Księgarnia PWN: Daniel T. Larose - Odkrywanie wiedzy z danych

Przedmowa

0.1. Co to jest eksploracja danych?

Według przewidywań internetowego magazynu *ZDNET News* z 8 lutego 2001 roku eksploracja danych (ang. *data mining*)[1] będzie „jednym z najbardziej rewolucyjnych osiągnięć następnej dekady". Rzeczywiście *MIT Technology Review* wybrało eksplorację danych jako jedną z dziesięciu nowych technologii, które zmienią świat. Według Gartner Group, „eksploracja danych jest procesem odkrywania ważnych nowych współzależności, wzorców, trendów dzięki przeszukiwaniu dużych ilości danych przechowywanych w bazach, za pomocą zarówno technik rozpoznawania wzorców, jak i metod statystycznych i matematycznych."

Eksploracja danych jest tak ważną dziedziną, że wydawnictwo Wiley-Interscience i dr Daniel T. Larose połączyli siły, aby wydać cykl publikacji na ten temat, składający się wstępnie z trzech tomów. Pierwszy tom serii, *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych* wprowadza czytelnika w tę szybko rozwijającą się dziedzinę.

0.2. Dlaczego ta książka jest potrzebna?

Ludzie są zasypywani danymi w większości dziedzin. Niestety, te cenne dane, których zebranie i zestawienie kosztuje firmy miliony, marnują się w hurtowniach i skarbnicach danych. *Problemem jest zbyt mała liczba dostępnych, wykwalifikowanych analityków, którzy potrafią przemienić te wszystkie dane w wiedzę*, czyli drzewo taksonomiczne w mądrość. Właśnie dlatego książka ta jest potrzebna. Umożliwia ona czytelnikowi:

- zapoznanie się z modelami i technikami odkrywania ukrytych cennych informacji,
- dogłębne zrozumienie, jak działają algorytmy eksploracji danych,
- wykonywanie rzeczywistej eksploracji danych na dużych zbiorach danych.

Eksploracja danych rozpowszechnia się z każdym dniem coraz bardziej, ponieważ umożliwia firmom odkrywanie przynoszących zyski wzorców i trendów z posiadanych przez nie baz danych. Firmy i instytucje wydały miliony dolarów, by zgromadzić megabajty i terabajty danych, ale nie czerpią korzyści z cennych informacji ukrytych głęboko

[1] W terminologii polskiej spotykane są również następujące określenia: odkrywanie wiedzy z baz danych, drążenie danych, zgłębianie danych, kopanie danych, ekstrakcja danych (*przyp. tłum.*).

w ich skarbnicach. Jednak wraz ze wzrostem popularności technik eksploracji danych firmom, które nie stosują tych technik, grozi upadek i utrata rynku, ponieważ ich konkurenci, wykorzystując techniki eksploracji danych, zyskują przewagę. W książce *Odkrywanie wiedzy z danych* opisane, krok po kroku, bezpośrednie rozwiązania rzeczywistych problemów biznesowych za pomocą szeroko dostępnych technik eksploracji danych zastosowane do rzeczywistych zbiorów danych przemówią do menadżerów, dyrektorów IT (tzw. CIO), dyrektorów naczelnych (tzw. CEO), finansistów (tzw. CFO) i innych, którzy muszą być na bieżąco z najnowszymi metodami zwiększającymi zwrot inwestycji.

0.3. Niebezpieczeństwo! Łatwo jest źle przeprowadzić eksplorację danych

Nadmiar nowego gotowego oprogramowania do przeprowadzania eksploracji danych stworzył nowy rodzaj zagrożenia. Łatwość, z jaką te programy oparte na GUI mogą manipulować danymi, połączona z potężną mocą algorytmów ekstrakcji danych wbudowanych w obecnie dostępnym oprogramowaniu typu „czarna skrzynka", powoduje, że ich złe użycie staje się jeszcze bardziej niebezpieczne.

Tak jak każdą nową technikę informacyjną *eksplorację danych jest łatwo źle przeprowadzić*. Mała wiedza jest szczególnie niebezpieczna w przypadku zastosowania potężnych modeli opartych na dużych zbiorach danych. Na przykład analizy przeprowadzone na nieprzygotowanych danych mogą prowadzić do błędnych wniosków lub też niewłaściwe analizy mogą zostać zastosowane do zbioru danych, który wymaga zupełnie innego podejścia, albo też można otrzymać modele oparte na błędnych założeniach. Jeżeli zostaną one wdrożone, błędy analizy mogą prowadzić do bardzo kosztownych niepowodzeń.

0.4. Podejście typu „biała skrzynka'': Zrozumienie podstawowych struktur algorytmów i modeli

Najlepszym sposobem na uniknięcie tych kosztownych błędów, które wywodzą się z podejścia typu „czarna skrzynka" do ekstrakcji danych, jest zastosowanie metodologii typu „biała skrzynka", w której kładzie się nacisk na zrozumienie algorytmicznych i statystycznych struktur leżących u podłoża oprogramowania. W niniejszej książce zostało zastosowane podejście typu biała skrzynka przez:

- przykładowe omówienie działania różnych algorytmów,
- dostarczenie przykładów działania algorytmów na rzeczywistych dużych zbiorach danych,
- sprawdzenie stopnia zrozumienia pojęć i algorytmów przez czytelnika,
- stworzenie czytelnikowi możliwości wykonania rzeczywistej eksploracji danych na dużych zbiorach danych.

0.4.1. Omówienie działania algorytmów

W książce *Odkrywanie wiedzy z danych* są opisane kolejne operacje i niuanse różnych algorytmów, z użyciem małej próbki zbioru danych, tak by czytelnik właściwie zrozumiał to, co się dzieje w algorytmie. Na przykład w rozdziale 8 widzimy, jak środki grup są uaktualniane, przesuwane w kierunku środka swoich grup. Również w rozdziale 9 widzimy, który typ wag sieci okaże się neuronem wygrywającym w danej sieci i dla danego rekordu.

0.4.2. Zastosowanie algorytmów do dużego zbioru danych

Odkrywanie wiedzy z danych dostarcza przykłady zastosowań różnych algorytmów do dużych zbiorów danych. Na przykład w rozdziale 7, zadanie klasyfikacji dla rzeczywistego zbioru danych jest rozwiązywane za pomocą modelu sieci neuronowej. Opisano, jak wynikowa topologia sieci razem z wagami jest sprawdzana przez oprogramowanie. Te zbiory danych są dostępne na stronach internetowych książki: `http://www.dataminingconsultant.com/DKD.htm` , zatem czytelnik może sam podążyć za analitycznymi krokami, używając wybranego oprogramowania.

0.4.3. Ćwiczenia do rozdziałów: Sprawdzenie, aby upewnić się, że rozumiesz

Odkrywanie wiedzy z danych zawiera ponad 90 ćwiczeń, które pozwolą czytelnikom ocenić stopień zrozumienia materiału oraz dostarczą nieco zabawy z liczbami i danymi. Składają się one z ćwiczeń pojęciowych, które pomogą wyjaśnić niektóre z ambitniejszych pojęć w eksploracji danych, i z ćwiczeń na „niewielkich zbiorach danych”, które zmobilizują czytelnika do zastosowania konkretnego algorytmu do małego zbioru danych, krok po kroku, by dotrzeć do rozwiązania liczbowego. Na przykład w rozdziale 6 dany jest mały zbiór danych i czytelnicy są proszeni o ręczne zbudowanie, za pomocą metod pokazanych w rozdziale, modelu drzewa decyzyjnego C4.5, jak również modelu drzewa klasyfikacji i modelu drzewa regresji, oraz porównanie zalet i wad każdego z nich.

0.4.4. Ćwiczenia praktyczne: Nauka eksploracji danych poprzez wykonywanie eksploracji danych

Rozdziały od 2 do 4 oraz od 6 do 11 zawierają praktyczne ćwiczenia umożliwiające czytelnikowi zastosowanie nowo poznanej wiedzy o eksploracji danych do rozwiązania rzeczywistych problemów na dużych zbiorach danych. Skuteczna jest nauka przez praktykę. *Odkrywanie wiedzy z danych* dostarcza konstrukcji (schematu), za pomocą której czytelnik może nauczyć się eksploracji danych, eksplorując dane. Zamiarem jest odzwierciedlenie rzeczywistego scenariusza eksploracji danych. W rzeczywistości „brudne” zbiory danych potrzebują czyszczenia, surowe dane potrzebują normalizacji, dane spoza dziedziny muszą zostać sprawdzone. W tej książce znajduje się ponad 70 praktycznych ćwiczeń. W ten sposób czytelnik może szybko „nauczyć się” i stosunkowo szybko samodzielnie przeprowadzać własne analizy eksploracji danych.

Na przykład w rozdziale 10 czytelnik ma za zadanie odkryć pewne, z wysokim poparciem, reguły do przewidywania, który klient zrezygnuje z usług firmy. W rozdziale 11 czytelnicy są proszeni o stworzenie wykresów przyrostu i zysku dla kilku metod klasyfikacji dużego zbioru danych, aby wybrać najlepszy model.

0.5. Eksploracja danych jako proces

Jednym z błędnych przekonań związanych z wdrożeniem eksploracji danych jest to, że eksploracja danych reprezentuje odizolowany, gotowy do użycia przez dział analiz zbiór narzędzi, nieistotny dla głównego projektu biznesu lub badań. Organizacje, które próbują wdrożyć eksplorację danych w ten sposób, będą miały bardzo ograniczone szanse na sukces. Dzieje się tak dlatego, że eksploracja danych powinna być postrzegana jako proces.

Odkrywanie wiedzy z danych przedstawia eksplorację danych jako dobrze zorganizowaną metodologię, silnie powiązaną z menadżerami, decydentami i tymi, którzy zajmują się wdrożeniem wyników.

Tak więc ta książka przeznaczona jest nie tylko dla analityków, ale również dla menadżerów, którzy będą musieli umieć porozumieć się w języku eksploracji wiedzy. Jedną z używanych metodologii jest CRISP-DM: Cross-Industry Standard Process for Data Mining. CRISP-DM wymaga, aby eksploracja wiedzy była postrzegana jako całkowity proces, od zrozumienia uwarunkowań biznesowych, przez zebranie i zarządzanie danymi, przygotowanie danych, modelowanie, ewaluację modelu do jego wdrożenia. Dlatego też, ta książka jest przeznaczona nie tylko dla analityków i menadżerów, ale również dla osób zarządzających danymi, projektantów baz danych oraz decydentów.

0.6. Podejście graficzne, podkreślające wstępną analizę danych

W książce położono nacisk na graficzne podejście do analizy danych. Znajduje się w niej ponad 80 zrzutów ekranu przedstawiających wyniki komputerowe i ponad 30 innych rysunków. Eksploracyjna analiza danych (ang. *Exploratory Data Analysis*, EDA) reprezentuje interesującą i fascynującą metodę „rób po swojemu" dla dużych zbiorów danych. Wykorzystując wizualizację i podsumowania numeryczne, analitycy stopniowo rzucają światło na złożone relacje ukryte w danych. Uwypuklenie techniki EDA w eksploracji danych idzie w parze z ogólnym podejściem graficznym.

0.7. Jak zorganizowana jest ta książka?

Odkrywanie wiedzy z danych jest obszernym wprowadzeniem do tej dziedziny. Pokazano, jak eksploracja wiedzy została wykorzystana pomyślnie (i nie tak pomyślnie). Popularne mity na temat eksploracji danych są obalane, a znane pułapki sygnalizowane, tak aby nowi adepci nie musieli się uczyć na własnych błędach.

Pierwsze trzy rozdziały są wprowadzeniem zgodnym z metodologią CRISP-DM, zwłaszcza podczas faz przygotowania i zrozumienia danych. Następnych siedem rozdziałów stanowi sedno książki i są one związane z fazą modelowania CRISP-DM. Każdy rozdział przedstawia metody i techniki eksploracji danych dla konkretnego zadania eksploracji danych.

- Rozdziały 5, 6 i 7 odnoszą się do **zadania klasyfikacji**; zostały w nich omówione algorytmy *k*-najbliższych sąsiadów (rozdział 5), drzew decyzyjnych (rozdział 6) i sieci neuronowych (rozdział 7).
- Rozdziały 8 i 9 dotyczą **zadania grupowania**, z metodami hierarchicznego grupowania, *k*-średnich (rozdział 8) oraz z sieciami Kohonena (rozdział 9).
- Rozdział 10 dotyczy **zadania skojarzeń**; omawiane są tu reguły asocjacyjne, tworzone poprzez algorytmy *A priori* oraz GRI.
- W końcu rozdział 11 zawiera opis technik oceny modelu, które należą do fazy ewaluacji modelu w metodologii CRISP-DM.

0.8. *Odkrywanie wiedzy z danych* jako podręcznik

Odkrywanie wiedzy z danych oczywiście spełnia funkcję podręcznika do wykładu z eksploracji danych na poziomie podstawowym. Wykładowcy mogą docenić:

- przedstawienie eksploracji danych jako procesu;
- podejście typu „biała skrzynka", kładące nacisk na zrozumienie podstawowych struktur algorytmu:
  - opisy algorytmów,
  - zastosowanie algorytmów do dużych zbiorów danych,
  - ćwiczenia do rozdziałów,
  - ćwiczenia praktyczne;
- podejście graficzne, podkreślające wstępną analizę danych;
- prezentacje logiczne, przechodzące płynnie od metodologii CRISP-DM do zbioru zadań eksploracji danych.

Odkrywanie wiedzy z danych może być wykorzystane przez studentów starszych lat studiów licencjackich lub młodszych lat studiów magisterskich. Poza jednym podrozdziałem z rozdziału 7, rachunek różniczkowy nie jest wymagany. Wykład ze statystyki na poziomie podstawowym może pomóc w opanowaniu materiału, ale nie jest wymagany. Nie jest wymagana umiejętność programowania oraz znajomość baz danych.

0.9. Podziękowania

Odkrywanie wiedzy z danych nie zostałoby napisane bez pomocy redaktora, Vala Moliere'a, koordynatora programowego wydawcy Kirstena Rohsteda, redaktora technicznego w wydawnictwie Wiley-Interscience Rosalyn Farkas oraz Barbary Zeiders, która adjustowała pracę. Dziękuję za wasze wskazówki i wytrwałość.

Chciałbym również podziękować doktorowi Chun Jin i doktorowi Danielowi S. Millerowi, moim kolegom z programu Master of Science in Data Mining w Central Connecticut State University; doktorowi Timothy'emu Craine'owi, dziekanowi Wydziału Matematyki, doktorowi K. Deyowi, dziekanowi wydziału statystyki na University of Connecticut i doktorowi Johnowi Judge'owi, dziekanowi Wydziału Matematyki w Westfield State College. Wasza pomoc była (i jest) bezcenna.

Dziękuję moim dzieciom, Chantal, Tristanowi i Ravel, za dzielenie ze mną komputera. W końcu chciałbym podziękować mojej wspaniałej żonie, Debrze J. Larose, za jej cierpliwość, wyrozumiałość i umiejętności korektorskie. Ale słowami nie można wszystkiego wyrazić…

Daniel T. Larose, Ph.D.
Director, Data Mining @CCSU
www.ccsu.edu/datamining