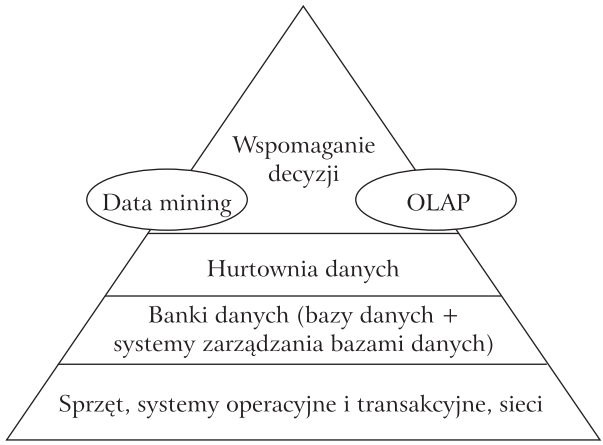
Wstęp

Praca przedstawia podstawy użytkowania programu *Enterprise Miner* firmy *SAS* oraz krótki opis metod eksploracji danych (*data mining*), których wykorzystanie umożliwia ten program. Pod pojęciem eksploracji danych rozumie się metody statystyczne i metody sztucznej inteligencji umożliwiające odkrywanie nieznanych zależności między danymi w nagromadzonych zbiorach danych [M. Lasek, 2004]. Są to metody, które pozwalają z danych tworzyć wiedzę, tzn. znajdywać zależności, wzorce, trendy „ukryte" w danych.

Rozwój metod eksploracji danych związany jest z rozwojem wielu dziedzin: informatyki, statystyki, ekonometrii, ekonomiki i organizacji przedsiębiorstw, zarządzania finansami, teorii i narzędzi wnioskowania w warunkach niepewności. To właśnie ich rozwój, a w szczególności burzliwy rozwój technologii informatycznych, spowodował, że coraz częściej punktem wyjścia w procesach decyzyjnych są dane, a nie hipotezy statystyczne lub koncepcje klasycznych modeli ekonometrycznych. Współczesny sprzęt i oprogramowanie umożliwiają gromadzenie i analizę olbrzymich zbiorów danych, obejmujących bazy mierzone dziesiątkami lub więcej gigabajtów [M. Lasek, M Pęczkowski, 2010(c)].

W komputerowych magazynach danych zwanych hurtowniami danych (*data warehouses*), tworzonych w przedsiębiorstwach przemysłowych, bankach, agencjach ubezpieczeniowych, firmach usługowych, znajdują się olbrzymie ilości danych. Dane te trudno poddają się znanym metodom analiz statystycznych i ekonometrycznych, tak aby służyło to budowaniu wiedzy, która mogłaby być przydatna do wspomagania w podejmowaniu decyzji w zarządzaniu, dokonywaniu wyborów ekonomicznych, znajdywaniu reguł i uogólnień. Narzędzia raportowania *OLAP* (***O**n**L**ine **A**nalytical **P**rocessing*), pomimo swoich niewątpliwych zalet: wielowymiarowości i wielopoziomowości (umożliwiającej

przechodzenie od ogółu do szczegółu i z powrotem) oraz pomimo analiz zapewniających wgląd w dane z różnych perspektyw badawczych, wynikających z potrzeb różnych użytkowników, okazały się nie w pełni wystarczające. Metody eksploracji danych stanowią analityczne rozszerzenie technik *OLAP* i prezentują podejście do analizy danych służące zasadniczo innym celom niż *OLAP*. Polegają one raczej nie na raportowaniu, ale na zdobywaniu nowej wiedzy. Miejsce *OLAP* i eksploracji danych w piramidzie systemów wspomagania decyzji ilustruje rys. 0.1.



Rysunek 0.1. Miejsce *OLAP* i metod eksploracji danych w piramidzie systemów wspomagania decyzji
Źródło: [Z. Chen, 2001, s. 360].

W języku polskim angielski termin *data mining methods* jest tłumaczony jako metody eksploracji danych, odkrywania wiedzy w bazach danych, zgłębiania danych, eksploatacji danych, drążenia danych. Jednak chyba najczęściej, zarówno w polskiej literaturze, na konferencjach, jak i wśród praktyków używana jest nazwa angielska: *data mining*.

Na temat metod eksploracji danych istnieje już od dosyć dawna bogata literatura (por. np. [M.J.A. Berry, G.S. Linoff, 2000; M.J.A. Berry, G.S. Linoff, 2004; P. Cabena i in., 1998; J. Han, M. Kamber, J. Pei, 2012; D.T.Larose, 2006; D.T.Larose, 2008; O. Maimon, L. Rokach (eds.), 2005; Z. Markov, D.T. Larose, 2009; R. Matignon, 2007; I.H. Witten, E. Frank, M.A. Hall, 2011; N. Ye (ed.), 2003]) dotycząca ich podstaw oraz wykorzystania w różnych zastosowaniach. Ukazują się także prace polskich autorów w języku polskim, głównie w postaci monografii, w polskich czasopismach informatycznych i w Internecie. Przygotowywane są materiały kursowe w języku polskim (por. np. [M. Pęczkowski, 2011 (b)]).

Należy podkreślić, że w Polsce prace dotyczące metod eksploracji danych są coraz intensywniej rozwijane. Są one prowadzone w licznych polskich ośrodkach. Należą do nich m.in. Szkoła Główna Han-

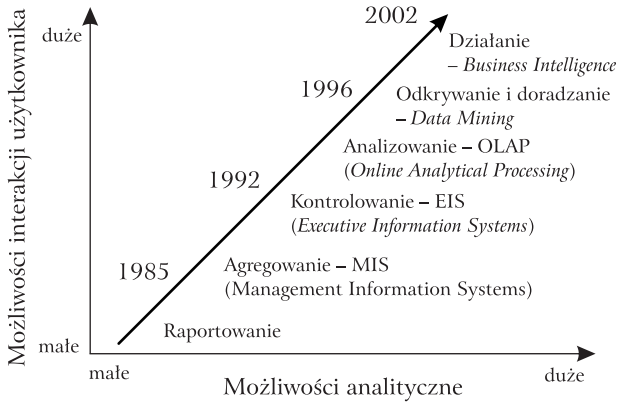
dlowa w Warszawie, Uniwersytet Gdański, Uniwersytet Ekonomiczny we Wrocławiu, Uniwersytet Ekonomiczny w Katowicach, Uniwersytet Technologiczno-Przyrodniczy w Bydgoszczy, Politechnika Białostocka, Uniwersytet w Białymstoku, Uniwersytet Ekonomiczny w Krakowie.

Nasze próby wykorzystywania metod eksploracji danych wskazały na ich dużą przydatność w przeprowadzaniu analizy danych [M. Lasek, M. Pęczkowski, 2010 (c)]. Dotyczyły m.in. analizy zróżnicowania pięciuset największych firm Rzeczpospolitej [M. Lasek, M. Pęczkowski, 2008 (b)], przeprowadzania segmentacji klientów na potrzeby prowadzenia kampanii reklamowych [M. Kutera, M. Lasek, 2010], analizowania i prognozowania kondycji ekonomicznej przedsiębiorstw [M. Lasek, 2007 (a)], analizy działalności inwestycyjnej gospodarstw agroturystycznych [M. Lasek, E. Nowak, M. Pęczkowski, 2008], analiz finansowych przedsiębiorstw [M. Lasek, M. Pęczkowski, 2008 (a)], przewidywania groźby upadłości lub konieczności prowadzenia postępowania układowego przedsiębiorstw [M. Lasek, M. Pęczkowski, D. Wierzba, 2009]. Wymienione prace prezentują wyniki prowadzonych przez nas analiz i różnorodnych badań, chociaż oczywiście w literaturze (zwłaszcza anglojęzycznej) można znaleźć wiele prac opisujących zarówno przydatność poszczególnych metod eksploracji danych do różnych celów, jak i całościowe projekty zastosowań. Często powołujemy się na własne artykuły dotyczące zagadnień eksploracji danych także z tego względu, że czytelnik może tam znaleźć wskazane przez nas dość liczne pozycje literatury odnoszące się do szczegółowych kwestii i problemów przedstawianych w oddzielnych, podejmujących odrębne tematy artykułach, jak np. wyznaczanie liczby skupień w grupowaniu obiektów, graficzna ocena jakości modeli, grupowanie zmiennych, sporządzanie prognoz. Uznaliśmy, że nie byłoby celowe powtarzanie już tam przytaczanych czy też cytowanych pozycji.

Obecnie metody eksploracji danych znajdują zastosowania praktyczne w ramach systemów oprogramowania zwanych systemami *Business Intelligence* – rys. 0.2.

Metody eksploracji danych dzielone są w różny sposób. W przypadku kryterium celu stosowania można wyróżnić metody:

- klasyfikacji – umożliwiające przydział obiektów do z góry zdefiniowanych klas (podzbiorów); należy do nich np. analiza dyskryminacyjna, w której dąży się do znalezienia funkcji umożliwiającej przewidywanie przynależności nowego obiektu do danej klasy;
- regresji – obejmujące znajdywanie związków opisujących wpływ jednej lub większej liczby cech (zmiennych objaśniających) na wybraną cechę (zmienną objaśnianą);



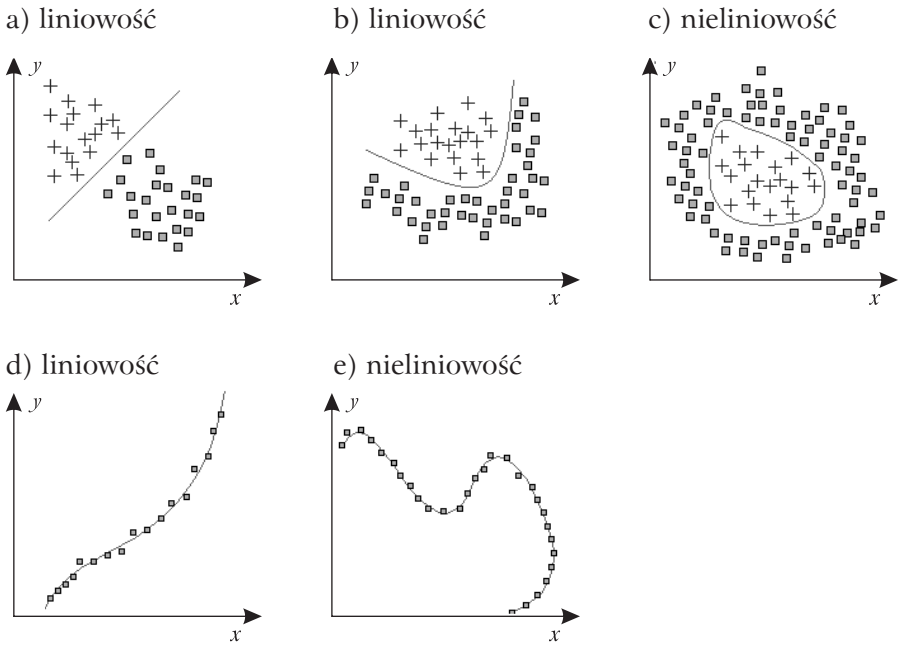
Rysunek 0.2. Ewolucja od statycznych raportów do metod eksploracji danych i systemów *Business Intelligence*
Źródło: [N. Rasmussen i in., 2002, s. 5].

- grupowania (analizy skupień) – dotyczące podziału zbioru obiektów na skończoną liczbę grup w ten sposób, aby obiekty podobne do siebie znalazły się w tej samej grupie;
- odkrywania charakterystyk – polegające na znajdywaniu opisu grup obiektów za pomocą skończonej, możliwie małej liczby cech (charakterystyk) określanych często mianem profili (np. klientów bankowych, osób kupujących określone towary, użytkowników określonych typów komputerów);
- odkrywania asocjacji – dotyczące odkrywania związków między obiektami lub grupami obiektów opisanych przez wiele cech ilościowych lub jakościowych;
- odkrywania sekwencji – służące do odnajdywania kolejności następowania zdarzeń lub pojawiania się obiektów;
- wykrywania zmian i odchyleń – polegające na poszukiwaniu wartości nietypowych (odstających, skrajnych), a także systematycznych błędów pomiaru.

Innym kryterium podziału metod eksploracji danych jest charakter związku lub podziału obiektów (cech), który może być liniowy lub nieliniowy.[1] Wyróżnia się związki liniowe oraz związki nieliniowe między obiektami, jak widać to na rys. 0.3. Przypadki *a*, *b* i *d* przedstawiają ujęcie liniowe w modelach analizy danych, chociaż linie z rys. 0.3*b* oraz 0.3*d* nie są liniami prostymi. Podział na metody liniowe i nieliniowe wynika z zależności stwarzanych przez parametry modeli (charakter dopasowania parametrów modeli).

Celem naszej pracy jest umożliwienie czytelnikowi w możliwie krótkim czasie opanowania zasad użytkowania programu *Enterprise Miner* oraz poznania podstawowych metod eksploracji danych. Praca nie

[1] Przyjęta tu terminologia; liniowy, nieliniowy, wynika z dosłownego tłumaczenia z języka angielskiego. Wydaje się, że poprawniejsze byłyby terminy: związki jednowymiarowe oraz związki wielowymiarowe.



Rysunek. 0.3. Liniowe i nieliniowe modele eksploracji danych
Źródło: na podstawie [H. Lohninger, 1999].

jest szczegółową instrukcją użytkowania programu *Enterprise Miner* ani podręcznikiem z zakresu metod eksploracji danych (takie podręczniki i materiały kursowe są opracowywane i udostępniane przez *SAS Institute*). Przedstawiamy jedynie podstawowe opcje programu oraz ogólną charakterystykę wybranych metod, zwracając uwagę na potrzeby w zakresie odpowiedniego przygotowania danych oraz na właściwą interpretację uzyskiwanych wyników. Niektóre możliwości programu, jak nam się wydaje rzadziej używane przez początkującego użytkownika, który dopiero zapoznaje się z programem i metodami, zostały jedynie zasygnalizowane – zainteresowany czytelnik może się z nimi zapoznać w dokumentacji programu dostarczanej przez firmę *SAS* lub uczestnicząc w specjalistycznych kursach organizowanych przez tę firmę. Aby umożliwić czytelnikowi wykorzystywanie tylko wybranych przez niego rozdziałów (bez potrzeby czytania całej pracy), musieliśmy powtórzyć niektóre informacje. W takich przypadkach staraliśmy się o dostosowanie opisów do potrzeb tematyki rozdziałów, w których są zamieszczane.

Możliwości programu i metod ilustrujemy przykładami analiz zbiorów danych z różnych dziedzin i źródeł:

- zbioru dotyczącego charakterystyki polskich gospodarstw domowych *F2007BD*, liczącego 37 121 obiektów (czyli gospo-

darstw domowych) oraz utworzonego na jego podstawie zbioru *F2007ZYWN*, przedstawiającego wydatki na żywność, alkohole i tytoń w gospodarstwach domowych, uwzględniającego 31 pozycji takich wydatków (zbiory danych pochodzą z badania budżetów gospodarstw domowych przeprowadzanego przez GUS i dotyczą danych z 2007 r.) [Budżety gospodarstw domowych …, 2008; Metodologia badań budżetów ..., 2011; M. Pęczkowski, 2011 (a)],

- zbioru *HMEQ* dotyczącego udzielania kredytów przez bank, liczącego 5960 obiektów (klientów banku), zbioru *HMEQ_Score* o wybieranej liczbie obserwacji (klientów lub potencjalnych klientów) dla prowadzenia na bieżąco badań skoringowych analizy ryzyka niewywiązania się klientów z płatności (badania ich ewentualnej niewypłacalności) oraz zbioru danych *BANK* o usługach świadczonych przez bank, liczącego 32 367 obserwacji reprezentujących transakcje bankowe – usługi świadczone klientom przez bank, realizowane przez około 8000 klientów banku (te dwa zbiory danych są używane na kursach prowadzonych przez *SAS Institute Inc.*, np. [Applied Analytics Using …, 2008]),
- zbioru *CHURN* z danymi dotyczącymi wykorzystywania usług przez klientów firmy telefonicznej, liczącego 3333 obiekty (klientów, którzy zrezygnowali lub nie zrezygnowali z usług firmy) oraz zbioru transakcji sprzedaży i zakupu towarów, którymi w analizowanym przypadku były rozmaite warzywa (w [D.T. Larose, 2006] podane są adresy do stron internetowych, gdzie można znaleźć zbiór *CHURN*; dane dotyczące handlu warzywami zamieszczone są bezpośrednio w tej książce).

Praca składa się z siedmiu rozdziałów.

W rozdziale 1. zapoznajemy czytelnika z podstawami posługiwania się programem *Enterprise Miner*. Opisujemy, jak rozpocząć pracę z programem, jak przygotować dane, które będą wykorzystywane w analizach, oraz jak utworzyć projekt analizy danych, który będzie złożony z diagramów definiujących poszczególne kroki składające się na analizę danych, począwszy od wprowadzenia danych i wstępnej ich analizy oraz obróbki, poprzez zastosowanie wybranych metod eksploracji danych, kończąc na interpretacji uzyskanych wyników.

W rozdziale 2. przedstawiamy metody przydatne do przeprowadzania analiz danych. Oddzielną część tego rozdziału (podrozdz. 2.2) poświęcamy metodyce analizy danych *SEMMA* (*Sample*, *Explore*, *Modify*, *Model*, *Assess*). Jest to oryginalna metodyka opracowana przez firmę *SAS Institute*, definiująca kolejne kroki i narzędzia analizy danych. Opisujemy udostępniane w *Enterprise Miner* narzędzia analizy danych, które są zalecane w ramach poszczególnych kroków metodyki. Narzędzia

te stanowią węzły diagramów analizy danych. Węzły na diagramach łączone są liniami zakończonymi strzałkami wskazującymi kolejność kroków analizy danych. W tym rozdziale omawiamy także sposoby budowania diagramu analizy danych i wskazujemy, jakich zasad budowy diagramów należy przestrzegać, aby zapewnić poprawność przeprowadzanych analiz.

Cały rozdział 3. poświęcony jest problematyce przygotowywania danych na potrzeby analizy związanej z eksploracją danych. Obejmuje on dostępne w *Enterprise Miner* metody wstępnej statystycznej analizy danych i opis narzędzi udostępnianych w celu przeprowadzenia takiej analizy. Przedstawiamy także zagadnienia losowania próby do przeprowadzania analiz, gdy uznamy, że nie ma potrzeby posłużenia się całym zbiorem. Omawiamy problem podziału danych na dane treningowe, walidacyjne i testowe (stosowanego na potrzeby budowania modeli) i przedstawiamy celowość jego wykonania. W tym rozdziale poruszamy także zagadnienia filtrowania danych, wyboru zmiennych, ze względu na które będziemy analizować dane, przeprowadzania transformacji zmiennych, zastępowania brakujących wartości wartościami wyliczanymi przez program lub przyjętymi przez użytkownika wartościami (tzw. imputacja danych) oraz wpływu wartości nietypowych (*outliers*) na wyniki analiz. Problematyce wyboru zmiennych, z uwagi na jej znaczenie w przypadkach wykorzystywania metod eksploracji danych, poświęcamy szczególną uwagę. Przedstawiamy specjalne narzędzie programu, które może być wykorzystywane, aby właściwie dobrać zmienne do tworzonego modelu – narzędzie o nazwie *Variable Selection*, które udostępnia metody doboru zmiennych. Ilustrujemy tu m.in. możliwość posłużenia się kryterium współczynnika determinacji R^2 oraz kryterium Chi^2 dla doboru zmiennych. W dalszej części pracy przedstawimy także inne metody, które mogą być pomocne w doborze zmiennych, mianowicie możliwości wykorzystywania drzew decyzyjnych oraz dokonywania selekcji zmiennych za pomocą zbudowanego przez użytkownika modelu regresji (liniowej lub logistycznej).

W rozdziale 4. omawiamy możliwości wykorzystania metod eksploracji danych na potrzeby prognozowania. Przedstawiamy zastosowanie w prognozowaniu metod: regresji liniowej i logistycznej, sieci neuronowych oraz drzew decyzyjnych. Podajemy przykłady zastosowania tych metod, posługując się programem *Enterprise Miner* i wykorzystując w możliwie szerokim zakresie domyślne ustawienia parametrów budowanych modeli regresji, sieci neuronowych i drzew decyzyjnych, tak aby umożliwić czytelnikowi opanowanie zasad budowy i wykorzystywania tych modeli. W praktycznych zastosowaniach potrzebne jest zwykle stopniowe udoskonalanie modeli metodą prób i błędów aż do dobrania takich parametrów, które pozwolą zbudować jak najlep-

szy model uwzględniający jego dostosowanie do potrzeb analizy, ale też i biorąc pod uwagę fakt, że budując model jesteśmy uzależnieni od właściwości i jakości posiadanego zbioru danych, który służy do jego utworzenia. Brak odpowiednich danych może stanowić poważną przeszkodę w zbudowaniu modelu przydatnego dla praktycznych zastosowań.

Opisujemy sposób przeprowadzania oceny poprawności prognoz oraz generowania tzw. kodu skoringowego (*score code*), który służy do przewidywania wartości cechy nowych obiektów, niewystępujących w dotychczas badanym zbiorze.

Skoring (*scoring*; *score*) jest to ocena przypisywana jakiemuś obiektowi (podobnie jak stopnie w szkole, ocena punktowa kredytobiorcy w banku lub punkty w konkurencji sportowej), która pozwala porównywać obiekty ze względu na daną cechę (zmienną wynikową). Liczbowa wartość tej oceny jest obliczana na podstawie modelu uzyskanego z treningowego zbioru danych (*trained model*). Skoring (ocenianie, punktacja) jest uogólnieniem pojęcia wartości teoretycznej zmiennej wynikowej i może być obliczany dla obiektów, dla których znamy wartości zmiennych objaśniających, ale nie znamy wartości zmiennej objaśnianej (*target*). Jest to więc przewidywana (prognozowana) wartość zmiennej objaśnianej dla tego obiektu. Skoring nowego zbioru danych (tzn. niebiorącego udziału w treningu) jest końcowym wynikiem większości problemów eksploracji danych. Na podstawie uzyskanej metody (wzoru) – tzw. *scoring formula* – obliczamy skoring (*ocenę skoringową*) nowego obiektu (np. respondenta ocenianego ze względu na prawdopodobieństwo pozytywnej odpowiedzi na ofertę marketingową).

W naszej pracy metodę zilustrujemy przykładem dotyczącym przewidywania spłaty kredytu bankowego przez klienta.

Rozdział 5. poświęcamy zastosowaniu metod eksploracji danych *i* programu *Enterprise Miner* w grupowaniu obiektów. Rozważamy zastosowanie metody grupowania należącej do niehierarchicznych metod tworzenia skupień oraz przedstawiamy możliwości tworzenia grup za pomocą sieci neuronowych Kohonena.

W rozdziale 6. opisujemy, jak przeprowadzić analizę asocjacji, jak wygenerować reguły przedstawiające zależności między zmiennymi, co umożliwia wyciąganie wniosków dotyczących współwystępowania określonych wartości zmiennych (reguły asocjacyjne) oraz następowania zjawisk w czasie (reguły sekwencji).

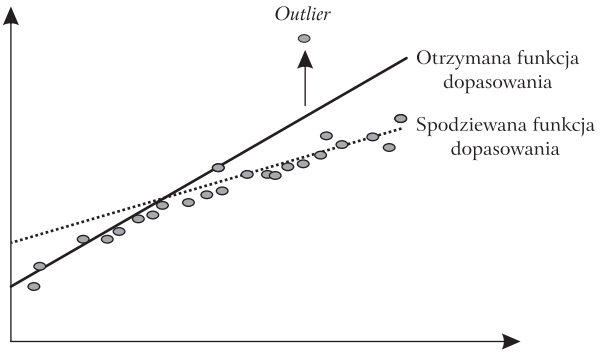
Rozdział 7. dotyczy grupowania zmiennych, które ma na celu ograniczenie redundancji informacji wprowadzanej przez zmienne, a także zmniejszenie wymiaru problemu, co ułatwia użytkowanie i interpretację modeli.

Metody eksploracji danych znajdują zastosowanie w prowadzeniu kampanii reklamowych, utrzymywaniu klientów i zdobywaniu nowych klientów, analizach i ocenach kredytobiorców bankowych oraz ocenach ryzyka kredytowego, promocji produktów i usług, prognozowaniu i planowaniu sprzedaży, przewidywaniu sukcesów kampanii reklamowych oraz sprzedaży produktów, badaniach rynku.

Należy pamiętać, że proponowane metody wymagają dogłębnego ich zrozumienia przed zastosowaniem, aby móc prawidłowo odczytać wygenerowane przez nie wyniki. Wyniki te są na ogół silnie uzależnione od danych wejściowych. Narzuca to konieczność krytycznego podejścia do uzyskanych wyników przy ich interpretacji. Potrzebne jest też nabycie umiejętności doboru proponowanych w ich ramach algorytmów i wyznaczania parametrów wejściowych (początkowych) określających sposób realizacji algorytmów specyficznych dla poszczególnych proponowanych metod, np. przyjęcia niektórych wartości początkowych lub warunków zakończenia działania algorytmów z iteracjami. Trzeba pamiętać, że wnioski w przypadku metod eksploracji danych powinny być formułowane raczej w postaci domniemań niż kategorycznych stwierdzeń.

Należy mieć także na uwadze, że metody eksploracji danych przed ich zastosowaniem wymagają wstępnej analizy danych, podczas której rozpatrywany jest charakter danych, np. występowanie wartości odstających (*outliers*) – por. rys. 0.4. O wynikach decyduje dobór cech, według których oceniane są obiekty. Aby uniknąć redundancji cech, wystarczająca okazuje się często analiza korelacji. Do niektórych analiz należy wybierać najsłabiej skorelowane zmienne. W wielu przypadkach należy dążyć do ograniczenia liczby cech, gdy ich liczba jest zbyt duża z uwagi na potrzeby analizy, charakter zastosowanej metody lub ilość dostępnych danych (obiektów ze znaną charakterystyką).

Wymogiem użycia określonej metody eksploracji danych może być posiadanie danych dla odpowiednio dużej liczby obiektów w stosunku



Rysunek 0.4. Efekt wpływu wartości odstających (*outliers*)

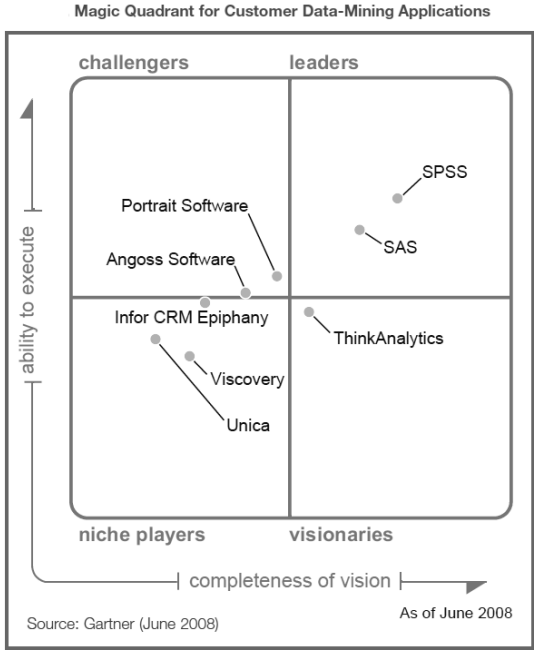
Źródło: na podstawie [A.P. Engelbrecht, 2002, s. 92].

do liczby cech. Możemy też po prostu uznać, że liczba cech opisujących dane dla obiektów jest zbyt duża (wręcz zastraszająco duża, np. wynosi kilkaset), co utrudni zbudowanie właściwego modelu predykcyjnego. Istnieją algorytmy doboru cech do analizy, które ograniczają liczbę zmiennych. Możliwe są dwie strategie postępowania przy stosowaniu tych algorytmów:

(1) rozpoczyna się od małej liczby cech, po czym są dodawane kolejne cechy, aż do momentu uzyskania pożądanej jakości modelu (określane mianem algorytmów doboru cech w przód);
(2) rozpoczyna się od zestawu wszystkich cech i kolejno eliminuje cechy, które mają najmniejszy wkład w uzyskanej jakości modelu, tj. z których rezygnacja w jak najmniejszym stopniu pogorszy jakość modelu (określane mianem algorytmów eliminacji wstecznej cech).

Kombinacją wymienionych powyżej strategii jest tzw. metoda krokowa, którą należy tu wymienić dla kompletności algorytmów i o której będzie jeszcze mowa w dalszej części tej pracy.

Zachętą do zapoznania się z programem *Enterprise Miner* firmy *SAS* może być wysoka pozycja, jaką zajmuje ta firma na rynku producentów narzędzi eksploracji danych. Przykładem potwierdzającym tę tezę są wyniki badań firmy *Gartner* prezentowane w Internecie w serii słynnych „magicznych kwadratów" (*magic quadrant*), co np. ilustruje rys. 0.5, ukazujący, że *SAS* znajduje się wśród liderów aplikacji do eksploracji danych wykorzystywanych do analizy danych o klientach.



Rysunek 0.5 Miejsce firmy *SAS* wśród dostawców narzędzi eksploracji danych

Źródło: wyniki badań Gartner, Inc., 2008 opublikowane w Internecie (*Gartner RAS Core Research Note G00158953*, podano za: http://bi.pl/publications/art, dostęp w dniu 18.04.2011 lub też http://www.spss.com.hk/PDFs/Gartner_Magic_Quadrant, dostęp w dniu 7.11.2011).