



## JAK WDROŻYĆ I STOSOWAĆ DATA MINING W PRAKTYCE?

**Tomasz Demski**

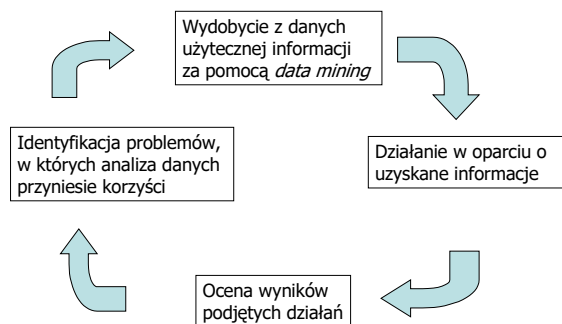
*StatSoft Polska Sp. z o.o.*

### Metodyka data mining

Realizacja złożonych projektów *data mining* (zgłębiania danych) w przedsiębiorstwach (i innych dużych organizacjach) wymaga skoordynowanego wysiłku ekspertów, właścicieli i różnych działów organizacji. W literaturze poświęconej *data mining* proponowane są różne metodyki, mogące służyć jako scenariusz, w jaki sposób należy zorganizować proces zbierania i analizy danych, rozpowszechniania wyników i sprawdzania korzyści z wdrażania projektu.

### *Virtuous Cycle of data mining (Właściwy cykl data mining)*

Jednym z modeli realizacji *data mining* jest *Virtuous Cycle of data mining* (zob. [3] i [4]) zaproponowany przez Berry'ego i Linoffa:



### Identyfikacja problemów

Na tym etapie należy określić problemy, których rozwiązanie z wykorzystaniem analizy danych przyniesie korzyści. Innymi słowy, osoby odpowiedzialne za pewną część działalności organizacji (np. z działu marketingu, zajmujące się programem lojalnościowym, z działu zapewnienia jakości) muszą zadać pytanie i określić swoje oczekiwania odnośnie odpowiedzi, a także wstępnie zdecydować, jak z tej odpowiedzi skorzystają. Nie wystarczy



np. zadać pytanie: „Które transakcje są podejrzane?”, ale trzeba też określić oczekiwania w stosunku do trafności przewidywań, że dana transakcja jest nadużyciem, i sposób wykorzystania modelu. Zauważmy, że inne są wymagania dla modelu stosowanego do bieżącego monitorowania wszystkich transakcji i alarmowania, że któraś z nich jest podejrzana, a inne do przygotowania ogólnych zaleceń dotyczących wykrywania nadużyć (np. należy zwrócić uwagę na wpłaty gotówki niewiele mniejsze od 10 000 \$).

Kluczową sprawą na etapie określania problemu jest dobra komunikacja między analitykami a praktykami. Całkowicie błędnym jest przekonanie, że analiza danych jest zagadnieniem technicznym i należy je pozostawić analitykom. Tylko osoby bezpośrednio zaangażowane w działalność przedsiębiorstwa rozumieją jego potrzeby i cel całego projektu.

Warto zwrócić uwagę, że spotyka się czasem wyobrażenie, iż *data mining* jest metodą automatyczną – wystarczy uruchomić program dla bazy danych i bez żadnego wysiłku dostaniemy użyteczną wiedzę. Jest ono oczywiście błędne – musimy przynajmniej zadać systemowi konkretne pytanie.

Na etapie identyfikacji problemu należy odpowiedzieć na pytania:

1. Czy projekt *data mining* jest istotnie potrzebny?
2. Czy wśród badanych obiektów da się wyróżnić szczególnie ważną i interesującą grupę?
3. Jakie są ogólne zasady prowadzenia działalności, które mogą wpłynąć na dostępność danych i możliwość podejmowania działań?
4. Jakie są właściwości danych? Jaka jest wiarygodność stosowanych źródeł danych? Gdzie, kiedy i w jaki sposób można uzyskać dane?
5. Jaka jest wiedza o problemie wynikająca z doświadczenia i intuicji praktyków?

Dosyć często projekt *data mining* stosuje się do problemu, który wcześniej był rozwiązywany innymi metodami (np. w oparciu o pisane lub niepisane reguły, czy wyniki prostych, jednowymiarowych analiz). Zawsze warto zebrać informacje o stosowanych wcześniej metodach, ich wadach i zaletach.

Na etapie formułowania problemu warto jest określić sposób oceny jego skuteczności.

### **Wydobycie z danych użytecznej informacji za pomocą *data mining*.**

Zasadniczą częścią *data mining* jest wydobycie z danych użytecznej wiedzy, która będzie stanowiła podstawę do podjęcia działań. Na tym etapie wykonujemy następujące czynności:

1. Identyfikacja i pozyskanie danych.
2. Sprawdzenie, zbadanie i oczyszczenie danych.
3. Uzyskanie właściwego układu danych.
4. Dodanie zmiennych wyliczonych na podstawie wartości cech.
5. Wybranie próby uczącej.
6. Wybranie metody modelowania.
7. Sprawdzenie dobroci dopasowania.



## Działanie na podstawie uzyskanych wyników

Cały projekt *data mining* jest realizowany po to, aby móc podjąć takie działania, które poprawią jakość funkcjonowania przedsiębiorstwa (lub innej organizacji). Odłożenie uzyskanej wiedzy na półkę oznacza klęskę całego projektu.

Przypomnijmy, że już na etapie formułowania problemu powinniśmy wstępnie określić działania, które podejmiemy po uzyskaniu modelu. Jednak czasami w trakcie procesu modelowania uzyskamy wiedzę, która pozwoli nam odkryć nieznane wcześniej drogi do polepszenia jakości działania przedsiębiorstwa, i to powinniśmy wykorzystać.

W pracy [4] wymieniono następujące rodzaje działań:

- ◆ Upowszechnienie wiedzy. W wyniku *data mining* zazwyczaj uzyskujemy wiele interesujących spostrzeżeń o działalności przedsiębiorstwa, o klientach itp. Wiedza ta powinna być upowszechniona wśród zainteresowanych osób (nawet jeśli nie są bezpośrednio zaangażowane w projekt).
- ◆ Jednorazowy wynik. Projekt może dotyczyć jednego, konkretnego działania, np. kampanii promocyjnej, akcji cross-sellingowej itp. W takim wypadku działanie to powinno być przeprowadzone z wykorzystaniem wiedzy wydobytej z danych.
- ◆ Zapamiętanie wyników. W wyniku projektu *data mining* możemy uzyskać informacje, które będą użyteczne w przyszłości, takie jak np. zyskowność klientów z poszczególnych segmentów itp. Informacje takie powinny zostać zapamiętane i upowszechnione przez system *business intelligence* przedsiębiorstwa.
- ◆ Regularne przewidywanie. Model może być używany wielokrotnie – np. co miesiąc możemy przewidywać prawdopodobieństwo tego, że dany kredyt przestanie być terminowo spłacany.
- ◆ Bieżące oceny. Model może zostać wbudowany w system informatyczny, tak aby na bieżąco wspierał podejmowanie decyzji. Przykładowo pracownik, przyznając kredyt, po wpisaniu informacji o kliencie uzyskuje prawdopodobieństwo, że będzie on bezproblemowo spłacał kredyt. Inny przykład to monitorowanie wszystkich transakcji kartami kredytowymi w celu wykrycia potencjalnych nadużyć.
- ◆ Poprawa jakości danych. Bardzo często w toku projektu wykryjemy błędne dane, a poprawa tych błędów może być jedną z korzyści z *data mining*. Czasami dane są w tak złym stanie, że po prostu nie można uzyskać użytecznego modelu i wtedy należy wrócić do drugiego etapu procesu *data mining*.

Zauważmy, że w toku projektu może się okazać, że jest inne miejsce, gdzie ukryte są „samorodki” użytecznej wiedzy i właśnie tam powinniśmy „kopać”. Opracowanie i wdrożenie nowego projektu może być również jednym z efektów przeprowadzonego projektu *data mining*.

## Ocena wyników podjętych działań

Jedną z niezbędnych czynności przy budowie modelu jest ocena jego skuteczności i praktycznej użyteczności. Jednak uzyskujemy wtedy tak naprawdę prognozy błędów, które



same mogą być błędne. Należy więc sprawdzić, jak nasze przewidywania mają się do rzeczywistości i jaki był skutek podjętych działań.

Dosyć często stosowaną techniką jest wydzielenie grupy kontrolnej, w stosunku do której nie podejmujemy działań sugerowanych przez wyniki *data mining*. Przykładowo możemy wysłać ofertę cross-sellingową do klientów wskazanych przez model i do losowo wybranej grupy klientów, po czym porównać stopę pozytywnych odpowiedzi w obu grupach.

### **Metodyka Six Sigma**

Innym podejściem jest metodyka Sześć Sigma (Six Sigma). Jest to dobrze zorganizowana, bazująca na danych strategia unikania wad i problemów z jakością we wszystkich rodzajach produkcji i usług, zarządzaniu i innej działalności biznesowej. Metodyka Sześć Sigma staje się ostatnio coraz bardziej popularna (ze względu na wiele udanych wdrożeń) w USA i na całym świecie. Metodyka Sześć Sigma zaleca następujące etapy (tzw. DMAIC):

- ◆ **Definiuj.** Na tym etapie określa się cele i ograniczenia, identyfikuje się zagadnienia, którymi trzeba się zająć, by osiągnąć wyższy poziom sigma.
- ◆ **Mierz.** W tej fazie planu Sześć Sigma zbiera się informacje o aktualnym stanie procesu, by ustalić poziom odniesienia oraz by rozpoznać skalę problemu.
- ◆ **Analizuj.** Celem tej fazy jest wskazanie krytycznych przyczyn kłopotów z jakością i potwierdzenie, z użyciem odpowiednich analiz, ich wpływu na proces.
- ◆ **Poprawiaj.** Na tym etapie wprowadza się rozwiązania usuwające analizowane wcześniej, krytyczne problemy.
- ◆ **Sprawdzaj.** W tej fazie sprawdza się i monitoruje wyniki osiągnięte w poprzednim etapie.

Wywodzą się one z tradycji doskonalenia jakości i sterowania procesami i szczególnie dobrze nadają się do zastosowania w produkcji i świadczeniu usług.

### **Podsumowanie**

Stosowanie metodyk data mining wymaga odpowiedniego zorganizowania wymiany informacji, dostępu do danych, procesu analitycznego i udostępniania wyników. Narzuca to konieczność zastosowania odpowiedniego oprogramowania analitycznego, które łatwo integruje się z systemem informatycznym przedsiębiorstwa.

Niektóre aplikacje *data mining* zostały zaprojektowane i udokumentowane tak, aby spełniać wymogi tylko jednej z opracowanych strategii *data mining*. Natomiast system *STATISTICA Data Miner* jest środowiskiem *data mining*, które ma zastosowanie w dowolnej organizacji, gałęzi przemysłu i kulturze organizacyjnej, bez względu na ogólny model procesu *data mining*, na który zdecydowała się dana organizacja. Przykładowo system *STATISTICA Data Miner* może zawierać pełny zakres narzędzi wymaganych do wdrożenia, metodyki Sześć Sigma w całej organizacji, a użytkownicy mogą korzystać ze środowiska zorientowanego na DMAIC (które jest jedną z wielu opcji do wyboru). System również dobrze spełni rolę części projektu CRM (*Customer Relationship Management*),



związanej z badaniami. Ponadto system *STATISTICA Data Miner* ma zalety ogólnego, przeznaczonego do *data mining* systemu, zawierającego narzędzia umożliwiające nie tylko stosowanie w projektach takich obiektów, jak: połączenia z bazami danych, interakcyjne zapytania do baz danych i własne algorytmy. Dodatkowo (korzystając z opcjonalnych aplikacji StatSoft, np. *WebSTATISTICA*) można korzystać z narzędzi pracy grupowej i tworzyć rozbudowane systemy korporacyjne obejmujące całą organizację i spełniające wymogi wybranej strategii.

Doświadczeni menedżerowie wiedzą, że programy i metodyki zarządzania i rozwoju organizacji podlegają ciągłym zmianom (dla przykładu porównajmy dzisiejsze środowisko organizacyjne, technologiczne, rynkowe i relacje z klientami z tymi z połowy lat 90., gdy komercyjne zastosowania Internetu były w powijakach). *STATISTICA Data Miner* jest również zmieniany wraz ze zmianą technologii i warunków rynkowych, tak aby był najaktualniejszym narzędziem *data mining* o najbardziej wygodnej i dostosowywalnej postaci. Legalni użytkownicy systemu bez dodatkowych opłat otrzymują aktualizacje i udoskonalenia, zawierające najnowsze procedury, które mogą bezpośrednio po ich przygotowaniu wykorzystywać w swoich projektach. Dzięki temu użytkownik ma zapewniony nieprzerwany dostęp do aktualnego zestawu narzędzi do „odkrywania zależności, wyjaśniania zaobserwowanych tendencji i przewidywania przyszłości”.

## Oprogramowania *data mining*

### *Wymagania*

Realizacja *data mining* nie jest możliwa bez zastosowania odpowiedniego oprogramowania. Oprogramowanie takie powinno umożliwiać:

- ◆ wydajny dostęp do danych w różnych formatach,
- ◆ przygotowanie danych dla potrzeb właściwej analizy,
- ◆ przeprowadzenie analizy *data mining* (nawet dla ogromnych zbiorów danych),
- ◆ przygotowanie raportu i wdrożenie uzyskanych wyników.

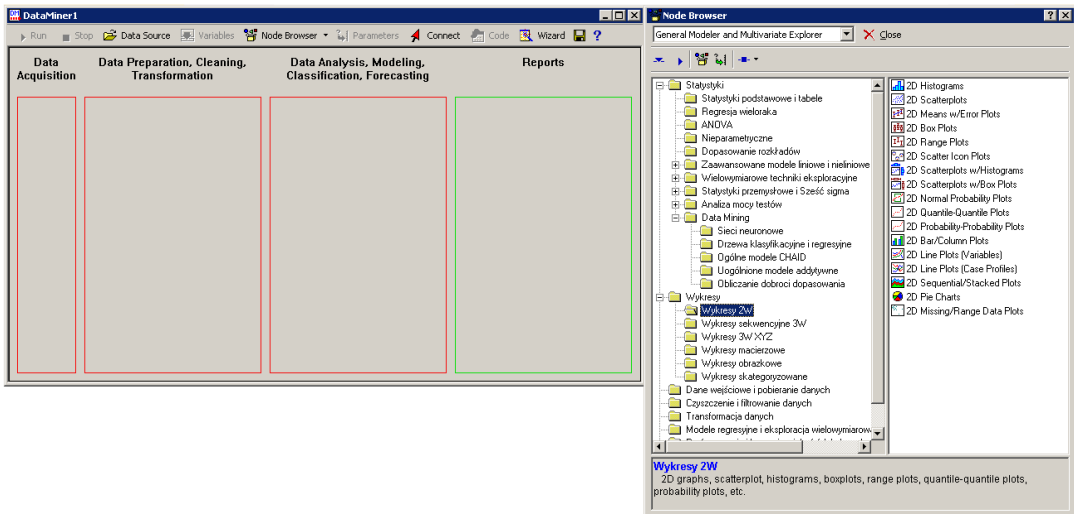
Powyższe wymagania stanowią minimalne kryterium, które musi być spełnione przez każdy w pełni funkcjonalny system *data mining* (zob.[5], [6]) Ponadto cennymi zaletami systemu są:

- ◆ prostota obsługi,
- ◆ możliwość współpracy z systemem informatycznym przedsiębiorstwa,
- ◆ skalowalność,
- ◆ możliwość dostosowania do konkretnych potrzeb i upodobań użytkownika,
- ◆ możliwość automatyzacji rutynowych zadań,
- ◆ bogactwo narzędzi analizy i wizualizacji danych.



## STATISTICA Data Miner

System *STATISTICA Data Miner* spełnia wymienione wyżej wymagania, a ponadto ma wiele unikalnych cech i funkcji.



Rys. 1. Przestrzeń robocza *STATISTICA Data Miner* wraz z Przeglądarką węzłów

Środowisko *STATISTICA Data Miner* zostało zaprojektowane tak, aby zapewnić:

- ◆ proste i intuicyjne budowanie oraz modyfikowanie projektów analiz,
- ◆ możliwość szybkiego zorientowania się w działaniu nawet bardzo złożonego projektu,
- ◆ naturalne i bezproblemowe stosowanie uzyskanych modeli dla nowych danych.

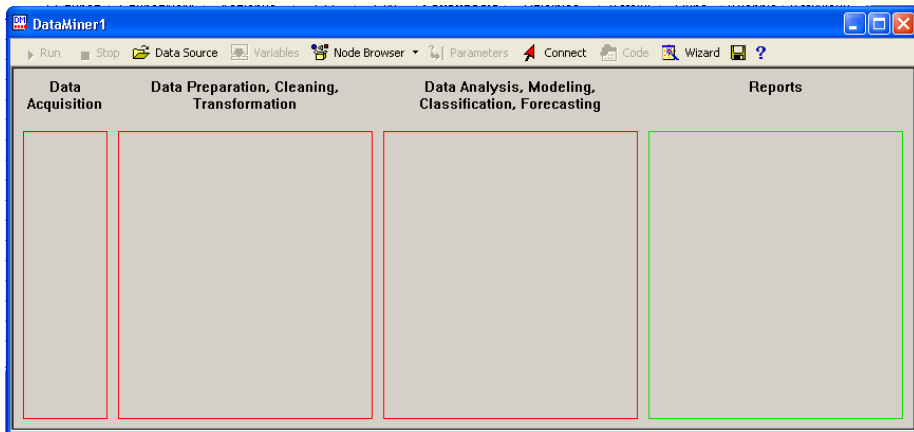
W środowisku systemu z projektami pracujemy w specjalnej przestrzeni roboczej, a cały projekt i przepływ danych między kolejnymi etapami analizy reprezentowany jest graficznie.

Proces *data mining* możemy podzielić na cztery części:

- ◆ pobranie danych,
- ◆ wstępna obróbka danych,
- ◆ wykonanie analizy,
- ◆ przygotowanie i udostępnienie raportów z analizy.

Zgodnie z tym podziałem przestrzeń roboczą podzielono również na cztery segmenty (zob. rysunek poniżej).

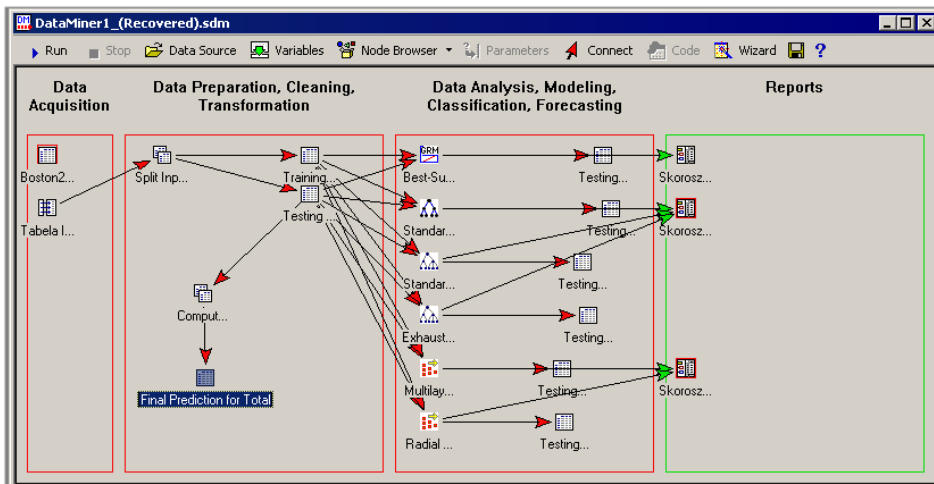
W środowisku systemu źródła danych, procedury przetwarzające dane i raporty reprezentowane są przez ikonę (tzw. węzeł). Przepływ danych obrazują strzałki łączące poszczególne węzły. Węzły zaprojektowane są tak, że dane wypływające z jednego z węzłów mogą stanowić wejście dla innych węzłów. Zapewnia to możliwość składania projektu analizy z poszczególnych węzłów, tak jak z klocków.



Dwa specjalne typy węzłów to: źródła danych i raporty. W przypadku źródeł danych nie mamy strzałek dochodzących do tych węzłów, lecz wyłącznie strzałki wychodzące z węzłów. W przypadku węzłów raportu mamy tylko strzałki dochodzące do węzła.

W postaci węzłów dostępne są setki procedur przekształcania, analizy i wizualizacji danych. Wszystkie węzły dostępne są w *Przeglądarce węzłów*, która ma uporządkowaną hierarchiczną strukturę. W systemie zdefiniowano wiele przeglądarek, zawierających tylko te węzły, które są potrzebne przy wykonywaniu zadań pewnego typu (np. prognozowania szeregu czasowego). Użytkownik może zdefiniować własną przeglądarkę, z tymi narzędziami, z których najczęściej korzysta.

Projekt *data mining* budujemy, wstawiając węzły z *Przeglądarki węzłów* do przestrzeni roboczej *STATISTICA Data Miner* i łącząc je ze sobą strzałkami. Dzięki takiemu trybowi pracy nawet skomplikowaną, wieloetapową analizę możemy łatwo zbudować i modyfikować, przeciągając obiekty myszką. Ponadto łatwo jest zorientować się w strukturze projektu. Przykładowy projekt widzimy poniżej:





Środowisko systemu możemy dostosować do własnych potrzeb i upodobań. Można m.in. zbudować własną przeglądarkę węzłów, tworzyć własne węzły analityczne (korzystając z makr nagranych podczas interakcyjnego specyfikowania analiz i wykresów), a nawet uruchamiać system i sterować nim z innych aplikacji.

Wersja klient-serwer *STATISTICA Data Miner* wykorzystuje przetwarzanie rozproszone i architekturę wielowarstwową, co pomaga optymalnie wykonywać bardzo złożone zadania obliczeniowe. Technologia ta umożliwia szybkie wykonywanie nawet bardzo dużych projektów z pełnym wykorzystaniem wielu procesorów serwera lub wielu serwerów pracujących równolegle. W przypadku wersji klient-serwer *STATISTICA Data Miner* wszystkie obliczenia odbywają się na serwerze (lub wielu serwerach) *STATISTICA*, a stacja robocza użytkownika obsługuje wyłącznie interfejs programu. Na komputerze użytkownika nie trzeba instalować żadnego oprogramowania – wystarczy przeglądarka internetowa.

Oprócz przedstawionego powyżej środowiska, możemy pracować w typowym środowisku graficznym, w którym analizy i wyniki wybieramy interakcyjnie. W szczególności na etapie wstępnej, eksploracyjnej analizy taki tryb pracy może być bardzo wygodny.

## Dostęp do danych

Proces *data mining* zazwyczaj dotyczy bardzo dużej ilości danych, przechowywanych w bazach i hurtowniach danych, często o bardzo złożonej i skomplikowanej strukturze. Standardowym narzędziem pobierania danych z bazy danych jest język SQL. Użytkownicy nieznający tego języka i zaawansowanej technologii informatycznej również potrzebują łatwego dostępu do danych gromadzonych w różnych repozytoriach danych. Każde narzędzie *data mining* powinno więc zawierać wbudowane mechanizmy dostępu do zewnętrznych baz danych, gwarantujące użytkownikowi interakcyjną budowę zapytań do bazy danych i łatwy dostęp do danych.

*STATISTICA Data Miner* umożliwia budowanie zapytań do baz danych w graficznym środowisku użytkownika, a także korzystanie z utworzonych wcześniej zapytań SQL (w trybie tekstowym).

System korzysta z trzech rodzajów źródeł danych:

- ◆ lokalnych plików o formacie *STATISTICA*,
- ◆ zdalnych baz danych podłączonych za pośrednictwem technologii IDP,
- ◆ węzłów zaprojektowanych przez użytkownika (reprezentujących utworzone przez użytkownika procedury generujące dane lub pobierające je ze specyficznego źródła, np. urządzenia pomiarowego).

Warto zwrócić szczególną uwagę na technologię IDP. Umożliwia ona pracę na bazie danych bez konieczności importowania danych i tworzenia pliku lokalnego. Technologia IDP jest użyteczna przy przetwarzaniu bardzo dużych zbiorów danych; w takich przypadkach jej zastosowanie daje duży wzrost wydajności i umożliwia przetwarzanie zbiorów danych o wielkości przekraczającej pojemność urządzeń lokalnych.



The screenshot shows the STATISTICA Query environment. On the left, a tree view displays the database schema with tables like 'Dostawcy', 'Kategorie', and 'Produkty'. The main window shows a query builder with a diagram of table relationships. Below the diagram, there are tabs for 'Sekwencja pól', 'Kryteria', 'Podgląd danych', and 'Wyrażenie SQL'. To the right, a window titled 'Tabela IDP1\*' displays a table with columns: 'Quantity', 'Product', 'Unit Price', 'Total', and 'Ship to State'. The table contains data for various products like 'Glove', 'Jersey', 'Baseball', and 'Jersey'. Below the table is a dialog box 'Opcje zapytania' (Query Options) with settings for cursor type, server, and buffer size.

Środowisko pobierania danych – STATISTICA Query

Dane przeniesione do tabeli IDP, oraz okno opcji umożliwiających wybór zdalnego przetwarzania danych

W wersji klient-serwer wszystkie obliczenia przenoszone są na serwer aplikacji STATISTICA, zmniejszając w ten sposób liczbę wykonywanych zadań na serwerze bazy danych i stacji roboczej użytkownika. W procesie pobierania danych z bazy danych można posłużyć się narzędziem STATISTICA Query, umożliwiającym interakcyjną budowę zapytania i pobranie lub połączenie w ten sposób danych ze środowiskiem STATISTICA Data Miner.

System STATISTICA Data Miner może przetwarzać praktycznie dowolnie duże zbiory danych (jedynym ograniczeniem są możliwości komputera), jednak w praktyce bardzo często potrzebne jest wyodrębnianie z ogółu danych prób losowych, np. przy podziale danych na próbę uczącą i testową. W związku z tym w system wbudowano bardzo wydajne, wysokiej jakości procedury tworzenia prób losowych (przeszły one niezwykle wymagające, niezależne testy DIEHARD).

Projekt data mining może być automatycznie aktualizowany przy każdej zmianie danych. Dzięki temu możemy np. zbudować system automatycznie wykrywający podejrzone transakcje i powiadamiający o tym odpowiednie osoby.

### Przekształcanie danych dla potrzeb analiz

Często dane w bazie danych gromadzone są automatycznie, dotyczyć one mogą (jak w hurtowni danych) wszystkich gałęzi przedsiębiorstwa. Czasem nie potrzebujemy aż tak bogatych danych, tylko opisu ściśle określonych cech. Po połączeniu systemu data mining z wybranymi danymi może okazać się, że konieczne jest wstępne ich przetworzenie.

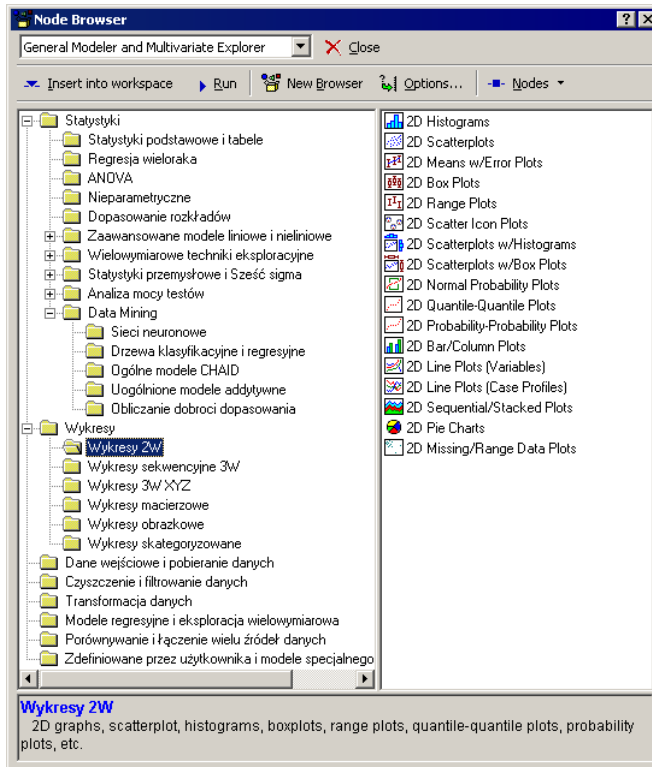
Na przykład gdy w modelu data mining korzystamy z sieci neuronowych, a dane, które pobraliśmy z bazy danych, raczej nie zawierają zmiennej grupującej, wybierającej podzbiór uczący i testowy, to konieczny jest algorytm umożliwiający taki podział. Trudno sobie bowiem wyobrazić ręczne wybieranie podzbioru z 5 000 przypadków.



*STATISTICA Data Miner* umożliwia szeroki wybór algorytmów czyszczenia i transformacji danych. Wszystkie te algorytmy łącznie z szerokim wyborem statystyk i wykresów dostępne są w przeglądarce węzłów (rys. 2).

Korzystając z węzłów czyszczenia, filtrowania i transformacji danych, możemy filtrować dane z użyciem określonej wartości minimalnej i maksymalnej, usuwać brakujące dane lub zastępować je średnią, eliminować i dobierać zmienne, które prawdopodobnie będą najlepszymi predyktorami dla bieżącego zbioru danych i nie wprowadzą szumów do budowanych modeli. Prócz tego mamy do dyspozycji szereg algorytmów umożliwiających transponowanie, sortowanie, rangowanie i standaryzowanie danych. Przy zagadnieniach regresyjnych i klasyfikacyjnych możemy korzystać z algorytmów podziału zbioru danych na próby: uczącą i testową (np. dla sieci neuronowych) lub w przypadku wielu modeli zawartych w jednym projekcie - obliczać dla nich najlepsze predyktory.

W każdym z powyżej opisanych przypadków jako wynik otrzymamy arkusz odpowiednio przetworzonych danych, który będzie wykorzystany w dalszych etapach projektu *data mining*.



Rys. 2. Przeglądarka węzłów systemu *STATISTICA Data Miner*

Możliwość wstępnego przetwarzania danych w narzędziach *data mining* gwarantuje wysoką jakość i spójność danych trafiających do modelu. Jest to równie istotne jak wybór samej analizy, gdyż zyskujemy pewność pełnego i prawidłowego wykorzystania modelu.



## **Analiza danych, modelowanie, klasyfikacja**

Kluczowym elementem każdego projektu *data mining* jest wybór analiz odpowiednich dla typu i treści danych. Technika *data mining* rozwiązujemy określone grupy zadań:

- ◆ opis danych,
- ◆ klasyfikacja,
- ◆ modelowanie zmiennej ciągłej,
- ◆ prognozowanie,
- ◆ analiza koszykowa (analiza asocjacji),
- ◆ analiza skupień (segmentacja),

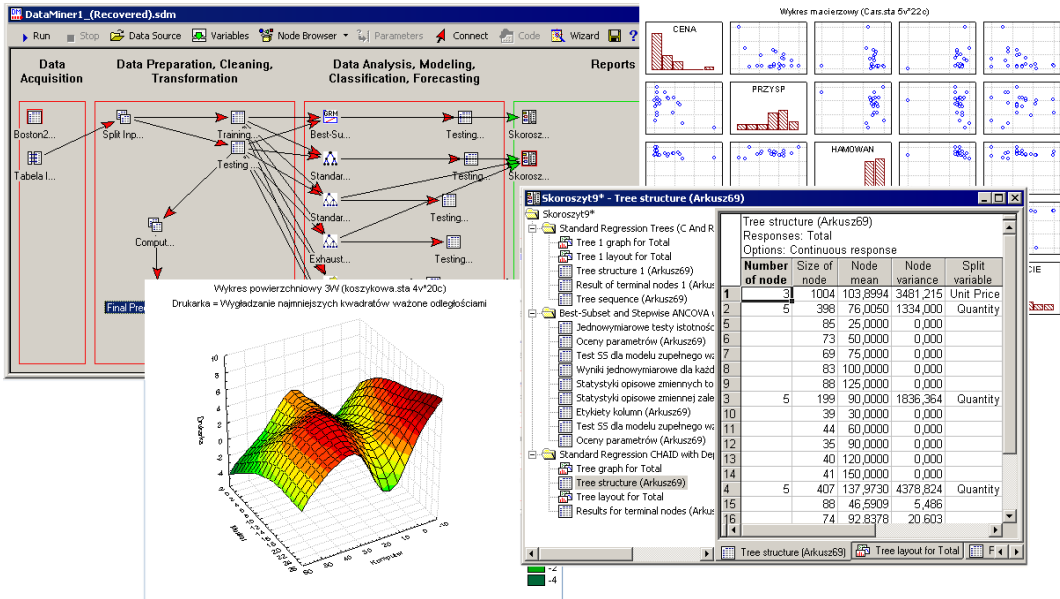
a z każdą z powyższych grup związane są określone procedury analityczne.

System *STATISTICA Data Miner* dysponuje wszystkimi procedurami analitycznymi z programu *STATISTICA*. Ponadto dysponuje najnowszymi, wyrafinowanymi technikami analizy danych, takimi jak np.: *Uogólnione modele addytywne*, *Ogólne modele CHAID*, *Drzewa klasyfikacyjne i regresyjne*, *Drzewa interakcyjne (C&RT, CHAID)*, *Sieci neuronowe*, *Interakcyjne drążenie danych*, *Obliczanie dobroci dopasowania*, *Dobór i eliminacja zmiennych*, *Analiza koszykowa*, *Algorytm krzywych składanych rekurencyjnego podziału przestrzeni cech do budowy modelu regresyjnego (MAR Splines)*, *Metoda SVM (Support Vector Matrix)*, *Metoda k najbliższych sąsiadów*, *Naiwny klasyfikator Bayesa*. W systemie dostępne są gotowe projekty dla poszczególnych typów zadań, do których wystarczy podłączyć źródło danych, aby móc z nich skorzystać.

## **Wizualizacja i rozpowszechnianie wyników**

*STATISTICA Data Miner* zawiera wszechstronny zestaw narzędzi przeznaczonych do graficznej eksploracji i analizy danych, służących do poszukiwania relacji, trendów i błędów systematycznych, „ukrytych” w nieuporządkowanych zbiorach danych. Techniki graficznego *data mining* obejmują: dopasowywanie i wykreślanie funkcji, wygładzanie danych, nakładanie i łączenie wielu obrazów, interakcyjne rozdzielanie danych na kategorie, podział i łączenie podzbiorów danych na wykresach, agregowanie danych, identyfikacje i zaznaczanie podzbiorów danych spełniających określone warunki, wykreślanie przedziałów i obszarów (elips) ufności, generowanie wykresów mozaikowych, płaszczyzn spektralnych, rzutowanych warstw, techniki redukcji obrazowanych danych, interakcyjne (i płynne) obracanie wykresów trójwymiarowych, selektywne podświetlanie określonych serii i bloków danych, interakcyjne wybieranie trójwymiarowych obszarów na wykresie, analityczne narzędzie powiększania, umożliwiające interakcyjny wybór fragmentu wykresu i przedstawienie go na oddzielnym wykresie oraz wiele innych.

Tak szeroki wybór graficznych technik prezentacji wyników i graficznego *data mining* umożliwia dostosowanie wizualizacji danych do indywidualnych potrzeb i oczekiwań użytkownika.



Rys. 3. Projekt *data mining* ze skoroszytem zawierającym wykresy i tabelę podsumowującą

Wyniki analiz możemy kierować do skoroszytu, oddzielnych węzłów (każda tabela lub wykres będą stanowiły odrębny obiekt w przestrzeni roboczej projektu) lub bezpośrednio do raportu. Tak sformatowane wyniki projektu *data mining*, lub ich część, możemy zapisać w postaci elektronicznej i przelać innym współpracownikom lub klientom pocztą elektroniczną. Alternatywnym i bardziej globalnym rozwiązaniem jest umieszczenie ich w Internecie lub lokalnym Intranecie w postaci plików html. Jeśli potrzebujemy autoryzowanego dostępu do raportów za pośrednictwem Internetu, to system *STATISTICA Data Miner* można rozbudować o aplikację *STATISTICA Knowledge Portal*.

Możliwość prezentacji samych wyników w środowisku Internetowym to nie wszystko. *STATISTICA Data Miner*, dzięki opcjonalnej aplikacji *WebSTATISTICA Server*, umożliwia wykonywanie wszystkich operacji *data mining* w oknie przeglądarki internetowej, na dowolnym komputerze połączonym z Internetem. Taka konfiguracja sprawia, że użytkownik nie musi mieć na swoim komputerze zainstalowanego systemu *STATISTICA Data Miner*, natomiast korzysta ze wszystkich jego funkcji poprzez przeglądarkę internetową. Dzięki temu można pracować nad projektami przez Internet i współpracować zarówno z osobami w tym samym biurze, jak i na innym kontynencie.

## Literatura

1. *STATISTICA Data Miner*, 2002, StatSoft Inc.
2. Weiss S.M, Indurkha N., 1998, Predictive data mining. A practical guide, Morgan Kaufman Publishers.



3. Berry M.J.A., Linoff G. 1997, Data mining techniques: for marketing, sales, and customer support, John Willey & Sons.
4. Berry M.J.A., Linoff G., 2000, Mastering data mining, , John Willey & Sons.
5. Data mining – metody i przykłady, 2002, Statsoft Polska.
6. Han, J, Kamber, M, Data mining: Concepts and Techniques, 2001, Academic Press.
7. Berson, A., Smith, S., Thearing, K., 1999, Building Data mining Applications for CRM, McGraw-Hill.