań rynkowych i marketingowych mowa o związkach przyczynowo-skutkowych i strukturalnych (tj. takich, które wynikają z rozmieszczenia przedmiotów w przestrzeni albo zdarzeń w czasie). Reguły skojarzeniowe są najczęściej wykorzystywane przez komórki merchandisingu w analizie koszykowej (*market basket analysis*).

Okrywanie sekwencji. Sekwencje to skojarzenia, w których poprzednik występuje znacznie wcześniej niż następnik. O ile reguła skojarzeniowa może brzmieć: „jeśli kupi wiertarkę, to kupi wiertła" (oba produkty kupiono w tym samym czasie – reguła dotyczy jednej transakcji), to reguła sekwencyjna będzie brzmieć: „jeśli kupi wiertarkę, to najpóźniej po 2 miesiącach kupi szlifierkę". Należy jednak pamiętać, że odkrywanie sekwencji może mieć praktyczne zastosowanie wyłącznie wówczas, gdy firma posiada dane o swych klientach i rejestruje wszystkie dokony-

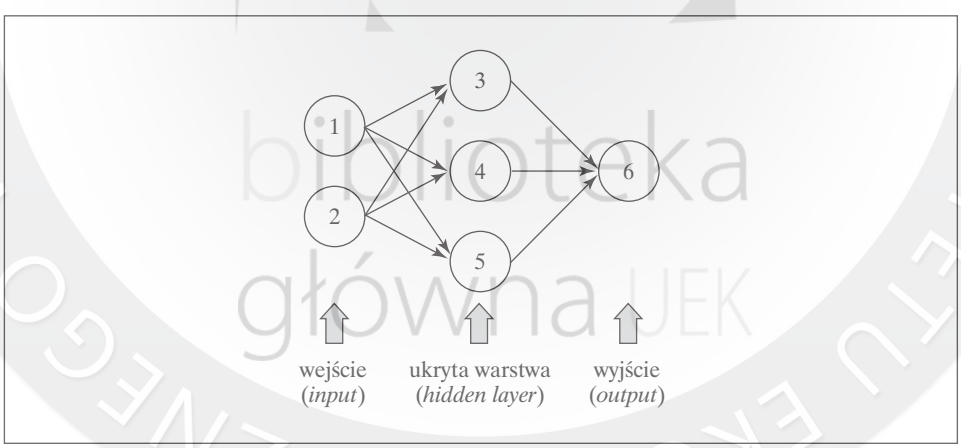
[29] P. Dittman, *Metody prognozowania sprzedaży w przedsiębiorstwie*, Wydawnictwo AE we Wrocławiu, Wrocław 1998, s. 19.

wane przez nich transakcje. Dotyczy to przede wszystkim instytucji finansowych (banków, firm ubezpieczeniowych), albo transakcji dokonywanych przy użyciu kart kredytowych czy kart stałego klienta.

4. Narzędzia *Data Mining*

Za narzędzia *Data Mining* uznaje się: sztuczne sieci neuronowe, drzewa klasyfikacyjne i regresyjne, reguły indukcyjne, metody wnioskowania opartego na doświadczeniach z przeszłości (*case-based reasoning*), algorytmy genetyczne oraz reguły skojarzeniowe i odkrywanie sekwencji. Czasami uznaje się również[30]: zbiory rozmyte (*fuzzy sets*), sieci Bayesowskie (*Bayesian networks*) i samoorganizujące się mapy (*self-organizing maps*), a czasami[31]: regresję logistyczną, analizę dyskryminacyjną i uogólnione modele addytywne (*Generalized Additive Models*).

Sztuczne sieci neuronowe. Sztuczna sieć neuronowa to narzędzie *Data Mining*, którego pierwowzorem jest mózg ludzki. Sieć taka składa się z wielu elementów podstawowych zwanych neuronami. Obliczenia wykonywane przez neurony nie są skomplikowane, jednak ich znaczna liczba, powiązania między nimi oraz układ warstwowy (rys. 1) powodują, że sztuczna sieć neuronowa nadaje się do przeprowadzenia bardzo złożonych operacji obliczeniowych. Neurony oznaczono jako koła ponumerowane od 1 do 6. Strzałki to powiązania między neuronami. Wszystkie zwrócone są w jedną stronę, gdyż jest to schemat sieci jednokierunkowej – znajdu-



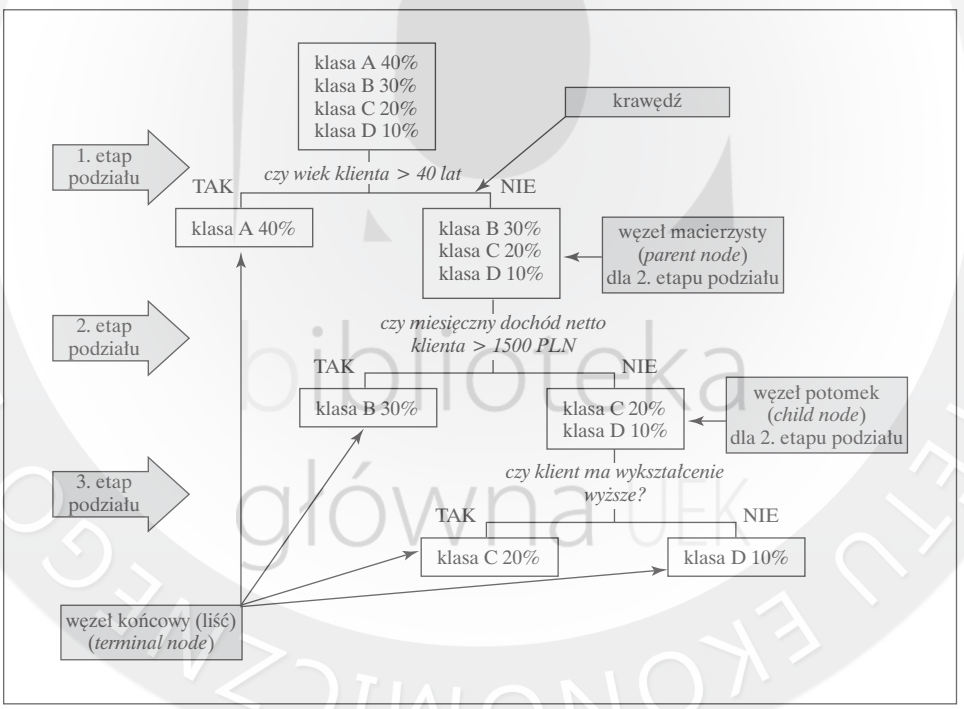
Rys. 1. Schemat sztucznej sieci neuronowej
Źródło: opracowanie własne.

[30] J.H. Friedman, *op.cit*.

[31] *Introduction to Data Mining and Knowledge Discovery – 2nd edition*, The Two Crows Corporation 1998, tekst pobrano ze strony www.twocrows.com

jącej najczęstsze zastosowanie w praktyce (sieci ze sprzężeniami zwrotnymi wykorzystywane są głównie w pracach badawczych). Zauważyć można ponadto trzy poziomy: warstwę wejściową, warstwę ukrytą i warstwę wyjściową. Istotną cechą sztucznych sieci neuronowych jest zdolność uczenia się. Warstwa ukryta (może być ich kilka) działa na zasadzie czarnej skrzynki. Przekształca sygnały wejściowe (np. informacje o klientach banku) w sygnały wyjściowe (np. w zmienną binarną: „spłacił kredyt" – „nie spłacił kredytu") automatycznie, bez ingerencji badacza. Co więcej użytkownik nie musi znać badanej dziedziny, a sieć uczy się na podstawie bardzo dużej liczby przypadków. Najpopularniejszym algorytmem obliczeniowym jest algorytm wstecznej propagacji błędów (*backpropagation*)[32], który w bardzo wielkim skrócie polega na rzutowaniu błędu pojawiającego się w danym neuronie do wszystkich neuronów, z których wychodzą sygnały do tego neuronu.

Drzewa klasyfikacyjne i regresyjne. Drzewo klasyfikacyjne (albo regresyjne) jest graficznym modelem powstałym w wyniku rekurencyjnego podziału zbioru obserwacji. Przykład drzewa klasyfikacyjnego przedstawiono na rys. 2. Podział rekuren-



Rys. 2. Przykład drzewa klasyfikacyjnego
Źródło: opracowanie własne.

[32] R. Tadeusiewicz, *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa 1993, s. 59 i nast.

cyjny (*recursive partitioning*) polega na podziale zbioru A na *n* rozłącznych podzbiorów A_1, A_2, A_3, ..., A_n. Podział rekurencyjny jest procesem wieloetapowym, przy czym na każdym etapie może być dokonywany na podstawie innej zmiennej niezależnej. Jeżeli zmienna zależna jest wyrażona na skalach słabych, to drzewo nazywa się drzewem klasyfikacyjnym, jeżeli na skalach mocnych – drzewem regresyjnym.

Do klasycznych algorytmów podziału drzew klasyfikacyjnych zalicza się: CART (*Classification and Regression Trees*) autorstwa L. Breimana i in., CHAID (*Chi-squared Automatic Interaction Detection*)[33] autorstwa G.V. Kassa oraz C4.5 autorstwa J.R. Quinlana. Inne, a jest ich ok. 20, to przede wszystkim modyfikacje 3 powyższych. Jeśli chodzi o implementacje tych algorytmów, to są obecne albo w powszechnie dostępnych pakietach statystycznych (Statistica, SPSS), albo jako odrębne oprogramowanie – implementacje pojedynczych algorytmów (CART, See5), albo jako część pakietów do eksploracyjnej analizy danych (DB Miner), albo w oprogramowaniu do analizy drzewkowej wykorzystującym wiele algorytmów podziału (Sipina).

Reguły indukcyjne. Reguły indukcyjne, to zdania warunkowe o postaci „jeżeli warunek, to klasa", które oznaczają, że spełnienie danego warunku świadczy o przynależności do klasy. Można je otrzymać poprzez bezpośrednie mechaniczne opisanie węzłów końcowych drzewa klasyfikacyjnego albo – w bardziej złożony sposób – poprzez wykorzystanie odpowiedniego algorytmu (np. CN2). Implementacją innego algorytmu – C4.5 – jest program See 5. Na rys. 3 przedstawiono wynik analizy wykonanej tym programem. „A" oznacza atrybut, czyli zmienną, zatem interpretacja reguły nr 10 będzie brzmieć następująco: „jeżeli zmienna nr 5 przyjmie wartość *q* i zmienna nr 7 przyjmie wartość *h* i zmienna nr 11 będzie ≤ 3, to obiekt będzie należał do klasy +". 96,4% obiektów, dla których zmienne: 5, 7 i 11 spełniały te warunki należało do tej klasy.

Metoda k-najbliższego sąsiedztwa (*case based reasoning*). Wnioskowanie oparte na doświadczeniach z przeszłości (*case based reasoning,* w skrócie *CBR*)[34] służy do rozwiązywania problemów teraźniejszości i przyszłości. Nie opiera się wyłącznie na ogólnej znajomości danej dziedziny, czy uogólnionych relacjach zachodzących między zmiennymi, ale na konkretnych przeszłych zdarzeniach (przykładach). W skrócie polega to na przypomnieniu sytuacji z przeszłości podobnej do aktualnie rozważanej i wykorzystaniu wiedzy o tym przypadku obecnie. Wyróżnia się kilka typów[35] wnioskowania opartego na doświadczeniach z przeszłości:

– wnioskowanie oparte na przykładach (*exemplar-based reasoning*),
– wnioskowanie oparte na przypadkach (*instance-based reasoning*),
– wnioskowanie oparte na pamięci (*memory-based reasoning*),

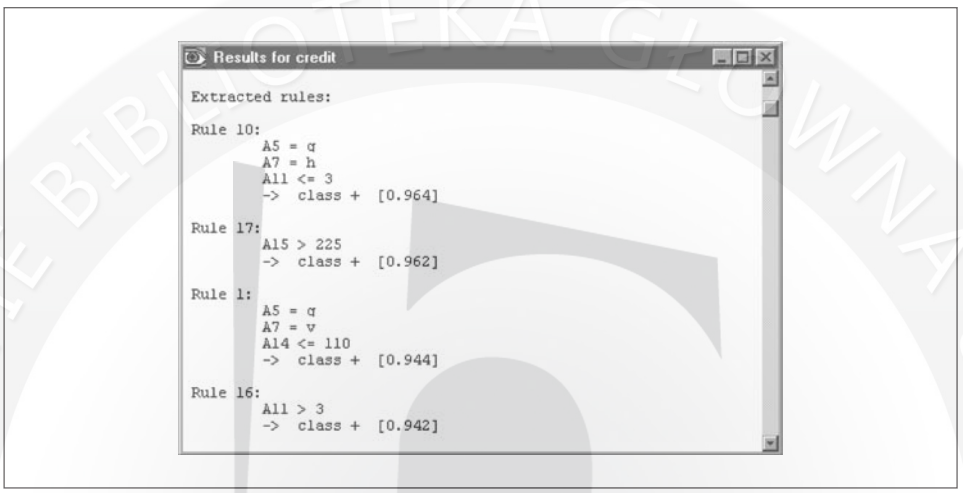
---

[33] Czasami skrót CHAID jest rozwijany jako Chi-squared Automatic Interaction Detector.

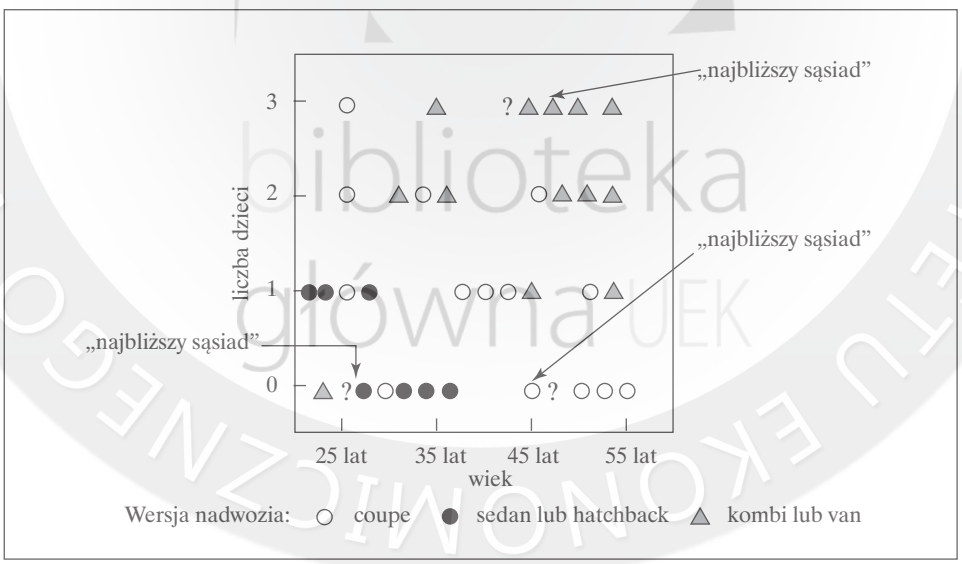
[34] Można też spotkać inne tłumaczenia tego terminu, np. rozumowanie oparte na precedensach, wnioskowanie na podstawie przykładów czy planowanie przez przykłady.

[35] A. Aamodt, E. Plaza, *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*, Artificial Intelligence Communications, IOS Press 1994, vol. 7:1, s. 39–59.

– wnioskowanie oparte na doświadczeniach (*case-based reasoning*),
– wnioskowanie oparte na analogii (*analogy-based reasoning*).



Rys. 3. Reguły indukcyjne otrzymane w programie See5
Źródło: opracowanie własne.



Rys. 4. Przykład wnioskowania opartego na doświadczeniach z przeszłości (wiek klienta i liczba dzieci a wersja nadwozia nabytego samochodu)
Źródło: opracowanie własne.

Choć podejścia te różnią się nieznacznie, to można powiedzieć, że każde z nich jest procesem cyklicznym składającym się następujących etapów (4R):

1) wyszukiwanie najbardziej podobnych przypadków (*retrieve*),

2) ponowne wykorzystanie informacji o tych przypadkach do rozwiązania bieżącego problemu (*reuse*),

3) korygowanie zaproponowanego rozwiązania (*revise*),

4) zachowanie bieżącego rozwiązanego problemu do rozwiązania przyszłych problemów (*retain*) – nowy przypadek staje się częścią zbioru przypadków z przeszłości.

Pierwszy etap tego cyklu jest też nazywany „poszukiwaniem najbliższego sąsiada" (*nearest neighbor*) w bazie danych zawierającej przypadki z przeszłości (rys. 4.)

Algorytmy genetyczne. O ile twórcy sztucznych sieci neuronowych wzorowali się na układzie nerwowym człowieka, o tyle twórcy algorytmów genetycznych obserwowali ewolucję gatunków. To właśnie selekcja naturalna i dziedziczenie zainspirowały pionierów tej metody. Algorytmy genetyczne przejęły część terminologii biologicznej[36], np. populacje, osobniki, chromosomy, geny, genotypy, fenotypy, allele czy *loci*. Wykorzystywane są w sztucznej inteligencji, współpracują z systemami rozmytymi, sztucznymi sieciami neuronowymi lub są metodą niezależną. Ich głównym zastosowaniem jest optymalizacja procesów, choć ich współpraca z sieciami oznacza, że pośrednio realizują cele im stawiane, tj. opisują zjawiska lub budują prognozy.

5. Wykorzystanie *Data Mining* w badaniach rynkowych i marketingowych

Termin *Data Mining* jest kojarzony z terminem *Database Marketing* (marketing oparty na bazach danych). Jako etap procesu odkrywania wiedzy w bazach danych ma z nimi wiele wspólnego. To właśnie liczba gromadzonych danych i moc obliczeniowa komputerów przyczyniły się do rozkwitu tej dziedziny na początku lat 90. ubiegłego stulecia. Przedsiębiorstwa posiadają dane o swoich klientach i transakcjach przez nich dokonywanych, które zbiera się w analityczne bazy danych (zwane hurtowniami danych) i targowiska danych (*data marts*, czyli tematyczne hurtownie danych). Ich zadaniem jest wspomaganie decydentów w procesie podejmowania decyzji. Firmy wprowadzają system CRM (*Customer Relationship Management*) – system zarządzania kontaktów z klientami. Analizę tych dużych zbiorów obserwacji ułatwiają narzędzia *Data Mining*.

Ścisły związek z bazami danych nie oznacza wcale, że narzędzia eksploracyjnych metod analizy danych mogą być stosowane wyłącznie do terabajtów informacji w hurtowniach danych. Z powodzeniem zastępują klasyczne statystyczne meto-

[36] D. Rutkowska, M. Piliński, L. Rutkowski, *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, PWN, Warszawa 1997, s. 130 i nast.

dy analizy danych ankietowych, gdzie zbiór obserwacji liczy z reguły mniej niż 2 tys. przypadków.

Analiza koszykowa (*market basket analysis*) opisuje transakcje dokonywane przez klientów supermarketów za pomocą reguł skojarzeniowych (*association rules*). Reguła skojarzeniowa przyjmuje postać: „jeśli zdanie Z_1, to zdanie Z_2" np. „jeśli kupił produkt A, to kupił produkt B". Z formalnego punktu widzenia jest to zdanie warunkowe, w którym produkt A jest poprzednikiem, a produkt B następnikiem. Reguły skojarzeniowe dzielimy na jakościowe (Boolowskie) i ilościowe. W wypadku tych pierwszych uzyskuje się informacje o tym, jaki produkt z następnika został kupiony razem z produktem z poprzednika. Co do ilościowych reguł skojarzeniowych, to poza informacją nt. rodzaju produktu dostarcza informację o nabytej ich liczbie, np. „jeśli kupił 2 kg produktu A, to kupił 3 sztuki produktu *B*". Istnieją też inne podziały reguł skojarzeniowych:

– na jednowymiarowe (zawierające jeden poprzednik) i wielowymiarowe (zawierające kilka poprzedników, np. „jeśli kupił produkt A i produkt C, to kupił produkt B"),

– na jednopoziomowe i wielopoziomowe (zawierające bardziej szczegółowe informacje o nabytych produktach, np. „jeśli kupił produkt A marki Z w opakowaniu 0,5 l, to kupił produkt B marki Y").

Analiza koszykowa znajduje uznanie u specjalistów ds. merchandisingu (pozwala na efektywne rozmieszczenie asortymentu), specjalistów ds. promocji (wiedzą, jakie produkty powinny być promowane razem) i specjalistów ds. logistyki (przewidzą, jak braki w zaopatrzeniu jednych produktów spowodują spadek sprzedaży innych).

Badania segmentacji i selektywności rynku to kolejny obszar zastosowań narzędzi *Data Mining*. Przykładowo drzewa klasyfikacyjne i regresyjne pozwalają analizować zmienne ilościowe i jakościowe. Zwłaszcza te drugie dość często występują w badaniach segmentacyjnych[37]. Identyfikuje się homogeniczne grupy konsumentów na podstawie ich stosunku do marek produktów lub innych instrumentów marketingowego oddziaływania. Coraz popularniejsza staje się segmentacja psychograficzna konsumentów. To właśnie niemierzalny charakter zmiennych opisujących segmenty i niemierzalne cechy produktów w badaniach selektywności powodują, że coraz częściej do profilowania segmentów używa się narzędzi eksploracyjnych metod analizy danych. Inne praktyczne zastosowanie spotyka się w marketingu bezpośrednim – chodzi o przewidywanie zachowań konsumentów na otrzymaną pocztą ofertę i racjonalne planowanie wysyłki tych ofert.

Identyfikacja przyczyn nielojalności klientów (*churn analysis*) to kolejny duży obszar zastosowań eksploracji danych. W skrócie chodzi o to, aby oprócz obliczenia wskaźnika retencji[38] określić przyczyny braku satysfakcji i przyczyny nielojal-

[37] A. Sagan, *Badania marketingowe. Podstawowe kierunki*, Wydawnictwo AE w Krakowie, Kraków 1998, s. 157 i nast.

[38] Wskaźnik retencji wyraża procentowy stosunek liczby klientów firmy pod koniec ustalonego okresu do liczby jej klientów na początku tego okresu.

ności konsumentów. Analiza danych o transakcjach pozwala zapobiec dalszej utracie klientów przedsiębiorstwa. Praktyczne jej zastosowanie występuje na rynku telekomunikacyjnym – to klienci operatorów sieci stacjonarnych i sieci komórkowych najczęściej zmieniają dostawców usług.

Innymi obszarami zastosowań narzędzi *Data Mining* są:
- prognozowanie kursów giełdowych,
- prognozowanie sprzedaży,
- optymalizacja działalności handlowej,
- ocena zdolności kredytowej klientów banku,
- analiza i modelowanie kursów walut,
- analiza prawidłowości na rynku obligacji,
- prognozowanie łączenia się korporacji,
- prognozowanie wpływów do budżetu państwa,
- analiza odwiedzin stron internetowych (*web mining*).

6. Zakończenie

Szybki rozwój informatyki na początku lat 90. przyczynił się do popularyzacji dziedziny zwanej *Data Mining*. To właśnie wzrost mocy obliczeniowej komputerów umożliwił implementację algorytmów, które powstawały już od początku lat 60. Jeśli chodzi o Polskę, to najwcześniej zainteresowano się sztucznymi sieciami neuronowymi, a najpóźniej drzewami klasyfikacyjnymi i regresyjnymi. *Data Mining* pozwala na budowę modeli predykcyjnych i opisowych na podstawie dużych zbiorów obserwacji gromadzonych w hurtowniach danych. Wykorzystuje się w tym celu zestaw narzędzi, w skład którego wchodzą m.in.: sztuczne sieci neuronowe, drzewa klasyfikacyjne i regresyjne, reguły indukcyjne, algorytmy genetyczne czy metoda *k*-najbliższego sąsiedztwa.

Duża elastyczność narzędzi *Data Mining* przejawiająca się możliwością analizy zmiennych wyrażonych na różnych poziomach pomiaru spowodowała, że szybko znalazły się one w kręgu zainteresowań badaczy marketingowych. Zaczęto je wykorzystywać nie tylko w marketingu opartym na bazach danych, ale również w analizie mniejszych zbiorów przypadków. Przedstawione w niniejszej pracy obszary zastosowań eksploracyjnych metod analizy danych w badaniach rynkowych i marketingowych są tylko wstępem do bardzo złożonego i wymagającego kolejnych opracowań zagadnienia.

Literatura

Aaamodt A., Plaza E., *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*, Artificial Intelligence Communications, IOS Press, 1994, vol. 7:1.

Berry M.J.A., Linoff G., *Data Mining Techniques For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc., 1997.

Fayyad U., Piatesky-Shapiro G., Smyth P., *From Data Mining to Knowledge Discovery in Databases*, Artificial Intelligence Magazine, Fall, 1996.
Friedman J.H., *Data Mining and Statistics: What's the Connection?* Proceedings of the 29th Symposium on the Interface: Computing Science & Statistics, Houston, Texas, May 1997.
Gargano M.L., Raggad B.G., *Data Mining – a powerful information creating tool*, OCLC Systems & Services, vol. 15, 1999, nr 2.
Gatnar E. *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa 2001.
Hand D.J., *Data Mining: Statistics and More?* „The American Statistician", 1998, nr 52.
Hand D.J., *Statistics and Data Mining: Intersecting Disciplines*, SIGKDD Explorations, ACM SIGKDD, vol. 1, Issue 1., June 1999.
Introduction to Data Mining and Knowledge Discovery – 2nd edition, The Two Crows Corporation 1998, tekst pobrano ze strony www.twocrows.com
Jajuga K., *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa 1993.
Kolonko J., *Analiza dyskryminacyjna i jej zastosowania w ekonomii*, PWN, Warszawa 1980.
Lee S.J., Siau K., *A Review of Data Mining Techniques*, Industrial Management & Data Systems, 2001, nr 101.
Maindonald J., *Data Mining from a Statistical Perspective*, artykuł pobrany z witryny internetowej Australian National University, wwwmaths.anu.edu.au, plik pobrano w październiku 2001.
Mannila H., *Theoretical Frameworks for Data Mining*, SIGKDD Explorations, ACM SIGKDD, vol. 1, Issue 2., January 2000.
Marketingowe testowanie produktów, pod red. S. Sudoł, J. Szymczak, M. Haffer, PWE, Warszawa 2000.
Rutkowska D., Piliński M., Rutkowski L., *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, PWN, Warszawa 1997.
Sagan A., *Badania marketingowe. Podstawowe kierunki*, Wydawnictwo AE w Krakowie, Kraków 1998.
Statistical Themes and Lessons for Data Mining, C. Glymour, D. Madigan, D. Pregibon, P. Smyth, Data Mining and Knowledge Discovery 1997, nr 1.
Tadeusiewicz R., *Sztuczne sieci neuronowe*, Akademicka Oficyna Wydawnicza RM, Warszawa 1993.

Data Mining in Market and Marketing Research

This paper is aimed as an introduction to Data Mining and to show areas of market and marketing research where it is applied. Attempts were made to translate this metaphor into Polish and to explain model types from the point of view of theories of picture recognition and taxonomy. The paper also presents a short description of Data Mining instruments i.e., artificial neuron networks, classification and regression trees, induction rules, inference based on past experiences as well as genetic algorithms. To sum up the paper devotes time to the usefulness of Data Mining instruments in market and marketing researches for example in basket analysis or in market segmentation and selectivity researches.