



Danuta Mendrala

Marcin Szeliga

Microsoft SQL Server

Modelowanie i eksploracja danych



Helion



Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiegokolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz Wydawnictwo HELION dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz Wydawnictwo HELION nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Redaktor prowadzący: Michał Mrowiec
Korekta merytoryczna: Radosław Łebkowski
Projekt okładki: Jan Paluch

Fotografia na okładce została wykorzystana za zgodą Shutterstock.com

Wydawnictwo HELION
ul. Kościuszki 1c, 44-100 GLIWICE
tel. 32 231 22 19, 32 230 98 63
e-mail: helion@helion.pl
WWW: <http://helion.pl> (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
<http://helion.pl/user/opinie?sqlsme>
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Kody źródłowe wybranych przykładów dostępne są pod adresem:
<ftp://ftp.helion.pl/przyklady/sqlsme.zip>

ISBN: 978-83-246-3440-8

Copyright © Helion 2012

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to! » Nasza społeczność](#)

Spis treści

Wstęp	9
Proces eksploracji danych	10
Instalacja i konfiguracja narzędzi	12
Serwer SQL	12
Arkusz kalkulacyjny Excel	15
Dodatek Data Mining do pakietu Office	15
Przykłady	16
Konwencje i oznaczenia	20
 Część I Modelowanie	 23
Rozdział 1. Eksploracja danych jako technika wspomagania decyzji	25
Modelowanie świata	25
Obiekty, zdarzenia i reguły	26
Dane	27
Informacje	27
Wiedza	29
Decyzje	31
Eksploracja danych	32
Hipotezy	32
Kłopoty ze sformulowaniem problemu	33
 Rozdział 2. Analiza biznesowa	 35
Cele modelowania i eksploracji danych	35
Opisywanie danych czy wspieranie decyzji?	36
Decydenci	38
Zakres projektu eksploracji danych	39
Dane źródłowe	40
Kontekst	40
Sprecyzowanie spodziewanych wyników	42
Modele deskrypcyjne	43
Modele predykcyjne	43
Prawdopodobieństwo sukcesu projektu eksploracji danych	44
Ocena ryzyka	45

Rozdział 3. Ocena danych	49
Dane źródłowe	49
Błędy pomiaru	50
Przypadki, czyli to, co badamy	51
Profilowanie danych za pomocą usługi SQL Server Integration Services	54
Atrybuty i ich stany	57
Atrybuty jednowartościowe i wielowartościowe	57
Atrybuty monotoniczne	59
Rozkład wartości	59
Integralność danych	62
Duplikaty	62
Zakres wartości	63
Zgodność ze wzorcem	63
Próbkowanie i reprezentatywność danych	64
Próbkowanie danych	64
Zbieżność do rzeczywistego rozkładu	65
Odchylenie standardowe	67
Zmienność atrybutów tekstowych	68
Brakujące dane	69
Model brakujących danych	70
Zależności pomiędzy atrybutami	73
Niezależne atrybuty	74
Nadmiarowe atrybuty	75
Anachronizmy	76
Mierzenie informacji	76
Bity	77
Zaskoczenie	77
Kontekst	78
Rozdział 4. Przygotowanie danych	79
Przestrzeń stanów	79
Atrybuty dyskretne	81
Grupowanie	81
Numerowanie stanów	84
Atrybuty porządkowe	85
Atrybuty okresowe	86
Atrybuty ciągłe	86
Wartości skrajne	87
Normalizacja zakresu	87
Dyskretyzacja	90
Serie danych	92
Trend	96
Okresowość i sezonowość	96
Szum	97
Rozdział 5. Poprawa jakości danych	99
Uzupełnienie wartości	99
Wzbogacenie danych	103
Redukcja wymiarów	105
Korelacje	106

Dane dla modeli deskrypcyjnych	108
Dane dla modeli predykcyjnych	109
Zmiana proporcji	109
Dane na potrzeby analizy wariantowej	111
Analiza wariantowa	111
Wydzielenie danych testowych	113
Część II Eksploracja	117
Rozdział 6. Techniki eksploracji danych	119
Zastosowania	119
Dodatek Data Mining do pakietu Office	121
Ocena i przygotowanie danych źródłowych	121
Techniki eksploracji danych	126
Klasyfikacja	126
Szacowanie	136
Asocjacja	141
Grupowanie	145
Analiza sekwencyjna	151
Analiza wariantowa	152
Prognozowanie	156
Rozdział 7. Serwer SQL jako platforma eksploracji danych	161
Excel jako klient SQL Server Analysis Services	162
Narzędzia eksploracji zewnętrznych danych	162
Praca z modelami eksploracji danych	184
Formuły arkusza Excel	191
Projekty eksploracji danych	192
Business Intelligence Development Studio	192
Źródła danych	195
Widoki danych źródłowych	196
Struktury eksploracji danych	199
Modele eksploracji danych	206
Zapytania predykcyjne	210
Zagnieżdżanie przypadków	213
Zarządzanie serwerem SSAS i modelami eksploracji danych poprzez SQL	
Server Management Studio	216
Usługi eksploracji danych serwera SQL	218
Architektura	219
Bezpieczeństwo	221
Integracja z pozostałymi usługami Business Intelligence	223
Rozdział 8. DMX	227
Terminologia	227
Atrybut	227
Wartość i stan	229
Przypadek	229
Klucze	230
Struktury eksploracji danych	231
Modele eksploracji danych	232

Składnia języka DMX	232
Tworzenie struktur eksploracji danych	233
Tworzenie modeli eksploracji danych	235
Przetwarzanie struktur i modeli eksploracji danych	239
Odczytywanie zawartości struktur i modeli eksploracji danych	243
Zapytania predykcyjne	245
Funkcje predykcyjne	251
Rozdział 9. Naiwny klasyfikator Bayesa firmy Microsoft	253
Omówienie	253
Ograniczenia	255
Parametry	256
Zastosowania naiwnego klasyfikatora Bayesa	258
Badanie zależności pomiędzy atrybutami	258
Klasyfikacja dokumentów	260
Rozdział 10. Drzewa decyzyjne firmy Microsoft i algorytm regresji liniowej firmy Microsoft	267
Omówienie	268
Ograniczenia	272
Parametry	273
Zastosowania drzew decyzyjnych	275
Klasyfikacja klientów	275
Szacowanie potencjalnych zysków	277
Asocjacja klientów i wypożyczanych przez nich filmów	279
Rozdział 11. Szeregi czasowe firmy Microsoft	281
Omówienie	281
Ograniczenia	285
Parametry	286
Zastosowania szeregów czasowych firmy Microsoft	288
Prognozowanie sprzedaży	289
Prognozowanie sprzedaży na podstawie przeplatanych serii danych	291
Prognozowanie sprzedaży na podstawie danych odczytanych z kostki wielowymiarowej	292
Prognozowanie sprzedaży na podstawie krótkich serii danych	293
Analiza wariantowa	295
Rozdział 12. Algorytm klastrowania firmy Microsoft	297
Omówienie	297
Ograniczenia	302
Parametry	303
Zastosowania algorytmu klastrowania	305
Analiza skupień komórek	305
Klasyfikacja komórek	309
Przygotowanie danych do dalszej eksploracji	312
Wykrywanie anomalii	314

Rozdział 13. Algorytm klastrowania sekwencyjnego firmy Microsoft	319
Omówienie	320
Ograniczenia	323
Parametry	323
Zastosowania algorytmu klastrowania sekwencyjnego	324
Analiza sekwencji odwiedzanych stron WWW	324
Klasyfikacja klientów na podstawie kolejności kupowanych przez nich towarów	327
Przewidywanie kolejnych zdarzeń	329
Wykrywanie nietypowych sekwencji zdarzeń	332
Rozdział 14. Algorytm odkrywania reguł asocjacyjnych firmy Microsoft	335
Omówienie	336
Ograniczenia	340
Parametry	341
Zastosowania reguł asocjacyjnych	341
Badanie zależności pomiędzy wartościami atrybutów	342
Analiza koszykowa	343
Analiza typu cross-selling	347
Rozdział 15. Sieci neuronowe firmy Microsoft i algorytm regresji logistycznej firmy Microsoft	351
Omówienie	352
Ograniczenia	358
Parametry	360
Zastosowania sieci neuronowych i regresji logistycznej	361
Szacowanie potencjalnych zysków	362
Klasyfikacja dokumentów	366
Rozdział 16. Ocena i poprawa modeli	369
Powrót do średniej	369
Kryteria porównawcze	371
Łatwość interpretacji	373
Dokładność predykcji	374
Wiarygodność predykcji	374
Wydajność i skalowalność	375
Przydatność	375
Metody oceniania modeli eksploracji danych	376
Wykresy podniesienia i zysku	376
Macierz klasyfikacji	384
Ocena dokładności modeli algorytmu szeregów czasowych firmy Microsoft	386
Walidacja krzyżowa	387
Odchylenie wewnątrz- i międzyklastrowe	390
Problemy	391
Niewłaściwie postawione zadania	391
Niewłaściwe dane źródłowe	392
Nieprzygotowane dane źródłowe	393
Niewłaściwe lub źle sparаметryzowane algorytmy eksploracji danych	394

Rozdział 17. Programowanie predykcyjne	397
Narzędzia programistyczne	397
Wizualizatory modeli eksploracji danych	398
Raporty usługi SSRS	399
Inteligentne aplikacje	401
Kontrola poprawności danych	401
Uzupełnianie brakujących danych	404
Adaptacyjny interfejs	406
Skorowidz	413

Rozdział 9.

Naiwny klasyfikator Bayesa firmy Microsoft

- ◆ Dlaczego klasyfikator Bayesa nazywany jest naiwnym?
- ◆ Jakie są wady i zalety naiwnego klasyfikatora Bayesa firmy Microsoft?
- ◆ Jak tworzyć modele eksploracji danych używające naiwnego klasyfikatora Bayesa firmy Microsoft?
- ◆ Jak parametryzować naiwny klasyfikator Bayesa firmy Microsoft?
- ◆ Jak za pomocą naiwnego klasyfikatora Bayesa firmy Microsoft badać zależności pomiędzy atrybutami?
- ◆ Jak zbudować klasyfikujący dokumenty filtr antyspamowy przy użyciu naiwnego klasyfikatora Bayesa firmy Microsoft?



Nazwy wszystkich przedstawionych algorytmów eksploracji danych zawierają określenie *firmy Microsoft* nie dlatego, że algorytmy te zostały wymyślone przez Microsoft, ale dlatego, że to ta firma stworzyła zastosowane w serwerze SQL implementacje tych algorytmów.

Omówienie

Opracowany przez brytyjskiego matematyka i prezbiteriańskiego duchownego Thomasa Bayesa klasyfikator należy do klasycznych algorytmów uczenia przez obserwację¹. Wyobraźmy sobie, że spędzamy wolny czas, obserwując klientów właśnie otwartego butiku. Interesuje nas, kto (kobieta czy mężczyzna) za chwilę wejdzie do tego sklepu.

¹ Będące podstawą opisywanego klasyfikatora twierdzenie Bayesa zostało opublikowane w wydanym w 1763 roku esej *Essay Towards Solving a Problem in the Doctrine of Chances*. Dokument ten jest dostępny pod adresem <http://www.stat.ucla.edu/history/essay.pdf>.

Ponieważ w naszym miasteczku mieszka więcej kobiet niż mężczyzn (60% mieszkańców to kobiety, a 40% — mężczyźni), początkowo prawdopodobieństwo, że klientem będzie kobieta, wynosi 60%. Jednak po pewnym czasie zebraliśmy więcej informacji o rozkładzie dnia klientów i zauważyliśmy m.in., że przed południem butik odwiedzają głównie (w 80%) kobiety, a po godzinie 15.00 75% klientów to mężczyźni. Jeżeli od teraz usłyszymy, że ktoś wchodzi do tego sklepu o 11.15, wiemy, że prawdopodobnie jest to kobieta ($60\% \cdot 80\% = 48\%$), a nie mężczyzna ($40\% \cdot 20\% = 8\%$). Natomiast gdybyśmy usłyszeli osobę wchodzącą do butiku o 15.30, mielibyśmy podstawy przypuszczać, że jest to mężczyzna ($40\% \cdot 75\% = 30\%$), a nie kobieta ($60\% \cdot 25\% = 15\%$). Ten uproszczony przykład pokazuje istotę naiwnego klasyfikatora Bayesa.

Naiwny klasyfikator Bayesa zlicza zależności występujące pomiędzy atrybutami wyjściowymi a poszczególnymi atrybutami wejściowymi, uwzględniając warunkowe i bezwarunkowe prawdopodobieństwo ich wystąpienia:

1. Prawdopodobieństwo bezwarunkowe (początkowe) zależy od rozkładu przypadków — w powyższym przykładzie reprezentowane jest ono przez fakt, że 60% mieszkańców miasteczka to kobiety.
2. Warunkowe prawdopodobieństwo zależy od zaobserwowanych zdarzeń — w powyższym przykładzie zaobserwowaliśmy, że 75% klientów odwiedzających butik po południu to mężczyźni.

Obliczone na podstawie twierdzenia Bayesa ($P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$)² prawdopodobieństwa są następnie mnożone, a więc każde z nich ma taki sam wpływ na ostateczny wynik. To założenie jest prawdziwe, o ile poszczególne atrybuty wejściowe są od siebie niezależne³. W przeciwnym razie wpływ atrybutu skorelowanego z innym atrybutem jest większy, niż być powinien. Ponieważ w praktyce atrybuty bardzo często są ze sobą powiązane (np. wykształcenie wpływa na dochód, wciąż też występuje statystyczna zależność pomiędzy płcią a zawodem), ignorujący je klasyfikator Bayesa nazywa się naiwnym.

Naiwność klasyfikatora Bayesa wynika również z tego, że gdy pewna zależność nie wystąpiła w przypadkach treningowych (np. dotychczas w sobotę butik odwiedzały wyłącznie kobiety), obliczone przez niego prawdopodobieństwo, że klientem butiku w sobotę będzie mężczyzna, wyniesie 0%. Problem ten możemy rozwiązać, dodając 1 do wszystkich przyporządkowań stanów atrybutów do klas wyjściowych, czyli stosując estymację Laplace’a.

Obliczanie wyniku poprzez mnożenie prawdopodobieństw ma jeszcze jedną wadę. Jeżeli te prawdopodobieństwa są bardzo małe, co ma miejsce, gdy lista atrybutów jest długa i gdy atrybuty przyjmują wiele stanów, błędy ich zaokrąglania zaczynają wpływać na wyniki.

² $P(A)$ oznacza prawdopodobieństwo a priori wystąpienia klasy A , tj. prawdopodobieństwo, że przypadek należy do klasy A ; $P(B|A)$ oznacza prawdopodobieństwo a posteriori, że przypadek A należy do klasy B , natomiast $P(B)$ — prawdopodobieństwo a priori wystąpienia przypadku B .

³ Nieprzyjęcie założenia o niezależności zmiennych wejściowych wymagałoby obliczenia k^p prawdopodobieństw, gdzie p jest liczbą zmiennych, a k — liczbą ich stanów. Na przykład dla 30 zmiennych binarnych trzeba by wykonać 2^{30} (1 073 741 824) operacji.

Ograniczenia

Pierwsze ograniczenie wynika ze sposobu działania naiwnego klasyfikatora Bayesa — policzenie prawdopodobieństwa wystąpienia danego stanu jest możliwe tylko dla atrybutów dyskretnych, a więc atrybuty ciągłe są ignorowane przez naiwny klasyfikator Bayesa firmy Microsoft.

Drugie ograniczenie jest mniej oczywiste — naiwny klasyfikator Bayesa należy do klasyfikatorów liniowych i nie nadaje się do rozwiązywania problemów nieliniowych, czyli takich, w których stan atrybutu wyjściowego zależy od kombinacji stanów atrybutów wejściowych. Problemem nieliniowym jest np. kwestia określenia koloru pól na szachownicy.

Połowa pól na szachownicy jest biała, druga połowa — czarna. Czy znając kolumnę i wiersz, jesteśmy w stanie określić kolor pola znajdującego się na ich przecięciu? Spróbujmy użyć naiwnego klasyfikatora Bayesa firmy Microsoft do znalezienia odpowiedzi na to pytanie.

1. Otwórz przykładowy skoroszyt Excela i przejdź do arkusza *Chessboard*.
2. Zaznacz znajdującą się w nim tabelę. Jej pierwsza kolumna zawiera litery kolumn, druga — numery wierszy, a trzecia kolory pól szachownicy.
3. Kliknij znajdujący się w sekcji *Data Modeling* przycisk *Classify*.
4. Jako parametr wyjściowy i wejściowy wybierz *Color*, a jako użyty do klasyfikacji algorytm wybierz *Microsoft Naive Bayes*.
5. Przeznacz wszystkie dane do treningu i zakończ działanie kreatora, tworząc tymczasowy model eksploracji danych.

Okaże się, że algorytm nie znalazł żadnych zależności pomiędzy kolumną i wierszem pola na szachownicy a kolorem pola znajdującego się na ich przecięciu — wszystkie zakładki wizualizatora będą puste, z wyjątkiem zakładki *Dependency Network*, w której znajdziemy wyłącznie wyjściowy atrybut *Color*.

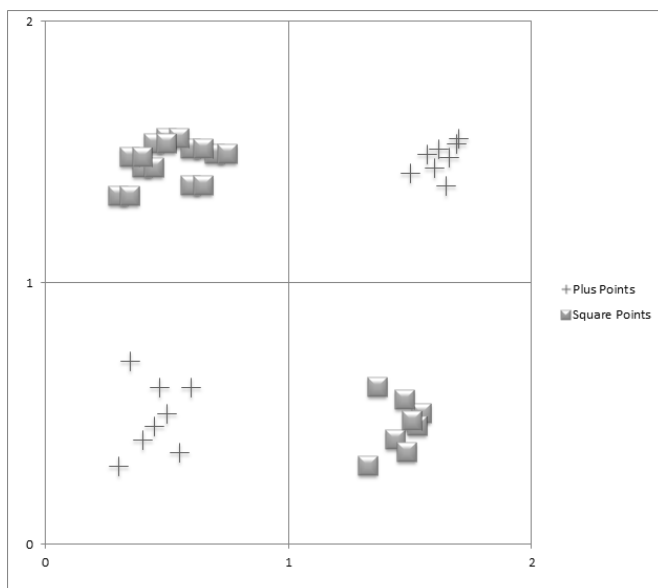
Zastanówmy się, od czego zależy kolor pól szachownicy. Czy zależy on od wierszy? Nie, w każdym wierszu 50% pól jest czarnych, a 50% białych. Nie zależy on również od kolumn, lecz od kombinacji wierszy i kolumn. Ponieważ naiwny klasyfikator Bayesa jest klasyfikatorem liniowym, nie znalazł powyższych zależności nieliniowych.

Tak postawiony problem nie zostałby rozwiązany również przez drzewa decyzyjne, czyli klasyfikator nieliniowy — w każdym wierszu i w każdej kolumnie białych pól jest dokładnie tyle samo co czarnych. Różnicę pomiędzy klasyfikatorami liniowymi i nieliniowymi pokazuje kolejny przykład. Tym razem kształt figury również nie zależy od jego poszczególnych współrzędnych, ale od ich kombinacji (rysunek 9.1).

1. Przejdź do arkusza *Linear*.
2. Przeprowadź klasyfikację znajdujących się w nim danych, wybierając na atrybuty wejściowe kolumny *RangeX*, *RangeY* i *Shape*, a na atrybut wyjściowy kolumnę *Shape*.
3. Jako użyty do klasyfikacji algorytm wybierz *Microsoft Naive Bayes*.

Rysunek 9.1.

W pierwszej i trzeciej ćwiartce znajdują się wyłącznie krzyżyki, w drugiej i czwartej — same kwadraty



4. Przeznacz wszystkie dane do treningu i zakończ działanie kreatora, tworząc tymczasowy model eksploracji danych.

Tym razem algorytm również nie znajdzie żadnych zależności pomiędzy współrzędnymi a kształtem figur.

Pomimo tych ograniczeń naiwny klasyfikator Bayesa firmy Microsoft dobrze radzi sobie z wykrywaniem zależności pomiędzy poszczególnymi atrybutami, a jego prostota (i związane z nią szybkość oraz małe zapotrzebowanie na pamięć i moc obliczeniową), jak również łatwość interpretacji wyników czynią z niego przydatny i często używany algorytm eksploracji danych.

Parametry

Naiwny klasyfikator Bayesa firmy Microsoft przyjmuje następujące parametry:

1. **MAXIMUM_INPUT_ATTRIBUTES** — parametr dostępny tylko w edycji Enterprise; określa maksymalną liczbę atrybutów wejściowych (objaśniających). Po jej przekroczeniu (domyślna wartość wynosi 255) analizowanych będzie tylko 255 atrybutów wejściowych najsilniej powiązanych z atrybutami wyjściowymi (objaśnianymi). Zmiana tego parametru na 0 spowoduje uwzględnienie wszystkich atrybutów wejściowych⁴.

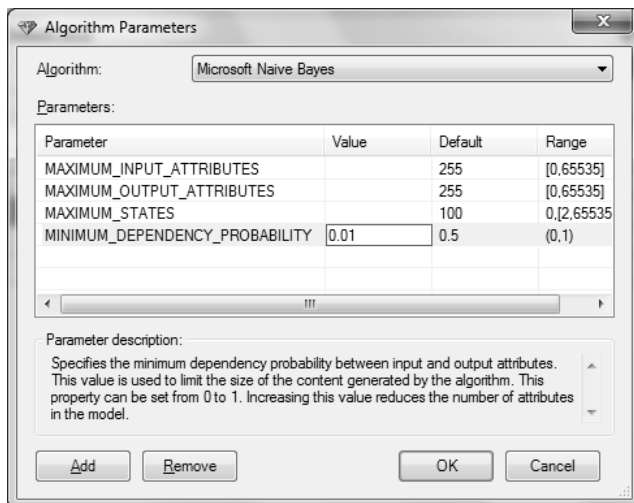
⁴ Maksymalna liczba atrybutów wynosi 65 535 i jest tak duża, że w praktyce nie spotkamy się z wynikającymi z niej ograniczeniami. Z pierwszej części książki wiadomo, że dane wejściowe powinny zawierać jak najwięcej informacji (a dokładnie, że entropia atrybutów wejściowych względem wyjściowych powinna być jak największa), tymczasem utworzenie kilkudziesięciu tysięcy atrybutów raczej zmniejszyłoby (a nie zwiększyło) ilość tych informacji. Ponadto dane właściwie reprezentujące wszystkie możliwe zależności pomiędzy tymi atrybutami byłyby liczone w milionach terabajtów.

2. `MAXIMUM_OUTPUT_ATTRIBUTES` — parametr dostępny tylko w edycji Enterprise; określa maksymalną liczbę atrybutów wyjściowych. Po jej przekroczeniu (domyślna wartość wynosi 255) zostanie uwzględnionych tylko 255 najczęściej występujących atrybutów wyjściowych. Zmiana tego parametru na 0 spowoduje uwzględnienie wszystkich atrybutów wyjściowych.
3. `MAXIMUM_STATES` — parametr dostępny tylko w edycji Enterprise; określa maksymalną liczbę uwzględnianych stanów atrybutów. Po jej przekroczeniu (domyślna wartość wynosi 100) analizowanych będzie tylko 100 najczęściej występujących stanów atrybutów, a pozostałe zostaną potraktowane jak wartości brakujące. Zmiana tego parametru na 0 spowoduje uwzględnienie wszystkich stanów atrybutów.
4. `MINIMUM_DEPENDENCY_PROBABILITY` — określa (w skali od 0 do 1) minimalne prawdopodobieństwo znalezienia zależności pomiędzy atrybutami wejściowymi a wyjściowymi. Zmiana tego parametru nie ma żadnego wpływu na trening algorytmu, a jedynie na liczbę zwracanych (znalezionych) zależności. Domyślna wartość wynosi 0,5 — jest to wartość, przy której wizualizatory tego algorytmu zwracają informacje tylko o zależnościach, których prawdopodobieństwo wystąpienia jest większe od prawdopodobieństwa ich braku.

Żeby przekonać się, jak zmiana parametru `MINIMUM_DEPENDENCY_PROBABILITY` wpłynie na zdolność naiwnego klasyfikatora Bayesa firmy Microsoft do rozwiązywania problemów nieliniowych, raz jeszcze przeprowadź klasyfikację kolorów pól na szachownicy, tym razem ustawiając wartość tego parametru na 0,01 (rysunek 9.2).

Rysunek 9.2.

*Naiwny klasyfikator
Bayesa firmy
Microsoft to prosty
algorytm eksploracji
danych; jego
działaniem możemy
sterować w bardzo
ograniczonym zakresie*



Zgodnie z oczekiwaniami obniżenie wartości tego parametru nie wpłynęło na otrzymane wyniki — algorytm nadal nie jest w stanie znaleźć żadnych zależności pomiędzy kolumną i wierszem pola na szachownicy a jego kolorem.

Zastosowania naiwnego klasyfikatora Bayesa

„Naiwność” klasyfikatora Bayesa ogranicza jego stosowanie w modelach klasyfikacyjnych, ale w żaden sposób nie zmniejsza jego wartości dla modeli opisowych. W szczególności jego szybkość i małe wymagania dotyczące pamięci czynią z niego doskonałe narzędzie do oceny danych wejściowych.

Drugi z opisanych poniżej przykładów demonstruje predykcyjne możliwości naiwnego klasyfikatora Bayesa — jeżeli tylko atrybuty wejściowe rzeczywiście są od siebie niezależne lub ewentualne zależności między nimi są nieistotne w ramach przyjętego modelu (jak ma to miejsce np. podczas oceniania wiadomości e-mail na podstawie poszczególnych słów, czy jest ona spamem), algorytm ten okazuje się szybkim i dokładnym klasyfikatorem.



Wskazówka

W serwerze SQL klasyfikator Bayesa firmy Microsoft stosowany jest do klasyfikacji i — z pewnymi ograniczeniami — asocjacji.

Badanie zależności pomiędzy atrybutami

Naiwny klasyfikator Bayesa firmy Microsoft doskonale nadaje się (o czym powiedziano w rozdziale 3.) do analizowania zależności pomiędzy atrybutami. W tym punkcie utworzymy model analizujący zależności pomiędzy atrybutami klientów firmy Adventure Works:

1. Uruchom konsolę SSMS i połącz się z serwerem SSAS.
2. Zaznacz bazę analityczną DataMining i wyświetl okno edytora DMX.
3. Utwórz w tej bazie poniższy model eksploracji danych (tworząc model za pomocą instrukcji CREATE MINING MODEL, automatycznie utworzymy strukturę o nazwie tworzonego modelu, uzupełnioną o sufiks _Structure):

```
CREATE MINING MODEL CustomersAnalysis (
    [ID] LONG KEY,
    [Age] LONG DISCRETIZED(CLUSTERS,5),
    [MaritalStatus] TEXT DISCRETE PREDICT,
    [Gender] TEXT DISCRETE PREDICT,
    [TotalChildren] LONG DISCRETE PREDICT,
    [NumberChildrenAtHome] LONG DISCRETE PREDICT,
    [Education] TEXT DISCRETE PREDICT,
    [Occupation] TEXT DISCRETE PREDICT,
    [YearlyIncome] LONG DISCRETIZED(CLUSTERS,8),
    [HouseOwnerFlag] TEXT DISCRETE PREDICT,
    [NumberCarsOwned] LONG DISCRETE PREDICT,
    [TotalAmount] LONG DISCRETIZED(CLUSTERS,8) PREDICT,
    [TotalQuantity] LONG DISCRETE PREDICT,
    [BikesQuantity] LONG DISCRETE PREDICT,
    [BikesAmount] LONG DISCRETIZED(CLUSTERS,8) PREDICT,
```

```

[ClothingQuantity]    LONG DISCRETE PREDICT,
[ClothingAmount]     LONG DISCRETIZED(CLUSTERS,8) PREDICT,
[AccessoriesQuantity] LONG DISCRETE PREDICT,
[AccessoriesAmount]  LONG DISCRETIZED(CLUSTERS,8) PREDICT,
[MonthsAsCustomer]   LONG DISCRETIZED(CLUSTERS,10) PREDICT )
USING Microsoft_Naive_Bayes

```

Zwróć uwagę, że wszystkie atrybuty są dyskretne lub poddane dyskretyzacji oraz że wszystkie one zostały użyte w roli atrybutów wejściowych i wyjściowych.

Utwórz, np. korzystając z dołączonego do książki skryptu XMLA, źródło danych Adventure Works DW i skonfiguruj nazwę i hasło użytkownika, z którego uprawnieniami serwer SSAS będzie łączył się z tym źródłem danych, a następnie przetwórz ten model, wykonując poniższą instrukcję:

```

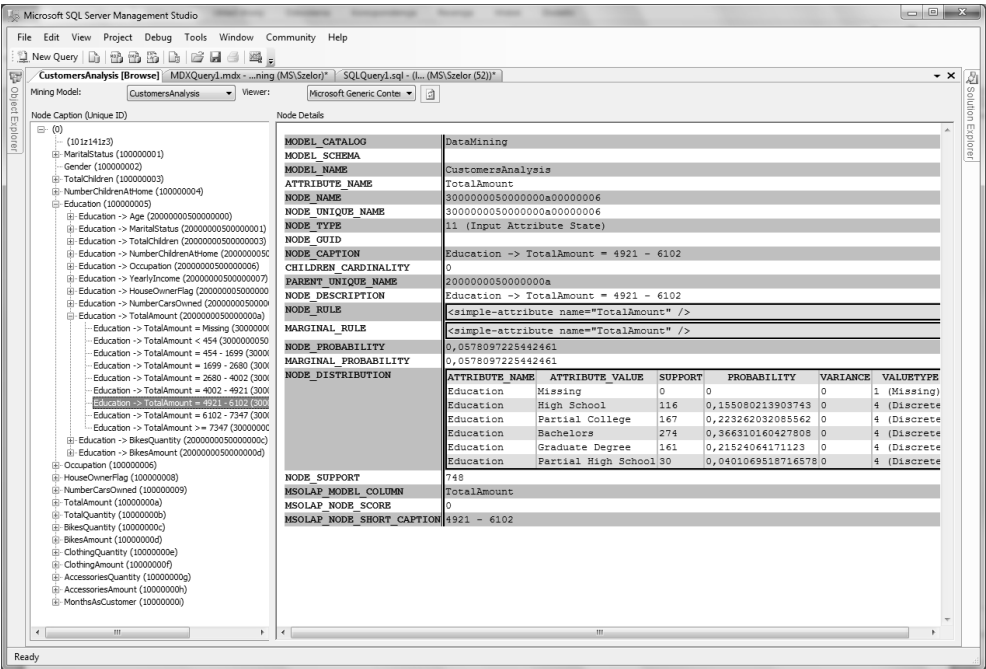
INSERT INTO CustomersAnalysis([ID], [Age], [MaritalStatus], [Gender], [TotalChildren]
,[NumberChildrenAtHome], [Education], [Occupation], [YearlyIncome], [HouseOwnerFlag]
,[NumberCarsOwned], [TotalAmount], [TotalQuantity], [BikesQuantity], [BikesAmount]
,[ClothingQuantity], [ClothingAmount], [AccessoriesQuantity], [AccessoriesAmount]
,[MonthsAsCustomer])
OPENQUERY ([Adventure Works DW], 'SELECT [ID], [Age], [MaritalStatus],
[Gender], [TotalChildren], [NumberChildrenAtHome], [Education], [Occupation],
[YearlyIncome], [HouseOwnerFlag], [NumberCarsOwned], [TotalAmount], [TotalQuantity],
[BikesQuantity], [BikesAmount], [ClothingQuantity], [ClothingAmount],
[AccessoriesQuantity], [AccessoriesAmount] ,[MonthsAsCustomer]
FROM [dbo].[CustomersHistoryTrain]')

```

Po wyświetleniu raportu *Dependency Network* (żeby wyświetlić okno z wizualizatorami bezpośrednio z konsoli SSMS, należy kliknąć model eksploracji danych i wybrać z menu kontekstowego *Browse*) przekonamy się, że używany w poprzednich modelach w roli atrybutu wyjściowego atrybut *TotalAmount* jest silnie powiązany nie tylko z atrybutami opisującymi klientów (takimi jak *Age*, *Occupation* czy *TotalChildren*), ale również z atrybutami opisującymi historię zakupów tych klientów (takimi jak *AccessoriesAmount*, *BikesAmount*, *ClothingAmount* czy *TotalQuantity*). Jednak te ostatnie atrybuty są silnie powiązane nie tylko z objaśnianym atrybutem *TotalAmount*, ale również ze sobą nawzajem. Z rozdziału 5. wiadomo, że w modelach klasyfikacyjnych nie należy używać w roli atrybutów wejściowych silnie powiązanych ze sobą atrybutów, dlatego atrybuty te nie były używane w utworzonych wcześniej modelach.

Raport zależności nie zawiera informacji na temat stanów poszczególnych atrybutów. Te dane znajdziemy w pozostałych raportach wizualizatora naiwnego klasyfikatora Bayesa firmy Microsoft lub odczytując strukturę modelu. Wizualizator każdego algorytmu eksploracji danych można zastąpić ogólnym wizualizatorem *Microsoft Generic Content Tree Viewer*, zwracającym informację na temat struktury modelu.

Wyświetl go, a następnie z listy węzłów modelu wybierz węzeł opisujący zależności pomiędzy atrybutem *Education* a poszczególnymi stanami atrybutu *TotalAmount* (rysunek 9.3).



Rysunek 9.3. Szczegółowe informacje na temat modeli eksploracji danych wraz z ich formatowaniem można skopiować do schowka i wkleić np. do dokumentu Worda

Modele naiwnego klasyfikatora Bayesa firmy Microsoft liczą tyle węzłów drugiego poziomu (węzłów typu 9.), ile jest zdefiniowanych atrybutów wejściowych (węzłem pierwszego poziomu jest sam model eksploracji danych). Listę tych węzłów wraz z ich identyfikatorami można odczytać, wywołując poniższą procedurę:

```
CALL GetPredictableAttributes ('CustomersAnalysis')
```

Na trzecim poziomie znajdują się węzły zawierające atrybuty wejściowe (węzły typu 10.), a na czwartym (w węzłach typu 11.) — znalezione zależności pomiędzy poszczególnymi atrybutami wejściowymi a atrybutem wyjściowym, nadrzędnym dla danego węzła.

Klasyfikacja dokumentów

Analiza dokumentów tekstowych wymaga ich wcześniejszego podzielenia na frazy — to występowanie lub brak w dokumencie poszczególnych fraz będzie podstawą ich oceny. Analiza dokumentów tekstowych przypomina więc analizę koszykową: koszyki zakupów analizowane są pod kątem występowania w nich poszczególnych towarów, dokumenty tekstowe — pod kątem występowania w nich poszczególnych fraz.

Podzielone na frazy dokumenty mogą być:

1. Klasyfikowane — frazy zapisane w tabeli zagnieżdżonej będą podstawą zaklasyfikowania dokumentu np. jako spam.

2. Dzielone na segmenty na podstawie częstotliwości występowania w nich poszczególnych fraz.
3. Kojarzone ze sobą na podstawie występujących w nich fraz.

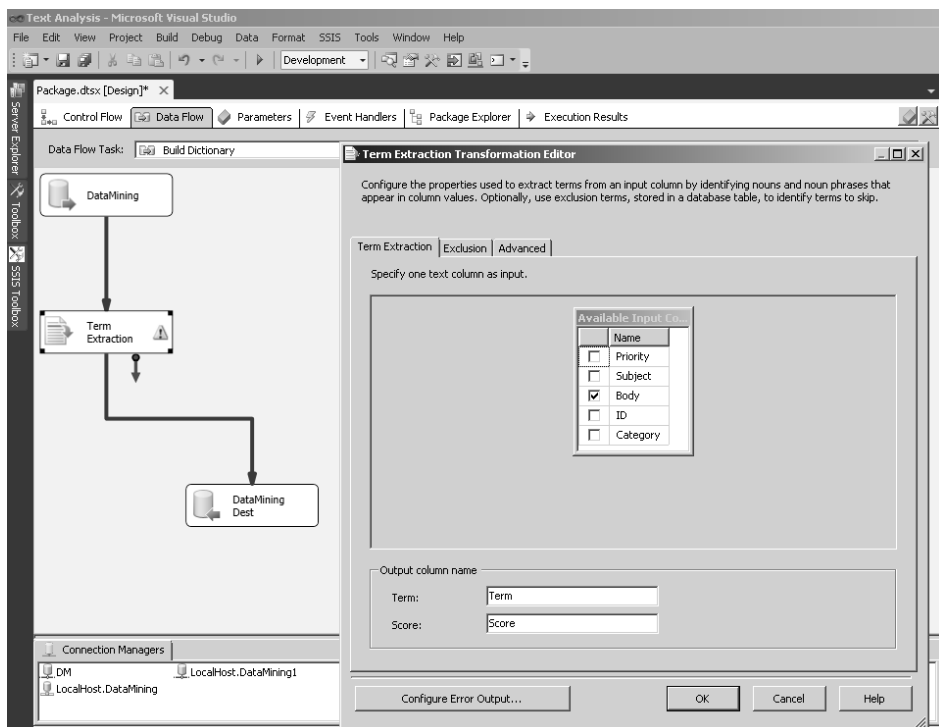
W tym punkcie przeprowadzimy klasyfikację wiadomości e-mail. Wymaga to:

1. Zbudowania słownika zawierającego wszystkie frazy występujące w tych dokumentach.
2. Dekompozycji poszczególnych dokumentów na frazy zapisane w słowniku.
3. Zbudowania modelu klasyfikacyjnego.
4. Rozłożenia ocenianych dokumentów na frazy i sklasyfikowania ich przy użyciu zbudowanego modelu.

Do zbudowania słownika użyjemy transformacji *Term Extraction* usługi SSIS:

1. Uruchom Business Intelligence Development Studio, utwórz nowy projekt typu Integration Services i nazwij go *Text Analysis*.
2. Dodaj do pakietu SSIS zadanie *Data Flow Task* i nazwij je *Build Dictionary*.
3. Przejdź na zakładkę *Data Flow*.
4. Dodaj do zadania *Build Dictionary* transformację *ADO.NET Source* i pobierz za jego pomocą dane z tabeli *EMails*, znajdującej się w przykładowej bazie danych *DataMiningDW*.
5. Dodaj do tego zadania transformację *Term Extraction* i połącz ją z domyślnym (zielonym) wyjściem utworzonego źródła danych.
6. Dwukrotnie kliknij tę transformację — wyświetli się okno edytora *Term Extraction Transformation Editor*:
 - a) Na zakładce *Term Extraction* wybierz kolumnę, w której przechowywane są treści wiadomości e-mail (rysunek 9.4).
 - b) Zakładka *Exclusion* pozwala wskazać tabelę zawierającą frazy wykluczone ze słownika.
 - c) Przejdź na zakładkę *Advanced*. Pozwala ona skonfigurować sposób rozkładania tekstu na frazy: m.in. to, czy ma on być dzielony na pojedyncze wyrazy lub tylko na frazy, wybrać sposób oceniania fraz (mogą być one oceniane tylko na podstawie częstotliwości występowania w danym dokumencie oraz z uwzględnieniem tego, jak często fraza występowała we wszystkich dokumentach⁵), minimalną liczbę wystąpień fraz oraz ich maksymalną długość w słowach.
7. Zamknij okno edytora przyciskiem *OK*.

⁵ Ocena frazy jest tym wyższa, im częściej występuje ona w dokumencie, ale metoda TFIDF dodatkowo obniża oceny fraz często występujących we wszystkich dokumentach.

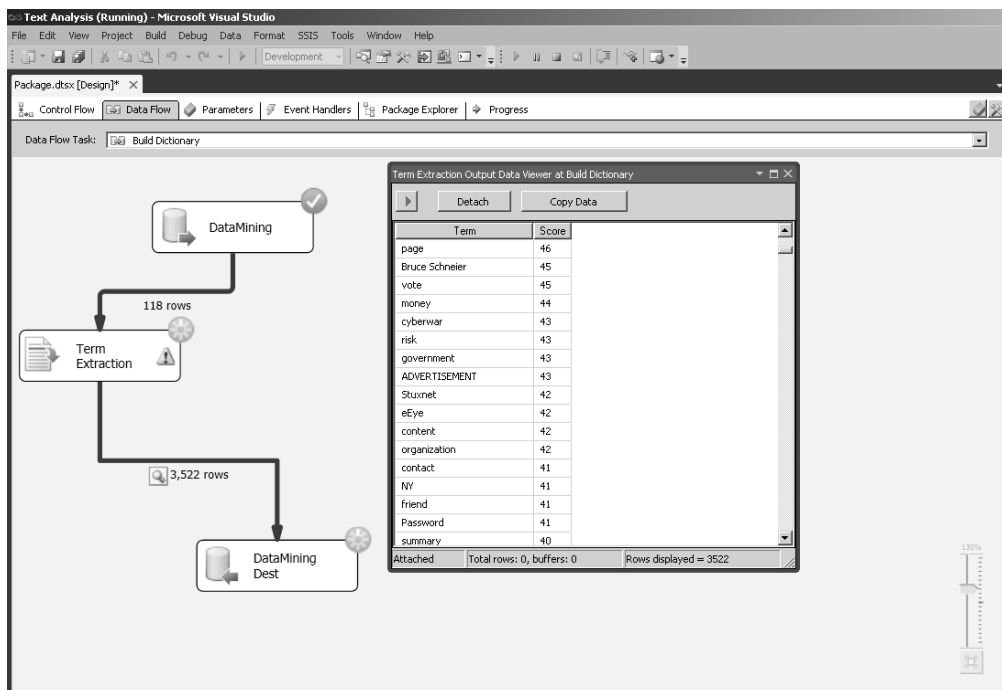


Rysunek 9.4. Wynikiem transformacji będą dwie nowe kolumny: w kolumnie o domyślnej nazwie *Term* zostaną zapisane frazy znaczeniowe, w kolumnie *Score* — punkty obliczone na podstawie częstotliwości ich występowania

8. Dodaj do zadania transformację *ADO.NET Destination* i utwórz za pomocą tego zadania w bazie danych DataMining tabelę Dictionary, w której zapisane zostaną frazy i ich oceny.
9. Uruchom pakiet SSIS (rysunek 9.5).

Po zbudowaniu słownika możemy rozłożyć poszczególne wiadomości e-mail na frazy:

1. Przejdź do zakładki *Control Flow*, dodaj do pakietu kolejne zadanie *Data Flow Task* i nazwij je *Decompose Documents*.
2. Połącz zadanie *Build Dictionary* z zadaniem *Decompose Documents* — w ten sposób najpierw zostanie utworzony słownik, który następnie zostanie użyty do dekompozycji wiadomości e-mail.
3. Kliknij dwukrotnie to zadanie lewym przyciskiem myszy — wyświetli się ono w edytorze przepływu danych.
4. Dodaj do zadania *Decompose Documents* transformację *ADO.NET Source* i pobierz za jego pomocą dane z tabeli *EMails* znajdującej się w przykładowej bazie danych DataMiningDW.
5. Dodaj do tego zadania transformację *Term Lookup* i połącz ją z domyślnym (zielonym) wyjściem utworzonego źródła danych.



Rysunek 9.5. Pakiet SSIS tworzący słownik fraz występujących w wiadomościach e-mail (frazy zostały pokazane poprzez kliknięcie strzałki łączącej dwa ostatnie zadania i wybranie opcji *Enable Data Viewer*)

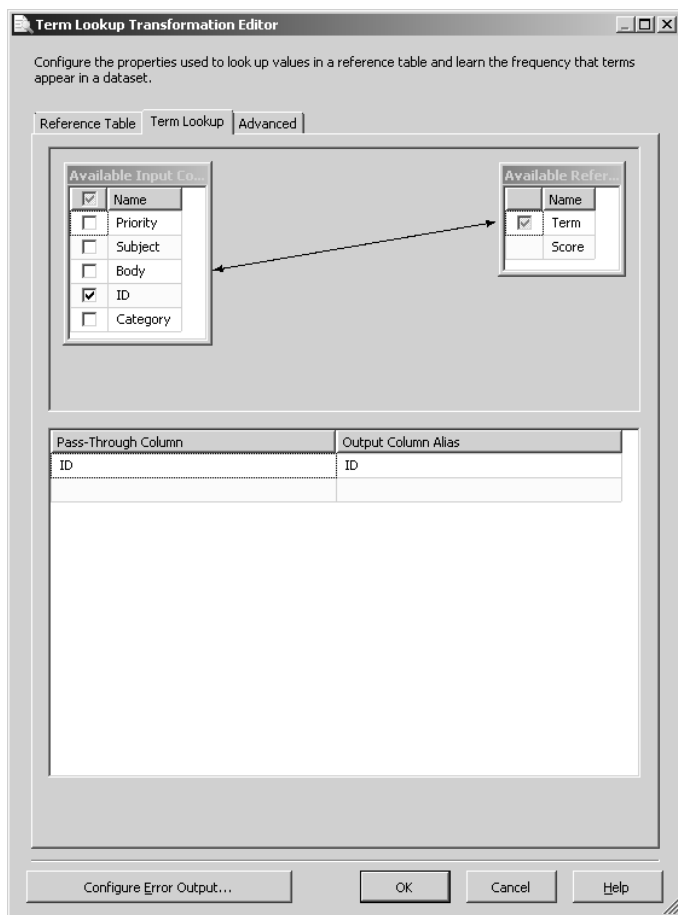
6. Dwukrotnie kliknij tę transformację — wyświetli się okno edytora Term Lookup Transformation Editor:
 - a) Zakładka *Reference Table* pozwala wskazać tabelę słownikową — połącz się z bazą DataMiningDW i wybierz tabelę Dictionary.
 - b) Przejdź na zakładkę *Term Lookup* i połącz kolumnę Body tabeli Emails z kolumną Term tabeli Dictionary. Ponieważ tabela utworzona za pomocą tej transformacji będzie musiała zostać powiązana z nadrzędną tabelą Emails, dodaj do jej wyniku zawartość kolumny ID (rysunek 9.6).
 - c) Zatwierdź zmiany przyciskiem *OK*.
7. Dodaj do zadania transformację *ADO.NET Destination* i utwórz za jego pomocą w bazie danych DataMinigDW tabelę EmailsFragments, w której zapisane zostaną zdekomponowane wiadomości e-mail.

Żeby ponowne uruchomienie pakietu nie powodowało duplikowania wierszy zapisanych w tabelach Dictionary i EmailsFragments:

1. Dodaj do niego zadanie *Execute T-SQL Statement Task*.
2. Połącz je z lokalnym serwerem SQL.

Rysunek 9.6.

Transformacja Term Lookup pozwoli nam zapisać w tabeli podrzędnej fraz informacje o tym, ile razy wystąpiły one w każdym dokumencie, oraz identyfikatory dokumentów, w których te frazy zostały znalezione



3. W polu *T-SQL Statement* wpisz poniższe instrukcje:

```
USE DataMiningDW
GO
IF EXISTS (SELECT * FROM sys.tables WHERE name='Dictionary')
TRUNCATE TABLE dbo.Dictionary
GO
IF EXISTS (SELECT * FROM sys.tables WHERE name='EmailsFragments')
TRUNCATE TABLE dbo.EmailsFragments
GO
```

4. Połącz to zadanie z zadaniem Build Dictionary.

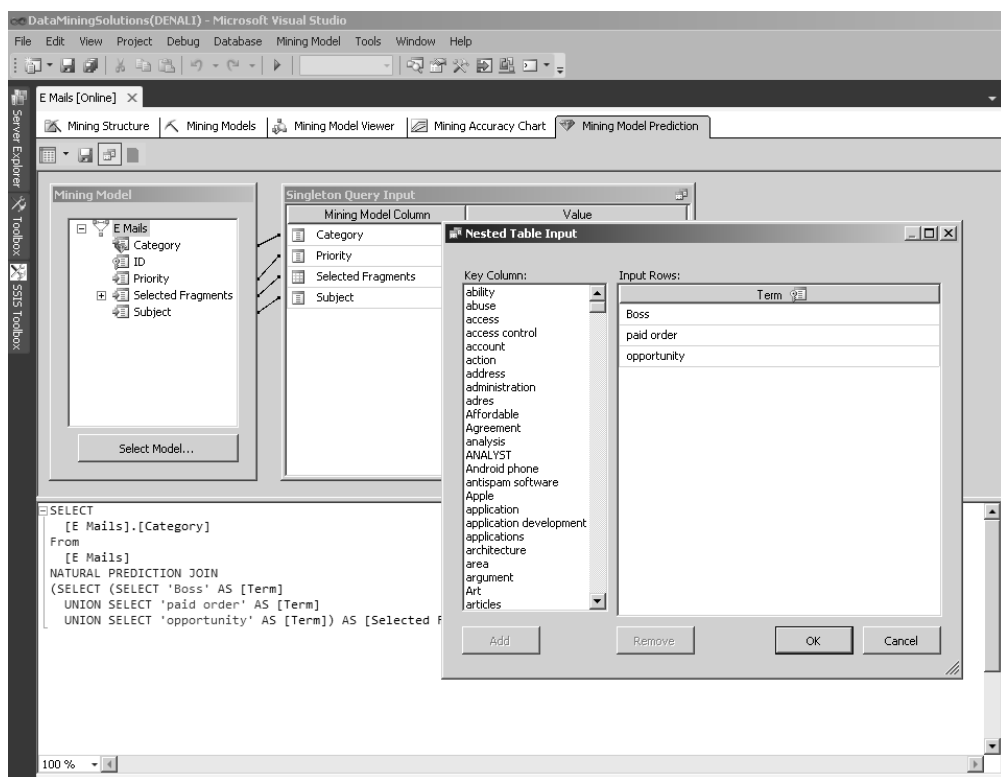
5. Uruchom i zapisz gotowy pakiet SSIS.

Dysponując przygotowanymi w ten sposób danymi źródłowymi, możemy już zbudować model klasyfikujący dokumenty. Nasz model będzie je klasyfikował wyłącznie na podstawie tematów i priorytetów wiadomości oraz znajdujących się w nich fraz — to, ile razy występuje w nich dana fraza, pominiemy. Dla odmiany model ten utworzymy w środowisku BIDS:

1. Połącz się z analityczną bazą danych DataMining.
2. Utwórz nowy widok danych źródłowych i dodaj do niego tabele Emails i EmailsFragments.
3. Połącz te tabele, przeciągając kolumnę ID tabeli EmailsFragments do kolumny ID tabeli Emails.
4. Analizując przykładowe dane, zwróć uwagę, że zaledwie 371 (1,5%) fraz pochodzi z wiadomości oznaczonych jako spam. Ponieważ nasz model ma klasyfikować dokumenty, musimy zmienić rozkład atrybutu wyjściowego, sztucznie zwiększając częstotliwość występowania fraz wskazujących na niechciane wiadomości:
 - a) Dodaj do widoku danych źródłowych nazwane zapytanie SelectedFragments.
 - b) Odczytaj w tym zapytaniu wszystkie fragmenty niechcianych wiadomości uzupełnione o 2% losowo wybranych fragmentów pozostałych wiadomości:

```
SELECT E.ID, Term, Frequency, NEWID() as n
FROM dbo.EmailsFragments AS F
JOIN dbo.Emails AS E ON E.ID=F.ID
WHERE E.Category='SPAM'
UNION ALL
SELECT TOP 2 PERCENT E.ID, Term, Frequency, NEWID()
FROM dbo.EmailsFragments AS F
JOIN dbo.Emails AS E ON E.ID=F.ID
WHERE E.Category<>'SPAM'
ORDER BY NEWID();
```
 - c) Połącz utworzone zapytanie z tabelą Emails, przeciągając jego kolumnę ID do kolumny ID tabeli Emails.
5. Zapisz zmiany i zamknij edytor widoku danych źródłowych.
6. Z wykorzystaniem kreatora utwórz nową strukturę i model eksploracji danych:
 - a) Pobierz dane z tabel relacyjnej bazy danych.
 - b) Wybierz naiwny klasyfikator Bayesa firmy Microsoft.
 - c) Wskaż widok danych źródłowych utworzony w poprzednich punktach.
 - d) Na tabelę nadrzędną (ang. *Case*) wybierz tabelę Emails, na tabelę zagnieżdżoną (ang. *Nested*) — nazwane zapytanie SelectedFragments.
 - e) Zaznacz kolumnę klucza zagnieżdżonego przypadku (kolumnę Term), dodaj do listy wejściowych atrybutów kolumny Category, Priority i Subject, a na atrybut wyjściowy wybierz kolumnę Category.
 - f) Użyj wszystkich danych jako przypadków treningowych.
 - g) Zwróć uwagę, że naiwny klasyfikator Bayesa firmy Microsoft nie umożliwia zaznaczenia opcji *Allow drill through* (przedstawiona w poprzednim punkcie struktura modeli tego algorytmu jest mało intuicyjna i nie pozwala w prosty sposób powiązać znalezionych zależności z poszczególnymi przypadkami). Zakończ pracę kreatora.

7. Przetwórz zbudowany model. Wyświetlą się dwa ostrzeżenia:
 - a) *Automatic feature selection has been applied to model, due to the large number of attributes. Set MAXIMUM_INPUT_ATTRIBUTES and/or MAXIMUM_OUTPUT_ATTRIBUTES to increase the number of attributes considered by the algorithm* — to ostrzeżenie dotyczy zagnieżdżonego atrybutu Term.
 - b) *Cardinality reduction has been applied on column, Subject of model, E Mails due to the large number of states in that column. Set MAXIMUM_STATES to increase the number of states considered by the algorithm* — to ostrzeżenie dotyczy atrybutu Subject.
8. Wyświetl okno właściwości algorytmu i ustaw na 0 wartości parametrów MAXIMUM_INPUT_ATTRIBUTES oraz MAXIMUM_STATES.
9. Ponownie przetwórz model eksploracji danych i zapoznaj się z jego wynikami.
10. Przejdź na zakładkę *Data Mining Prediction* i wykonaj zapytanie czasu rzeczywistego, oceniające, czy e-mail zawierający wybrane frazy będzie sklasyfikowany jako spam (rysunek 9.7).



Rysunek 9.7. Business Intelligence Development Studio pozwala wybrać z listy wartości zagnieżdżonego atrybutu *te*, których chcemy użyć w zapytaniach czasu rzeczywistego

Skorowidz

A

- abstrakcja, 30
- adaptacyjny interfejs, 406
- AdventureWorksDW, 16, 52
- algorytmy
 - CART, 268
 - drzew decyzyjnych, 72, 166
 - ID3, 268
 - klastrowania, 297
 - klastrowania sekwencyjnego, 319
 - odkrywania reguł asocjacyjnych, 335
 - regresji liniowej, 267
 - regresji logistycznej, 351
- anachronizmy, 76
- analiza
 - biznesowa, 35
 - dokumentów tekstowych, 260
 - koszykowa, 142, 335, 343
 - odwiedzin stron WWW, 324
 - sekwencyjna, 151
 - skupień komórek, 305
 - typu cross-selling, 347
 - wariantowa, 111, 152, 295
 - zależności pomiędzy atrybutami, 73, 258, 342
- anomalie, 149, 314, 332
- aplikacje inteligentne, 401
- architektura SSAS, 219
- asocjacja, 141, 177, 279
- atrybuty, 57, 227
 - bez wartości, 237
 - ciągłe, 57, 86
 - dyskretne, 57, 81
 - grupowanie, 81
 - jednowartościowe, 57
 - monotoniczne, 59
 - nadmiarowe, 75
 - niezależne, 74
 - okresowe, 86

- porządkowe, 85
- rozkład wartości, 59
- stany, 57
- tekstowe, 68
- wielowartościowe, 57
- zależności, 73, 258, 342
- AUTO_DETECT_PERIODICITY, 286

B

- Bayesa naiwny klasyfikator, 72, 253
 - analiza dokumentów tekstowych, 260
 - analiza zależności pomiędzy atrybutami, 258
 - ograniczenia, 255
 - omówienie, 253
 - parametry, 256
 - zastosowania, 258
- bazy danych
 - AdventureWorksDW, 16
 - DataMiningDW, 17
 - DataMiningSolutions, 19
- bezpieczeństwo SSAS, 221
- Bias, 50
- BIDS, 162, 192
 - interfejs, 193
 - tryb offline, 194
 - tryb online, 194
- bity informacji, 77
- błędy
 - pomiaru, 50
 - przypadkowy, 51
 - systematyczny, 50
- brakujące dane, 69
- uzupełnienie, 404
- Breiman, Leo, 268
- Business Intelligence Development Studio,
Patrz BIDS

C

cele

- eksploracji danych, 35
- modelowania, 35
- ciągłe atrybuty, 57, 86
- CLUSTER_COUNT, 303, 323
- CLUSTER_SEED, 303
- CLUSTERING_METHOD, 304
- Co będzie, jeżeli?, 155
- COMPLEXITY_PENALTY, 273, 286
- CRISP-DM, 11
- cross-selling, 347
- czynniki kluczowe, 128, 129

D

dane

- brakujące, 69, 404
- diagnostyczne, 115
- dla modeli deskrypcyjnych, 108
- dla modeli predykcyjnych, 109
- duplikaty, 62
- integralność, 62
- kontrola poprawności, 401
- korelacja, 106
- modelowanie, 27
- na potrzeby analizy wariantowej, 111
- ocena, 49, 121
- oczyszczanie, 122
- odchylenie standardowe, 67
- opisywanie, 36
- podział, 124
- poprawa jakości, 99
- profilowanie, 54
- próbkiowanie, 64
- przygotowanie, 79
 - do dalszej eksploracji, 312
- reprezentatywność, 64
- serie, 92
 - krótkie, 293
 - przeplatane, 291
- testowe, 113
- treningowe, 114
 - filtrowanie, 209
- wyniki eksploracji, 42
- wzbogacenie, 103
- zakres wartości, 63
- zbieżność, 65
- zewnętrzne, 162
- zgodność ze wzorcem, 63
- źródła, 195, 240
- źródłowe, 40, 49, 121
 - nieprzygotowane, 393

- niewłaściwe, 392
- widoki, 196

Data Mining, 15, 121, 162

- analiza koszykowa, 142
- dane źródłowe, 121
- instalacja, 15
- ocena danych, 121
- oczyszczanie danych, 122
- podział danych, 124

Data Mining Extensions to SQL, *Patrz* DMX

Data Profile Viewer, 55

DataMiningDW, 17

DataMiningSolutions, 19

decydenci, 38

decyzje

- mapa, 37
- modelowanie, 31
- typy, 36
- wspomaganie, 25, 36

deskrypcyjne modele eksploracji danych, 43, 108

diagnostyczne dane, 115

diagramy Ishikawy, 40

DMCONTENTQUERY, 191

DMPREDICT, 191

DMPREDICTTABLEROW, 191

DMX, 227

- funkcje predykcyjne, 251
- modele eksploracji danych, 235
 - odczytywanie zawartości, 243
 - przetwarzanie, 239
- składnia, 232
- struktury eksploracji danych, 233
 - odczytywanie zawartości, 243
 - przetwarzanie, 239
- wstawianie wierszy, 242, 243
- wybieranie przypadków, 239
- zagnieżdżanie przypadków, 236
- zagnieżdżanie tabel, 234
- zapytania predykcyjne, 245
- źródła danych, 240

dokładność predykcji modeli eksploracji danych, 374

dokumenty tekstowe, 260

drzewa decyzyjne, 72, 166, 267

- asocjacja, 279
- klasyfikacja, 275
- ograniczenia, 272
- omówienie, 268
- parametry, 273
- szacowanie, 277
- zastosowania, 275

Dudek, Daniel, 398

duplikaty, 62

dyskretne atrybuty, 57, 81

dyskretyzacja, 90

E

eksploracja danych, 9, 25, 32, 117
cele, 35
dane źródłowe, 40
etapy, 10
formułowanie problemu, 33
hipotezy, 32
kontekst, 40
modele, 182, 184, 206, 232, 235
dane brakujące, 70
deskrypcyjne, 43
dokładność predykcji, 374
korzystanie, 185
kryteria porównawcze, 371
łatwość interpretacji, 373
ocena, 369, 376
odczytywanie zawartości, 243
poprawa, 369
powrót do średniej, 369
predykcyjne, 43
problemy, 391
przetwarzanie, 210, 220, 239
przydatność, 375
skalowalność, 375
wiarygodność predykcji, 374
wizualizatory, 398
wstawianie wierszy, 242
wydajność, 375
zarządzanie, 190
zarządzanie poprzez SSMS, 216
narzędzia, 162
ocena ryzyka, 45
proces, 10, 11
serwer SQL, 218
struktury, 182, 199, 231
odczytywanie zawartości, 243
przetwarzanie, 204, 220, 239
wstawianie wierszy, 242
sukces projektu, 44
techniki, 119, 126
wyniki, 42
zakres projektu, 39
zastosowania, 119
entropia, 78
etapy eksploracji danych, 10
Excel, 15
asocjacja, 177
formuły, 191
grupowanie, 173
jako klient SSAS, 162
klasyfikacja, 163
modele eksploracji danych, 182, 184
narzędzia eksploracji danych, 162
prognozowanie, 179

struktury eksploracji danych, 182
szacowanie, 170
wersja demonstracyjna, 15

F

filtrowanie danych treningowych, 209
FORCE_REGRESSOR, 273
FORECAST_METHOD, 286
formułowanie problemu, 33
formuły arkusza Excel, 191
Friedman, Jerome, 268
funkcje
Co będzie, jeżeli?, 155
predykcyjne, 251
szukania wyniku, 153
uzupełniania, 132, 136
wykrywania anomalii, 149
wykrywania kategorii, 146
Fuzzy Grouping, 82

G

Garbage In, Garbage Out, 49
grupowanie, 81, 145, 173
funkcja wykrywania kategorii, 146
rozmyte, 82

H

HIDDEN_NODE_RATIO, 360
hipotezy, 32
HISTORIC_MODEL_COUNT, 286, 386
HISTORIC_MODEL_GAP, 286, 386
HOLDOUT_PERCENTAGE, 360
HOLDOUT_SEED, 360
Hopfield, John, 352

I

informacje
bity, 77
kontekst, 78
mierzenie, 76
modelowanie, 27
zaskakujące, 77
INSTABILITY_SENSITIVITY, 287
instalacja
Data Mining, 15
serwera SQL, 13
integracja serwera SQL
z SSAS, 223
z SSIS, 226
z SSRS, 226

integralność danych, 62
inteligentne aplikacje, 401
 adaptacyjny interfejs, 406
 kontrola poprawności danych, 401
 uzupełnianie brakujących danych, 404
interfejs adaptacyjny, 406
Ishikawy diagramy, 40

J

jakość danych, 99
jeden do wielu, 84
jednowartościowe atrybuty, 57

K

kalkulator predykcyny, 138
kategorie, 146
klastrowanie, 297
 analiza skupień komórek, 305
 klasyfikacja, 309
 ograniczenia, 302
 omówienie, 297
 parametry, 303
 przygotowanie danych do dalszej eksploracji, 312
 szacowanie, 309
 wykrywanie anomalii, 314
 zastosowania, 305
klastrowanie sekwencyjne, 319
 analiza odwiedzin stron WWW, 324
 klasyfikacja, 327
 ograniczenia, 323
 omówienie, 320
 parametry, 323
 przewidywanie kolejnych zdarzeń, 329
 wykrywanie anomalii, 332
 zastosowania, 324
klasyczna standaryzacja, 89
klasyfikacja, 109, 126, 163, 275, 309, 327, 366
 funkcja uzupełniania, 132
 wykrycie kluczowych czynników, 128, 129
 zapytanie predykcyjne, 134
klasyfikator naiwny Bayesa, 72, 253
 analiza dokumentów tekstowych, 260
 analiza zależności pomiędzy atrybutami, 258
 ograniczenia, 255
 omówienie, 253
 parametry, 256
 zastosowania, 258
klucze, 230
kluczowe czynniki, 128, 129
kłopoty ze sformułowaniem problemu, 33

kodowanie
 jeden do wielu, 84
 wiele do wielu, 85
kontekst
 eksploracji danych, 40
 informacji, 78
kontrola poprawności danych, 401
korelacja danych, 106
korzystanie z modeli eksploracji danych, 185
kostka wielowymiarowa, 292
kryteria porównawcze modeli eksploracji
 danych, 371

L

łańcuch Markowa, 320
łatwość interpretacji modeli eksploracji danych, 373

M

macierz klasyfikacji, 384
mapa decyzji, 37
Market Basket Analysis, 142
Markowa łańcuch, 320
MAXIMUM_INPUT_ATTRIBUTES, 273, 304, 361
MAXIMUM_ITEMSET_COUNT, 341
MAXIMUM_ITEMSET_SIZE, 341
MAXIMUM_OUTPUT_ATTRIBUTES, 273, 361
MAXIMUM_SEQUENCE_STATES, 323
MAXIMUM_SERIES_VALUE, 287
MAXIMUM_STATES, 304, 323, 361
MAXIMUM_SUPPORT, 341
McCulloch, Warren, 352
metody oceny modeli eksploracji danych, 376
 macierz klasyfikacji, 384
 odchylenie międzyklastrowe, 390
 odchylenie wewnątrz-klastrowe, 390
 walidacja krzyżowa, 387
 wykres podniesienia, 376
 wykres punktowy, 381
 wykres zysku, 376
Microsoft
 drzewa decyzyjne, 267
 klastrowanie, 297
 klastrowanie sekwencyjne, 319
 naiwny klasyfikator Bayesa, 253
 odkrywanie reguł asocjacyjnych, 335
 regresja liniowa, 267
 regresja logistyczna, 351
 sieci neuronowe, 351
 szeregi czasowe, 281
mierzenie informacji, 76

MINIMUM_IMPORTANCE, 341
MINIMUM_ITEMSET_SIZE, 341
MINIMUM_PROBABILITY, 341
MINIMUM_SERIES_VALUE, 287
MINIMUM_SUPPORT, 273, 287, 304, 323, 341
MISSING_VALUE_SUBSTITUTION, 287, 386
modele eksploracji danych, 182, 184, 206, 232, 235
 dane brakujące, 70
 deskrypcyjne, 43, 108
 dokładność predykcji, 374
 korzystanie, 185
 kryteria porównawcze, 371
 łatwość interpretacji, 373
 ocena, 369, 376
 odczytywanie zawartości, 243
 poprawa, 369
 powrót do średniej, 369
 predykcyjne, 43, 109
 problemy, 391
 przetwarzanie, 210, 220, 239
 przydatność, 375
 skalowalność, 375
 wiarygodność predykcji, 374
 wizualizatory, 398
 wstawianie wierszy, 242
 wydajność, 375
 zarządzanie, 190
 zarządzanie poprzez SSMS, 216
MODELING_CARDINALITY, 304
modelowanie, 23, 25
 abstrakcja, 30
 cele, 35
 dane, 27
 decyzje, 31
 informacje, 27
 obiekty, 26
 paradygmaty, 29
 reguły, 26
 symbole, 30
 wiedza, 29
 wzorcowe, 30
 zdarzenia, 26
monotoniczne atrybuty, 59

N

nadmiarowe atrybuty, 75
naiwny klasyfikator Bayesa, 72, 253
 analiza dokumentów tekstowych, 260
 analiza zależności pomiędzy atrybutami, 258
 ograniczenia, 255
 omówienie, 253
 parametry, 256
 zastosowania, 258
narzędzia eksploracji danych, 162

nieprzygotowane dane źródłowe, 393
nietypowe przypadki, 149
niewłaściwe
 algorytmy eksploracji danych, 394
 dane źródłowe, 392
niewłaściwie postawione zadania, 391
niezależne atrybuty, 74
Noise, 51
normalizacja zakresu, 87
numerowanie stanów, 84

O

obiekty, 26
ocena
 danych, 49, 121
 modeli eksploracji danych, 369
 dokładność predykcji, 374
 kryteria porównawcze, 371
 łatwość interpretacji, 373
 metody, 376
 powrót do średniej, 369
 przydatność, 375
 skalowalność, 375
 wiarygodność predykcji, 374
 wydajność, 375
ryzyka, 45
oczyszczanie danych, 122
odchylenie
 międzyklastrowe, 390
 standardowe, 67
 wewnątrzklastrowe, 390
odkrywanie reguł asocjacyjnych, 335
ograniczenia
 drzew decyzyjnych, 272
 klastrowania, 302
 klastrowania sekwencyjnego, 323
 naiwnego klasyfikatora Bayesa, 255
 regresji logistycznej, 358
 reguł asocjacyjnych, 340
 sieci neuronowych, 358
 szeregów czasowych, 285
okresowe atrybuty, 86
okresowość, 96
OLE DB/DM, 232
Olshen, Richard, 268
opisywanie danych, 36

P

paradygmaty, 29
parametry
 drzew decyzyjnych, 273
 klastrowania, 303
 klastrowania sekwencyjnego, 323

- parametry
 - naiwnego klasyfikatora Bayesa, 256
 - regresji logistycznej, 360
 - reguł asocjacyjnych, 341
 - sieci neuronowych, 360
 - szeregów czasowych, 286
- Pearsona współczynnik korelacji liniowej, 106
- PERIODICITY_HINT, 287
- Pits, Walter, 352
- podział danych, 124
- poprawa
 - jakości danych, 99
 - modeli eksploracji danych, 369
- poprawność danych, 401
- porządkowe atrybuty, 85
- powrót do średniej, 369
- prawdopodobieństwo sukcesu projektu eksploracji danych, 44
- PREDICTION_SMOOTHING, 287
- predykcja, 109, 111
- predykcjne
 - funkcje, 251
 - modele eksploracji danych, 43, 109
 - programowanie, 397
 - zapytania, 245
- problem, formułowanie, 33
- problemy z modelami eksploracji danych, 391
 - nieprzygotowane dane źródłowe, 393
 - niewłaściwe algorytmy, 394
 - niewłaściwe dane źródłowe, 392
 - niewłaściwie postawione zadania, 391
 - źle sparametryzowane algorytmy, 394
- proces eksploracji danych, 10, 11
- profilowanie danych, 54
- prognozowanie, 156, 179, 289
 - kostka wielowymiarowa, 292
 - krótkie serie danych, 293
 - przeplatane serie danych, 291
- programowanie predykcyjne, 397
 - inteligentne aplikacje, 401
 - narzędzia, 397
 - raporty usługi SSRS, 399
 - wizualizatory modeli eksploracji danych, 398
- projekt eksploracji danych
 - dane źródłowe, 40
 - kontekst, 40
 - ocena ryzyka, 45
 - sukces, 44
 - zakres, 39
- proporcja, zmiana, 109
- próbkiwanie danych, 64
- przestrzeń stanów, 79
- przetwarzanie
 - modeli eksploracji danych, 210, 220
 - struktur eksploracji danych, 204, 220

- przewidywanie kolejnych zdarzeń, 329
- przydatność modeli eksploracji danych, 375
- przygotowanie danych, 79
 - do dalszej eksploracji, 312
- przykładowe bazy danych
 - AdventureWorksDW, 16
 - DataMiningDW, 17
 - DataMiningSolutions, 19
- przypadki, 51, 229
- wybijanie, 239
- zagnieżdżanie, 213, 236

Q

Quinlan, John Ross, 268

R

- raporty usługi SSRS, 399
- redukcja wymiarów, 105
- regresja liniowa, 267
- regresja logistyczna, 351
 - klasyfikacja, 366
 - ograniczenia, 358
 - omówienie, 352
 - parametry, 360
 - szacowanie, 362
 - zastosowania, 361
- reguły, 26
- reguły asocjacyjne, 335
 - analiza koszykowa, 343
 - analiza typu cross-selling, 347
 - analiza zależności pomiędzy atrybutami, 342
 - ograniczenia, 340
 - omówienie, 336
 - parametry, 341
 - zastosowania, 341
- reprezentatywność danych, 64
- Rosenblatt, Frank, 352
- rozkład wartości atrybutów, 59
- ryzyko, 45

S

- SAMPLE_SIZE, 304, 361
- SCORE_METHOD, 274
- serie danych, 92
 - krótkie, 293
 - przeplatane, 291
- serwer SQL, 12
 - eksploracja danych, 161, 218
 - instalacja, 13
 - integracja z SSAS, 223

- integracja z SSIS, 226
- integracja z SSRS, 226
- usługi, 12
- wersja demonstracyjna, 13
- wymagane składniki, 14
- sezonowość, 96
- sieci neuronowe, 351
 - klasyfikacja, 366
 - ograniczenia, 358
 - omówienie, 352
 - parametry, 360
 - szacowanie, 362
 - zastosowania, 361
- Silesian Code Camp, 398
- skalowalność modeli eksploracji danych, 375
- skalowanie
 - liniowe, 88
 - logistyczne, 89
- składniki serwera SQL, 14
- skrajne wartości, 87
- skupienia komórek, 305
- SPLIT_METHOD, 274
- SQL Server Analysis Services, *Patrz* SSAS
- SQL Server Database Engine, 12
- SQL Server Integration Services, *Patrz* SSIS
- SQL Server Reporting Services, *Patrz* SSRS
- SSAS, 12, 126, 162
 - architektura, 219
 - bezpieczeństwo, 221
 - zarządzanie poprzez SSMS, 216
- SSIS, 12, 54
 - profilowanie danych, 54
- SSMS, 162, 216
- SSRS, 13
 - raporty usługi, 399
- stałe, 57
- standaryzacja klasyczna, 89
- stany
 - atributów, 57, 229
 - numerowanie, 84
 - przestrzeń, 79
- Stone, Charles, 268
- STOPPING_TOLERANCE, 304
- struktury eksploracji danych, 182, 199, 231, 233
 - odczytywanie zawartości, 243
 - przetwarzanie, 204, 220, 239
 - wstawianie wierszy, 242
- sukces projektu eksploracji danych, 44
- symbole, 30
- szacowanie, 136, 170, 277, 309, 362
 - funkcja uzupełniania, 136
 - kalkulator predykcji, 138
- szeregi czasowe, 281
 - analiza wariantowa, 295

- ocena dokładności, 386
- ograniczenia, 285
- omówienie, 281
- parametry, 286
- prognozowanie, 289
 - kostka wielowymiarowa, 292
 - krótkie serie danych, 293
 - przeplatane serie danych, 291
 - zastosowania, 288
- sztuczna inteligencja, 352
- szukanie wyniku, 153
- szum, 97

T

- tabele zagnieżdżone, 234
 - wstawianie wierszy, 243
- TABLESAMPLE, 115
- Targeted Mailing Decision Tree, 134
- techniki eksploracji danych, 119, 126
 - analiza sekwencyjna, 151
 - analiza wariantowa, 152
 - asocjacja, 141
 - grupowanie, 145
 - klasyfikacja, 126
 - prognozowanie, 156
 - szacowanie, 136
- tekstowe atrybuty, 68
- testowe dane, 113
- trend, 96
- treningowe dane, 114
 - filtrowanie, 209
- typy decyzji, 36

U

- usługi serwera SQL, 12
 - eksploracja danych, 218
- uzupełnienie
 - brakujących danych, 404
 - wartości, 99

W

- walidacja krzyżowa, 116, 387
- wartości
 - atributów, 59, 229
 - skrajne, 87
 - uzupełnienie, 99
 - zakres, 63
- wersje demonstracyjne
 - Excelsa, 15
 - serwera SQL, 13

What-If, 155
wiarygodność predykcji modeli eksploracji
danych, 374
widoki danych źródłowych, 196
wiedza, 29
wiele do wielu, 85
wielowartościowe atrybuty, 57
wielowymiarowa kostka, 292
Wightman, Charles, 352
wizualizatory modeli eksploracji danych, 398
wspomaganie decyzji, 25, 36
współczynnik korelacji liniowej Pearsona, 106
wstawianie wierszy
do modeli eksploracji danych, 242
do struktur eksploracji danych, 242
do tabel zagnieżdżonych, 243
wybieranie przypadków, 239
wydajność modeli eksploracji danych, 375
wydzielenie danych testowych, 113
wykresy
podniesienia, 376
punktowy, 381
zysku, 376
wykrywanie
anomalii, 149, 314, 332
kategorii, 146
wymiały, redukcja, 105
wyniki eksploracji danych, 42
wzbogacenie danych, 103
wzorce, 30, 63

X

xml, 55

Z

zagnieżdżanie
przypadków, 213, 236
tabel, 234
zakres
normalizacja, 87
wartości danych, 63
zależności pomiędzy atrybutami, 73, 258, 342
zapytanie predykcyjne, 134, 210, 245
zarządzanie modelami eksploracji danych, 190
zaskakujące informacje, 77
zastosowania
drzew decyzyjnych, 275
eksploracji danych, 119
klastrowania, 305
klastrowania sekwencyjnego, 324
naiwnego klasyfikatora Bayesa, 258
regresji logistycznej, 361
reguł asocjacyjnych, 341
sieci neuronowych, 361
szeregów czasowych, 288
zbieżność danych, 65
zdarzenia, 26
zewnętrzne dane, 162
zgodność danych ze wzorcem, 63
zmiana proporcji, 109
zmiennne, 58
zmiennność atrybutów tekstowych, 68

Ż

źle sparametryzowane algorytmy eksploracji
danych, 394
źródła danych, 195, 240
źródłowe dane, 40, 49, 121
nieprzygotowane, 393
niewłaściwe, 392
widoki, 196

PROGRAM PARTNERSKI

GRUPY WYDAWNICZEJ HELION



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW
w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA WYDAWNICZA

 **Helion SA**

Poznaj sprawdzone techniki profesjonalnej eksploracji danych!

Eksploracja danych jest najmłodszą i najszybciej rozwijającą się dziedziną informatyki. Łączy zaawansowane algorytmy analizowania danych i znajdowania ukrytych w nich wzorców z klasycznymi technikami statystycznymi, rachunkiem prawdopodobieństwa i technologiami bazodanowymi. Dziedzina ta zyskuje na znaczeniu wraz z lawinowo rosnącą ilością informacji cyfrowych, które każdego dnia na całym świecie są wytwarzane, zapisywane i przeszukiwane przez stale zwiększającą się grupę użytkowników technologii informatycznych. Trzydzieści lat temu świat zrewolucjonizowały komputery PC, a dziś w ogarnięciu tego chaosu pomóc mogą jedynie najskuteczniejsze narzędzia do modelowania i eksploracji danych.

Microsoft SQL Server. Modelowanie i eksploracja danych to książka, z której analitycy, informatycy i biznesmeni dowiedzą się, jak tworzyć właściwe modele, odpowiednio przygotowywać dane i prawidłowo je eksplorować, a także jak należy analizować i oceniać otrzymane wyniki przy użyciu narzędzi oferowanych przez oprogramowanie Microsoft SQL Server. Publikacja przybliży zagadnienia związane z poszczególnymi etapami procesu modelowania i eksploracji, prezentując przy tym zastosowanie różnych metod i technik analizy do rozwiązywania praktycznych problemów naukowych i biznesowych.

- Podstawowe techniki i narzędzia wykorzystywane w eksploracji danych
- Instalacja i konfiguracja niezbędnego oprogramowania
- Analiza biznesowa projektu eksploracji danych
- Ocena, przygotowywanie i poprawianie jakości danych
- Przegląd technik eksploracji danych
- Wykorzystywanie serwera SQL w procesie eksploracji danych
- Zasada działania zaawansowanych algorytmów eksploracji danych
- Programowanie predykcyjne

Naucz się wykorzystywać zaawansowane narzędzia do inteligentnej zamiany dużych zbiorów danych w przydatne informacje!

Nr katalogowy: 7481



Księgarnia internetowa
<http://helion.pl>



Zamówienia telefoniczne:

0 801 339900



0 601 339900



Helion

Sprawdź najnowsze promocje:

🔗 <http://helion.pl/promocje>

Książki najchętniej czytane:

🔗 <http://helion.pl/bestsellery>

Zamów informacje o nowościach:

🔗 <http://helion.pl/novosci>

Helion SA

ul. Kościuszki 1c, 44-100 Gliwice

tel.: 32 230 98 63

e-mail: helion@helion.pl

<http://helion.pl>

helion.pl
księgarnia
internetowa

Cena 69,00 zł

ISBN 978-83-246-3440-8



9 788324 634408

Informatyka w najlepszym wydaniu