POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

# ROZWÓJ I ZASTOSOWANIA TECHNOLOGII I SYSTEMÓW INFORMATYCZNYCH

pod redakcją:
Jana Studzińskiego
Ludosława Drelichowskiego
Olgierda Hryniewicza

ROZWÓJ I ZASTOSOWANIA TECHNOLOGII
I SYSTEMÓW INFORMATYCZNYCH

Polska Akademia Nauk • Instytut Badań Systemowych

# Seria: BADANIA SYSTEMOWE
tom 28

**Redaktor naukowy:**

**Prof. dr hab. Jakub Gutenbaum**

Warszawa 2001

# ROZWÓJ I ZASTOSOWANIA TECHNOLOGII I SYSTEMÓW INFORMATYCZNYCH

pod redakcją
Jana Studzińskiego, Ludosława Drelichowskiego
i Olgierda Hryniewicza

Wydano z wykorzystaniem dotacji KOMITETU BADAŃ NAUKOWYCH

Książka zawiera wybór artykułów poświęconych omówieniu aktualnego stanu badań w kraju w zakresie rozwoju technologii, modeli i systemów informatycznych oraz ich zastosowań w różnych dziedzinach gospodarki narodowej. Wyodrębnioną grupę stanowią artykuły aplikacyjne omawiające wyniki projektów badawczych i celowych KBN.

Recenzenci artykułów:

Dr hab. inż. Ryszard Budziński, prof. US
Prof. dr hab. inż. Janusz Kacprzyk
Dr hab. Adam Kopiński, prof. AE we Wrocławiu
Doc dr hab. inż. Marek Libura
Prof. dr hab. inż. Andrzej Straszak



ISBN 83-85847-59-6
ISSN 0208-8028

Rozdział 3

# Metody i algorytmy obliczeniowe w systemach informatycznych

# ANALIZA NARZĘDZI AUTOMATYCZNEGO WYSZUKIWANIA DANYCH (DATA-MINING)

***Witold Chmielarz***
*Wydział Zarządzania Uniwersytetu Warszawskiego*

*Zasadniczym celem niniejszego referatu jest analiza możliwości zastosowań narzędzi Data-Mining dla celów zarządzania organizacjami oraz ich analiza porównawcza. Podstawowymi przesłankami wykorzystania tej techniki jest masowość danych, gromadzonych w hurtowniach danych oraz lawinowo rosnąca popularność Internetu i związany z tym dostęp do niezwykle bogatych zasobów informacji. Na początku artykułu zdefiniowano podstawowe pojęcia związane z Data-Mining i ich typologie. Następnie zanalizowano trzy wybrane (IBM Intelligent Miner for Data, SAS Enterprise Miner, SGL Mine Set), najbardziej popularne, komercyjne pakiety Data-Mining wraz z przykładowymi zakresami zastosowań oraz skrótową analizą porównawczą tych pakietów.*

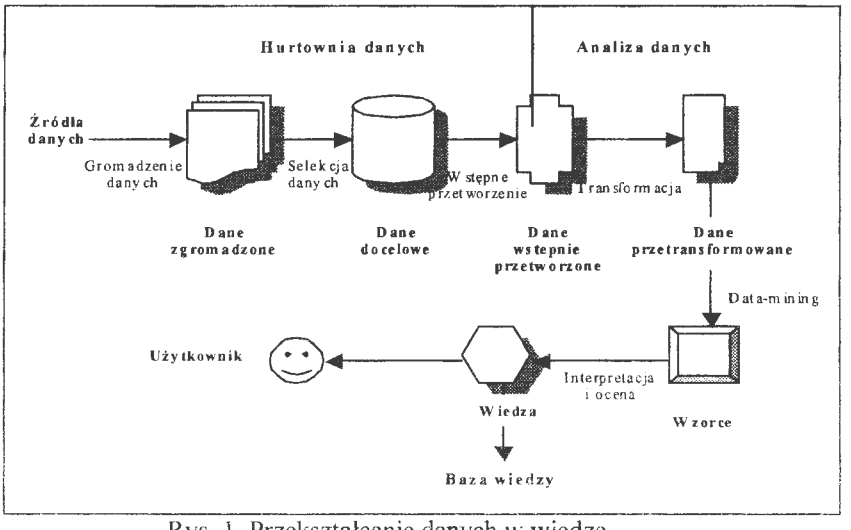
## 1. Wprowadzenie

Gromadzenie danych dla zarządzania jest jednym z najtrudniejszych problemów przed jakimi staje obecnie organizacja. Dzieje się tak z powodów następujących [Chmielarz00a]:

- ilość danych rośnie obecnie wykładniczo, wiele z nich musi być przechowywanych przez długi czas, ciągle i w sposób masowy są dodawane nowe dane,
- tylko mała część z nich jest używana dla podejmowania określonych decyzji,
- istotny wpływ na decyzje zewnętrzne ma coraz większa ilość informacji napływających z zewnątrz.
- niezbędne dla podejmowania decyzji dane mogą być gromadzone w różnych systemach komputerowych, bazach danych, formatach i językach zarówno programowania jak i ludzkich,
- wymagania prawne związane z gromadzeniem danych są różne w różnych krajach i bardzo często ulegają zmianie,
- występuje ogromna ilość narzędzi wspomagających selekcję danych dla celów zarządzania,
- bezpieczeństwo, jakość i integralność danych w systemie stanowią krytyczny czynnik sukcesu implementacji systemów w rzeczywistości gospodarczej.

Transformacja danych w wiedzę – w istocie potrzebna do podejmowania decyzji – może się odbywać bardzo różnymi drogami. Uogólniony sposób transformacji pokazuje Rys.1. Początkowo dane są gromadzone w bazie danych. Następnie po wstępnym przetworzeniu umieszcza się je w hurtowniach danych. W celu umożliwienia wykorzystania wiedzy w nich zawartej do celów zarządzania dane te przechodzą przez proces transformacji, przygotowujący je do analizy szczegółowej. Analiza ta dokonywana jest narzędziami automatycznego wyszukiwania danych *(Data-Mining)*. Ostatecznym krokiem przetworzenia jest porównanie wyszukanych

danych ze wzorcami (zachowań, reakcji) przechowywanymi w inteligentnych systemach, pozwalających na interpretacje uzyskanych porównań. Rezultatem ostatecznym tych porównań jest uzyskanie oceny przydatności uogólnionej informacji dla celów zarządzania oraz jej zgromadzenie, razem z danymi w bazie wiedzy.




Rys. 1. Przekształcanie danych w wiedzę
Źródło: opracowanie własne

Proces wydobywania użytecznej wiedzy z danych masowych nazywa się zarządzaniem wiedzą. Zarządzanie wiedzą to efektywne wykorzystanie przez użytkownika mechanizmów manipulacji informacją w celu usprawnienia procesów kierowania organizacją. W swojej historii przechodził on cztery podstawowe stadia, ukazane w Tab.1.

Tab.1. Etapy rozwoju zarządzania wiedzą

| Stopień rozwoju | Dostępne technologie | Charakterystyka danych |
|---|---|---|
| Gromadzenie danych – lata sześćdziesiąte | Komputery, pamięci taśmowe i dyskowe | Retrospektywne, statyczne pozyskiwanie danych |
| Uzyskanie dostępu do danych – lata osiemdziesiąte | Relacyjne bazy danych (RDBMS), Strukturalne języki zapytań (SQL) | Retrospektywne, dynamiczne pozyskiwanie danych na poziomie rekordu |
| Hurtownie danych i systemy wspomagania decyzji – wczesne lata dziewięćdziesiąte | Techniki przetwarzania analitycznego w czasie rzeczywistym – on-line analytic proccsing, wielowymiarowe bazy danych, hurtownie danych | Retrospektywne, dynamiczne pozyskiwanie danych na wielu poziomach |
| Inteligentne, automatyczne wyszukiwanie danych (intelligent data-mining) | Zaawansowane algorytmy, komputery wieloprocesorowe, masowe bazy danych | Dostarczanie aktywnych informacji prospektywnych |

Źródło: opracowanie własne na podstawie [Turban98]

Najbardziej efektywnym narzędziem zarządzania wiedzą jest automatyczne wyszukiwanie informacji *(data-mining). Data-Mining* jest to proces automatycznej ekstrakcji użytecznej, wartościowej i uprzednio nieznanej wiedzy z dużych baz danych. Ujawnione w ten sposób ukryte trendy, korelacje i wzorce w danych wspomagać mogą procesy decyzyjne w przedsiębiorstwie.

Podstawa zarządzania wiedzą w bazach danych wymaga gromadzenia masowych danych, potężnych mocy przetworzeniowych i skutecznych algorytmów wyszukiwawczych. Data-mining polega głównie na automatycznym przewidywaniu trendów na podstawie danych uzyskanych z baz przemysłowych oraz automatycznym odkrywaniu nieznanych uprzednio wzorców zależności i zachowań. Spowodowane jest to faktem, że w wielkich bazach danych niezbędne dla zarządzania dane są głęboko ukryte, oraz faktem, że dane dotyczące organizacji mogą być konsolidowane lub trzymane na serwerach intranetowych lub internetowych. Skutkiem działania „poszukiwacza danych" – użytkownika końcowego - jest wyabstrahowanie niezbędnej informacji z baz danych.

Data-mining powoduje powstanie pięciu typów informacji wynikających z relacji pomiędzy danymi uzyskanymi z bazy danych: asocjacji (skojarzenia), sekwencji (kolejności), klasyfikacji (segregowanie), klasteryzacji (gromadzenie, grupowanie), prognozowania (przewidywanie). Opiera się na technikach wnioskowania na podstawie przypadku, badania sieci neuronowych, algorytmów genetycznych, kreatorów *(Intelligent Agents)*, drzew decyzyjnych, metod najbliższego sąsiedztwa itp. Elementy heurystyki – oparte często na analizie częstości korzystania z informacji – przyspieszają procesy wyszukiwawcze.

## 2. Przegląd narzędzi Data-Mining

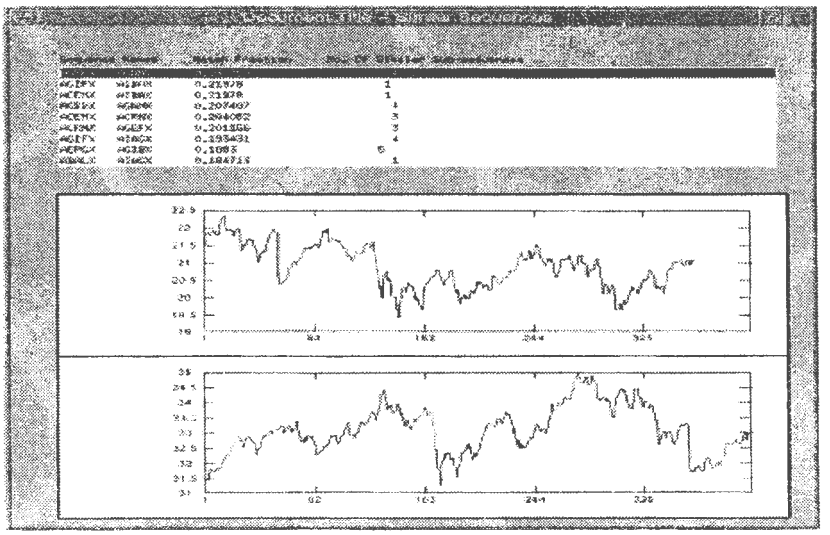
Obecnie na rynku znajduje się ogromna ilość pakietów oprogramowania, które można określić mianem narzędzi Data-Mining. Poniżej przedstawiono w skrócie możliwości trzech z nich, uznawane często za najbardziej uniwersalne [Westphal99].

### *2.1. IBM Intelligent Miner for Data*

Intelligent Miner for Data jest produktem firmy IBM będącym częścią „platformy Business Intelligence" w skład, której wchodzą także systemy zarządzania dokumentacją i analizy lingwistycznej (Intelligent Miner for Text), systemy CRM, a także uniwersalna baza danych DB2 [Gawrysiak01]. W przeciwieństwie do pozostałych omówionych w tej pracy pakietów, produkt IBM nie posiada właściwie żadnych możliwości wstępnej obróbki danych., co wynika to ze ścisłej integracji z bazą danych DB/2, na podstawie której dokonywany jest preprocesing analizowanych zbiorów danych [IBM2000]. Interfejs użytkownika udostępniany przez pakiet jest prosty, lecz dość funkcjonalny. Większość operacji może być wspomagana przez *kreatory* ułatwiające ustalanie parametrów poszczególnych algorytmów analizy danych [Chmielarz00]. Prezentacja wyników dokonywana jest w formie graficznej, głównie w postaci histogramów i wykresów kołowych. Intelligent Miner for Data udostępnia użytkownikom następujące narzędzia Data-Mining:

- Reguły asocjacyjne – algorytm zastosowany w pakiecie jest bardzo efektywny, co zostało to osiągnięte kosztem znacznego ograniczenia jego uniwersalności. Możliwe jest, bowiem wykrywanie jedynie reguł asocjacyjnych jednowartościowych (tj. takich gdzie prawa i lewa strona reguły są zbiorem jednoelementowym), dodatkowo zaś baza danych przeznaczona do analizy tym algorytmem musi mieć specyficzną postać, dostosowaną do zastosowań związanych ze sprzedażą detaliczną towarów.
- Grupowanie – Intelligent Miner wykorzystuje niestandardowe algorytmy grupowania, opracowane i opatentowane przez ośrodek badawczy IBM Almaden Research Center. Wyniki grupowania przedstawiane są graficznie, przy czym dla każdej zmiennej podawany jest zarówno jej rozkład w całej populacji jak i w danej grupie.
- Klasyfikacja – pakiet pozwala na klasyfikację danych przy użyciu dwóch metod – drzew decyzyjnych, oraz sieci neuronowych. Zastosowano proste algorytmy – użytkownik dostarcza jedynie do nich dane wejściowe. Ich niewątpliwą zaletą jest natomiast duża – w porównaniu z pozostałymi opisywanymi pakietami – szybkość działania. Sieć neuronowa w wyniku swojego działania daje model, który możemy następnie użyć do klasyfikacji elementów zbioru danych nie wchodzących w skład zbioru trenującego, drzewo decyzyjne może być zaś dodatkowo przedstawione w formie graficznej.
- Analiza sekwencji czasowych – jest unikalną cechą pakietu. W jej skład wchodzą klasyczne metody statystyczne (m.in. regresja liniowa) wykorzystywane do prognozowania przebiegów czasowych, oraz narzędzie *Similar Time Sequences* pozwalające na odkrywanie powtarzających się, charakterystycznych przebiegów w dużych zbiorach danych. Na poniższym rysunku przedstawiono wyniki działania tego narzędzia, kolorem czerwonym oznaczone zostały te okresy, w których dany atrybut charakteryzował się bardzo zbliżoną zmiennością.

W zamierzeniach swych projektantów pakiet Intelligent Miner for Data nie powinien być raczej używany jako samodzielne narzędzie analizy danych, lecz raczej jako podstawa do budowy dedykowanych systemów korporacyjnych, zorientowanych na obsługę konkretnego zadania biznesowego. Dlatego też pakiet ten wyposażono w bardzo zaawansowany interfejs programowania API, pozwalający na wykorzystanie wszystkich narzędzi analitycznych w programach tworzonych przez użytkownika.



Rys. 2. Automatyczne wykrywanie podobnych sekwencji czasowych w pakiecie Intelligent Miner for Data. Źródło: opracowanie własne

### *2.2. SGI Mine Set*

Mine Set został opracowany w kalifornijskiej firmie Silicon Graphics, znanej przede wszystkim z produkcji sprzętu i oprogramowania wykorzystywanego do prac graficznych. Z tego też powodu jedną z najbardziej istotnych cech pakietu Mine Set jest bardzo rozbudowany moduł wizualizacji trójwymiarowej danych. Nie jest to jednak jego jedyna funkcja, w skład pakietu wchodzą wszystkie istotne narzędzia, jakie mogą być przydatne w procesie obróbki i analizy danych np.:

- Moduł generacji reguł asocjacyjnych,
- Moduł budowy klasyfikatorów w oparciu o algorytmy drzew decyzyjnych i regresji,
- Automatyczna dyskretyzacja danych ciągłych,
- Moduł grupowania działający w oparciu o algorytm K-means,
- Import danych ze źródeł lokalnych (pliki tekstowe, arkusze MS Excell, pliki CSV),
- Import danych z baz danych poprzez interfejs ODBC,
- Narzędzia wstępnej obróbki danych,
- Narzędzia wizualizacyjne.

Wszystkie funkcje pakietu MineSet wywoływane są z narzędzia *ToolManager*, którego interfejs przedstawiono na poniższej ilustracji:

Rys. 3. Interfejs programu SGI Mine Set uruchomionego w systemie Unix.
Źródło: opracowanie własne

Z poziomu tego interfejsu dokonujemy uruchamiania pozostałych modułów programu dokonujących bądź to analizy danych, bądź też ich wizualizacji. Sam ToolManager jest jednak także przydatnym narzędziem, ponieważ wyposażono go w zaawansowane mechanizmy wstępnej obróbki i importu danych. Funkcje wstępnej obróbki danych są dość imponujące – możliwe jest obliczanie wielu funkcji matematycznych, dokonywanie automatycznej dyskretyzacji, wybór próbek danych, filtrowanie i tak dalej. Jest to istotne, gdyż w przypadku innych pakietów te operacje muszą być wykonane w programie zewnętrznym, jakim jest zwykle arkusz kalkulacyjny, a który może nie działać odpowiednio efektywnie z bardzo dużymi zbiorami danych. Uzupełnieniem wspomnianych narzędzi jest przeglądarka danych, mogąca efektywnie wyświetlać (w postaci tabelarycznej), filtrować i przeszukiwać zbiory danych o praktycznie nieograniczonej objętości.

Mine Set udostępnia następujące metody Data-Mining:

- Grupowanie – dysponujemy tu algorytmem *K-means (oraz iterative K-means),*
- Kategoryzacja – dostępna jest metoda wykorzystująca algorytm ID3 oraz jego ulepszoną wersję ID4.5. Narzędzia kategoryzacji Mine Set mogą także dokonać budowy modelu predykcyjnego wykorzystując metodę regresji wielorakiej.
- Generacja reguł asocjacyjnych – możliwa jest generacja klasycznych reguł asocjacyjnych. Możliwe jest także ograniczenie postaci generowanych reguł tak, aby po lewej i prawej stronie reguły zawsze znajdował się tylko jeden atrybut (tzw. reguły jednowartościowe). Gdy zaś generowane są reguły bez

tych ograniczeń, do ich prezentacji wykorzystywana jest klasyczna postać tabelaryczna.

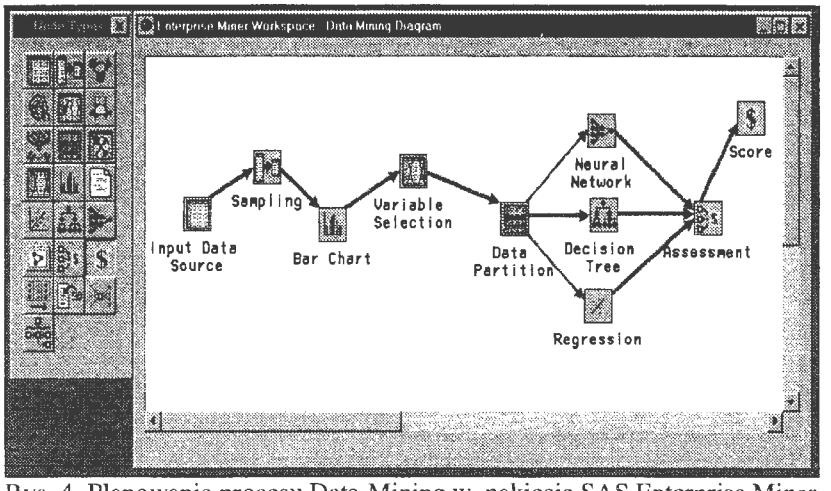
Możliwości wizualizacji danych oferowane przez Mine Set są bardzo duże. W szczególności wszystkie narzędzia wizualizacyjne generują obrazy trójwymiarowe. Dzięki temu znacznie zwiększa się przejrzystość wszelkiego rodzaju wykresów i diagramów, jednocześnie zwiększa wymagania sprzętowe pakietu. Dostępne funkcje wizualizacyjne są następujące:

- Tree Visualizer – pozwala na zobrazowanie drzew decyzyjnych. Liście drzewa przedstawiane są jako obiekty na płaszczyźnie, z każdym z liści związany jest histogram obrazujący rozkład części zbioru danych związanej z danym liściem.
- Rules Visualizer – tworzy opisane powyżej wizualizacje wyniku odkrywania reguł asocjacyjnych.
- Scatter Visualizer – pozwala na tworzenie wykresów wielowymiarowych. Dostepne są trzy wymiary przestrzenne, kolor obiektów, ich wielkość i typ – ich przykładowe wykorzystanie zaprezentowano na poniższym zdjęciu:
- Map Visualizer – pozwala na powiązanie danych z informacjami geograficznymi i tworzenie trójwymiarowych map terenu z nałożonymi obiektami reprezentującymi dane z analizowanego zbioru.

### *2.3. SAS Enterprise Miner*

SAS Institute znany jest głównie dzięki swojemu pakietowi analizy statystycznej SAS. Enterprise Miner jest dodatkiem do tego pakietu, wzbogacającym go o metody Data-Mining. Firma SAS Institute opracowała metodologię Data-Mining zwaną SEMMA. Zgodnie z jej założeniami zanim rozpoczniemy analizę danych powinniśmy zaprojektować – przynajmniej w ogólnym zarysie – przebieg procesu analizy i określić z jakich narzędzi będziemy korzystać. Do tego celu służy narzędzie o nazwie Data-Mining Workspace, gdzie w sposób graficzny budujemy sieć działań, wiążąc moduły analityczne pakietu Enterprise Miner ze źródłami danych.

Poniższa ilustracja pokazuje przykładową sieć działań. Jak widać początkowy zbiór danych poddany zostanie najpierw próbkowaniu, by ograniczyć jego objętość (*Sampling*). Następnie dokonana zostanie wizualizacja pobranej próbki w postaci histogramu (*Bar Chart*) na podstawie której użytkownik wybierze istotne dla analizy zmienne (*Variable Selection*). Następnie próbka zostanie podzielona na trzy części (*Data Partition*), z których każda zostanie poddana analizie przy użyciu innego narzędzia. Zbudowane zostaną modele wykorzystujące sieć neuronową (*Neural Network*), analizę regresji (*Regression*) oraz drzewo decyzyjne (*Decision Tree*). Na koniec modele te poddane zostają porównaniu i ocenie (*Assessment* i *Score*).



Rys. 4. Planowanie procesu Data-Mining w pakiecie SAS Enterprise Miner. Źródło: [SAS99]

Ponieważ Enterprise Miner jest integralną częścią pakietu SAS, toteż podczas analizy danych można korzystać ze wszystkich dostępnych w nim narzędzi. Odnosi się w szczególności do wstępnej obróbki danych. Każdy program napisany w języku *SAS code* może zostać umieszczony na diagramie procesu Data-Mining i uruchomiony. Dzięki temu do naszej dyspozycji staje bogaty zbiór procedur i algorytmów statystycznych z których możemy skorzystać podczas czyszczenia, filtrowania i próbkowania wyjściowego zbioru danych. Ponieważ pakiet SAS może odczytywać dane z własnych plików o niestandardowym formacie (tzw. *SAS library*) oraz z plików tekstowych i baz danych, toteż te same możliwości udostępnia Enterprise Miner.

Spośród trzech opisanych w niniejszej pracy pakietów SAS Enterprise Miner dysponuje niewątpliwie najbardziej rozbudowanymi narzędziami analizy danych. Niektóre narzędzia (np. kategoryzacja, regresja) generują w wyniku swego działania model matematyczny, który obrazuje zależności istniejące w zbiorze wejściowym. Modele te mogą być automatycznie porównywane przez system w celu wyboru tego, który najlepiej opisuje charakterystykę danych wyjściowych. Algorytmy Data-Mining dostępne w pakiecie są następujące:

- Reguły asocjacyjne – Enterprise Miner udostępnia jedynie algorytmy odkrywania reguł jednowartościowych. Dodatkowo jednak pozwala na analizę czasową, pod warunkiem, iż analizowany zbiór danych zawiera informacje o czasie dokonywanych transakcji. W ten sposób możliwe jest odkrywanie zależności nie tylko pomiędzy towarami zakupywanymi jednocześnie przez klientów, ale także takimi które kupowane są w podobnych odstępach czasu.
- Kategoryzacja – dostępne są tu dwie metody: drzewa decyzyjne oraz sieci neuronowe. Algorytm tworzący drzewa decyzyjne oparto na standardowej metodzie ID4.5, natomiast w ciekawy sposób rozwiązano wizualizację wyników jego działania. Zbudowane drzewo przedstawione jest jako koło po-

dzielone na pierścienie, każdy z pierścieni reprezentuje jeden poziom drzewa, ich wycinki zaś odpowiadają poszczególnym węzłom. Poza budowaniem drzew decyzyjnych możemy także tworzyć modele wykorzystujące sieci neuronowe. Dostępnych jest kilka typów sieci, od prostych struktur jednowarstwowych, aż do perceptronów wielowarstwowych z warstwami ukrytymi.

- Grupowanie – Enterprise Miner pozwala na dokonanie automatycznego grupowania zbioru danych. Dostępny jest algorytm grupowania hierarchicznego, przy czym użyto w nim klasycznej miary euklidesowej. Użytkownik może dokonać wstępnego ograniczenia maksymalnej liczby wygenerowanych grup, a także maksymalnej i minimalnej dopuszczalnej liczby elementów w jednej grupie.
- Regresja i inne metody statystyczne – bezpośrednio z poziomu interfejsu Enterprise Minera dostępna jest jedynie regresja liniowa. Można jednak w prosty sposób odwołać się do dowolnej funkcji statystycznej pakietu SAS.

Przedsatwione trzy pakiety są obecnie najczęściej stosowanymi uniwersalnymi narzędziami Data-Mining. IBM Intelligent Miner for Data wydaje się być stosunkowo najsłabszym systemem, przynajmniej jeśli weźmiemy pod uwagę jakość i elastyczność narzędzi analizy danych. Jego interfejs API i ścisła integracja z dość popularną bazą DB/2 powodują natomiast, iż stanowi on dobre rozwiązane korporacyjne, mogące stać się podstawą do tworzenia bardziej zaawansowanych systemów, tym bardziej iż charakteryzuje się wysoką wydajnością. Enterprise Miner posiada niewątpliwie największe możliwości analizy danych, tym bardziej iż pozwala na stosowanie wszystkich metod programu SAS. Jest on jednak bardzo skomplikowany w obsłudze, zaś komponenty wizualizacji danych są niezbyt zaawansowane. Wydaje się zatem, iż będzie on dobrym rozwiązaniem dla organizacji, które już wykorzystują w swej pracy system SAS. SGI Mine Set pozwala na bardzo efektowną wizualizację danych oraz jest bardzo łatwy w obsłudze, jednak efektywność wbudowanych algorytmów Data-Mining jest niezbyt duża. To, oraz niewielkie wymagania sprzętowe predestynują ten pakiet do roli osobistego narzędzia DataMining.

Poniższe tabele podają najważniejsze parametry opisanych powyżej pakietów oraz zawierają ich analizę porównawczą.

Tabl.2. Parametry pakietów Data-Mining.

| Nazwa pakietu | Systemy operacyjne | Źródła importu danych (pliki) | Dostęp do baz danych | Interfejs użytkownika |
|---|---|---|---|---|
| IBM Intelligent Miner for Data | Win32, Unix, Linux, OS/390 | Pliki tekstowe | DB/2 | Graficzny, API C++ |
| SGI Mine Set | Win32, Unix, | Pliki tekstowe, MS Excell, CSV | ODBC | Graficzny, brak API |
| SAS Enterprise Miner | Win32, Unix | Pliki tekstowe, pliki SAS, CSV | ODBC | Graficzny, niestandardowe API |

Źródło [Goebell99]

Tab. 3. Analiza porównawcza algorytmów stosowane w pakietach Data-Mining.

| Nazwa pakietu | Metody analizy danych | | | | | |
|---|---|---|---|---|---|---|
| | Preprocesing | Reguły asocjacyjne | Grupowanie | Sekwencje czasowe | Kategoryzacja | Wizualizacja |
| IBM Intelligent Miner for Data | Brak | Tylko jednowartościowe | Tak | Tak | Tak | Nie |
| SGI Mine Set | Tak | Tak | Tak | Nie | Tak | Tak, także 3D |
| SAS Enterprise Miner | Przy wykorzystaniu pakietu SAS | Tylko jednowartościowe | Tak | Tak | Tak | Tak |

Źródło: opracowanie własne

## 3. Zastosowania metodologii data-mining

Pomimo że metodologia Data-Mining jest bardzo młoda, to obszar jej zastosowań praktycznych jest już dość obszerny. Wynika to między innymi z szybko wzrastającej popularności systemów hurtowni danych oraz gwałtownego rozwoju handlu elektronicznego, tworząc bardzo duży popyt na wszelkiego rodzaju techniki analizy dużych zbiorów danych.

*Segmentacja rynku*

Techniką Data-Mining, która może zostać bezpośrednio zastosowana w segmentacji rynku jest grupowanie. Dzięki niemu, posiadając np. bazę adresową naszych klientów liczącą nawet kilkaset tysięcy rekordów, możemy w krótkim czasie zbadać, czy występują w niej grupy o charakterystycznych wartościach pewnych atrybutów. O ile dotychczas zwykle największym problemem było samo dokonanie segmentacji – to jest wydzielenie grup – to przy użyciu narzędzi Data-Mining najbardziej skomplikowana staje się interpretacja dokonanego podziału. Dużą pomocą w prowadzeniu segmentacji, czy też bardziej ogólnie – w analizie sytuacji rynkowej przedsiębiorstwa – są też narzędzia wizualizacji danych, dzięki którym szybko można odkryć np. klientów nietypowych, o wyraźnie odmiennych wartościach atrybutów, którzy mogą wymagać specjalnego traktowania. Jak widać te zastosowania ściśle wiążą się także z inną techniką szybko zyskującą ostatnio popularność, tj. z CRM.

*Przewidywanie zachowania klientów*

Obszarem zastosowań ściśle powiązanym z tym opisanym powyżej jest analiza zachowania rynku, co w większości przypadków sprowadza się do analizy zachowania klientów. Jak łatwo się domyśleć podstawowym narzędziem stosowanym tutaj są reguły asocjacyjne, pomagające w projektowaniu wszelkiego rodzaju pakietów typu *bundle* i analizie wyników sprzedaży. Równie istotna jest tu także kategoryzacja. Posiadając odpowiednio bogatą bazę informacji o dotychczasowych klientach i ich zachowaniu, firmy mogą próbować budować modele pozwalające

przewidzieć, to co czynić będą nowi klienci. Można także badać wzajemne powiązania pomiędzy poszczególnymi atrybutami opisującymi charakterystykę klientów.

*Wykrywanie oszustw*

Postępująca automatyzacja usług bankowych i telekomunikacyjnych wprowadza coraz większe możliwości dokonywania oszustw finansowych (np. pranie pieniędzy) czy też przestępstw teleinformatycznych (np. włamania do systemów komputerowych). Dzięki możliwości błyskawicznego dokonywania przelewów bankowych operacje takie jak np. „rozpłynięcie" gotówki z jednego konta na wiele innych kont, które dawniej wymagały dość żmudnej pracy i pomocy ze strony osoby zatrudnionej w instytucji finansowej, dziś mogą być dokonane w przeciągu kilku sekund. W wykrywaniu i śledzeniu większości nielegalnych operacji finansowych pomocne są głównie narzędzia wizualizacji danych, w tym szczególnie narzędzia analizy grafowej. Nie oznacza to oczywiście iż pozostałe metody Data-Mining nie mogą być tu użyteczne – wszystko zależy oczywiście od konkretnego problemu. Dla przykładu znane są przypadki stosowania algorytmów odkrywania reguł asocjacyjnych do wykrywania „podejrzanych" sekwencji transakcji finansowych [Wesphal99]. Z kolei w przypadku instytucji telekomunikacyjnych Data-Mining jest coraz częściej wykorzystywany w walce ze zjawiskiem zwanym *churning*. Mianem tym określa się "ucieczkę" klientów od jednego operatora do innych, w wyniku np. „zwabienia" ich szczególnie atrakcyjną promocją. Środkiem zaradczym może być tu zastosowanie metod Data-Mining, takich jak kategoryzacja - do zidentyfikowania potencjalnych ofiar *churningu* wśród własnych klientów, a następnie zastosowanie odpowiednich środków zaradczych takich jak selektywne obniżki cen, budowanie zaufania poprzez wysyłkę personalizowanej korespondencji itp. (patrz np. [Mattison97]).

*Inne zastosowania*

Data-Mining jest niezwykle szybko rozwijającą się metodologią i dlatego też nie sposób wymienić jej wszystkich potencjalnych obszarów zastosowań. Obszarem, w którym z pewnością będzie wykorzystywana w coraz większym stopniu jest obróbka naukowych danych eksperymentalnych, szczególnie w naukach biologicznych. Ostatnie osiągnięcia w biotechnologii umożliwiają bezpośrednie stosowanie informacji genetycznej w badaniach, ponieważ jednak DNA jest strukturą bardzo skomplikowaną, toteż wszelkie automatyczne techniki analizy danych będą tu bardzo pomocne. Pakiety Data-Mining są zresztą już stosowane podczas prowadzenia badań nad mapowaniem ludzkiego genomu, przykładem może być tu choćby kalifornijska firma Incyte, która opracowała narzędzie analizy sekwencji DNA przy wykorzystaniu pakietu SGI Mine Set.

Bardzo obiecującą dziedziną – nie tylko z punktu widzenia Data-Mining, ale także sztucznej inteligencji, jest zarządzanie portfelami inwestycyjnymi. Oczywiście ponieważ umiejętność skutecznej „gry na giełdzie" wiąże się z możliwością osiągania znacznych zysków, toteż problem ten analizowany jest już bardzo długo. Dopiero jednak stosunkowo niedawno pojawiły się nowe rozwiązania, takie jak algorytmy genetyczne.

## 4. Podsumowanie

Metodologia Data-Mining, pomimo,że jest bardzo młoda, jest dziedziną posiadającą ogromną dynamikę rozwoju. Praca ta żadną miarą nie dostarcza pełnego obrazu tej dziedziny, wydaje się jednak, iż pozwala na szybkie zapoznanie się z jej podstawowymi pojęciami i technikami oraz porównanie jej najczęściej używanych narzędzi. Wraz z rozwojem informatyki, komputery oraz ich oprogramowanie przestają służyć jedynie do obliczeń, zaczynają zaś zwiększać nasze możliwości poznania świata. Data-Mining jest metodologią, dzięki której możliwe jest zauważenie zależności, które normalnie umykają naszym – bardzo przecież ograniczonym zmysłom. Trudno w tej chwili przewidzieć jak rozwiną się techniki tej metodologii, związanej z procesami uczenia się maszyn i sztucznej inteligencji, ale jej rozwój na pewno warto śledzić.

## Literatura


[Chmielarz00a] ChmielarzW.: "Rola tendencji integracyjnych w kształtowaniu systemów informatycznych zarządzania", rozdział III książki pod red. Naukową T. Kasprzaka "Integracja i architektury systemów informacyjnych przedsiębiorstw", Katedra Informatyki Gospodarczej i Analiz Ekonomicznych Wydziału Nauk Ekonomicznych UW, Warszawa, 2000,

[Chmielarz00b] Chmielarz W.: „Aspekty zarządzania wiedzą w systemach wspomagających organizację", w materiałach konferencji naukowej "Systemy Wspomagania Organizacji" – SWO'2000, pod red. J. Gołuchowskiego i H. Sroki, Wydawnictwa Akademii Ekonomicznej w Katowicach, Katowice, 2000, str. 35-46,

[Dhar97] Dhar V., Stein R. : „Intelligent Decision Support Methods", Prentice Hall, 1997,

[Fayyad96] Fayyad U., Piatetsky-Shapiro G., Smyth P.: "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of ACM, vol. 39, New York, 1996.

[Gawrysiak01] Gawrysiak P.: „Data-mining – automatyczna eksploracja baz danych", WZ UW, Warszawa, 2001,

[Goebel99] Goebel M., Gruenwald L.: „A survey of Data-Mining and knowledge discovery software tools", SIGKDD Explorations, czerwiec 1999, str. 20-33,

[IBM00] IBM Corporation: „Intelligent Miner for Data 6.1 Instruction Manual", New York, 2000,

[Mattison97] Mattison R.: „Data Warehousing and Data-Mining for Telecommunications", Artech House, 1997,

[SAS99] SAS Institute White Papers: „Enterprise Miner Product Overview", „BI Systems and Data-Mining", 1999,

[Turban98] Turban E., McLean E., Wetherbe J.: "Information Technology for Management", Wiley&Sons, NY, 1998.

[Westpha99] Westphal C., Blaxton T.: „Data-Mining Solutions", Wiley and Sons, NY, 1999,

[Zaliwski00] Zaliwski A.: „Korporacyjne Bazy Wiedzy", PWE, Warszawa, 2000.

ISSN 0208-8028
ISBN 83-85847-59-6

W celu uzyskania bliższych informacji i zakupu dodatkowych egzemplarzy prosimy o kontakt z Instytutem Badań Systemowych PAN
ul. Newelska 6, 01-447 Warszawa
tel. 837-35-78 w. 241 e-mail: bibliote@ibspan.waw.pl