Zeszyty Naukowe PWSZ w Płocku
Nauki Ekonomiczne, t. XXI, 2015.

Katarzyna Racka
Państwowa Wyższa Szkoła Zawodowa w Płocku

METODY EKSPLORACJI DANYCH I ICH ZASTOSOWANIE

Wprowadzenie

Współczesne firmy przechowują i przetwarzają ogromne ilości informacji w bazach danych i hurtowniach danych. Zgromadzone dane opisujące działania przedsiębiorstwa i jego klientów pozwalają na analizę trendów, anomalii rozwoju firmy, oceny klienta a także przewidywania istotnych zagrożeń finansowych za pomocą metod data mining.

Data mining nazywana eksploracją danych, lub odkrywaniem wiedzy w bazach danych, to proces odkrywania nowych reguł, wzorców i zależności.

Głównym celem artykułu jest omówienie metod eksploracji danych, przedstawienie ich zastosowań oraz zaprezentowanie przykładów stosowanych oprogramowań.

1. Metody eksploracji danych i ich zastosowanie

Postępujący rozwój informatyzacji, coraz większy dostęp do sieci komputerowych i powszechne gromadzenie informacji w bazach i hurtowniach danych prowadzi do stałego wzrostu ilości przechowywanych danych. Codziennie sklepy, banki, firmy np.: finansowe, ubezpieczeniowe, telekomunikacyjne, turystyczne, agencje marketingowo-reklamowe, portale internetowe, ośrodki medyczne lub naukowo-badawcze wykonują i zapisują tysiące operacji handlowych, transakcji, raportów i opisów. Człowiek sam nie jest w stanie szybko analizować tak dużej ilości danych. W tym celu korzysta się z metod eksploracji danych (data mining), które umożliwiają pozyskiwanie nowej wiedzy wspomagającej procesy decyzyjne.

Pojęcie eksploracji danych definiowane jest jako proces odkrywania wzorców, reguł, zależności w dużych zbiorach danych (hurtownie danych). Zasadniczym celem eksploracji danych jest wydobycie nowej – nieznanej informacji z baz danych.

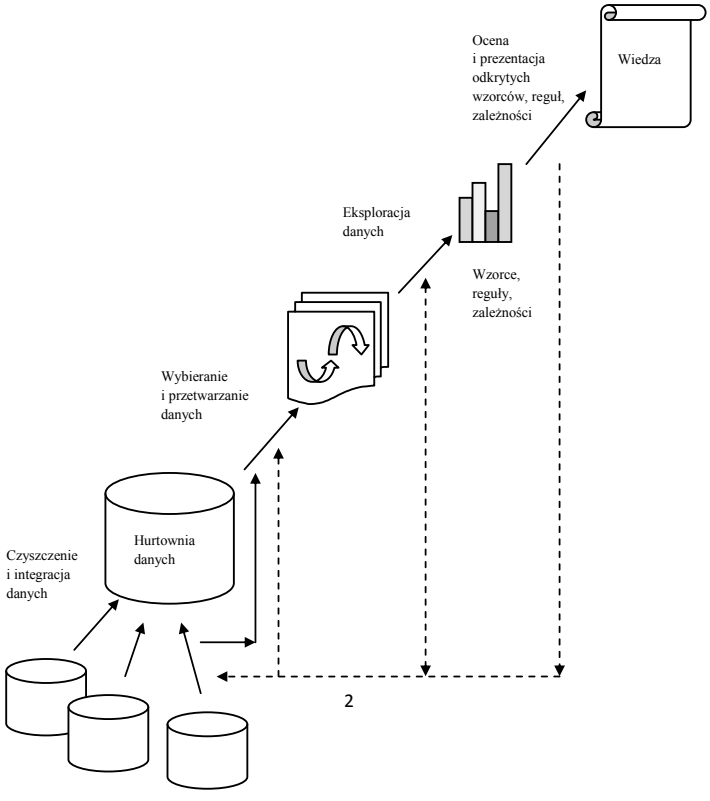
Eksploracja danych jest jednym z etapów procesu odkrywania wiedzy z baz danych (ang. Knowledge Discovery in Databases, KDD), który składa się z następujących kroków:

I. Czyszczenie danych (ang. data cleaning) – usuwanie błędów wynikających z pomyłek operatora (ortograficzne, literówki), różnych form danych

oznaczających te same informacje, niezgodności wartości pola i jego opisu, brakujących wartości, błędów wynikających z ograniczeń systemu (brak pól na niektóre ważne dane), wielokrotnego wprowadzenia tej samej danej.

II. Integracja danych – łączenie danych pochodzących z różnych źródeł (baz danych), posiadających różną strukturę oraz różne modele danych.
III. Wybieranie danych – z bazy danych pobierane są dane do przeprowadzenia analiz.
IV. Transformacja danych – przetwarzanie lub łączenie danych w formach odpowiednich dla eksploracji, wykonując np. operacje podsumowania lub agregacji.
V. Eksploracja danych – stosowanie metod eksploracji danych w celu wydobycia z danych wzorców, reguł, zależności.
VI. Ocena odkrytych wzorców, reguł, zależności – identyfikacja najbardziej interesujących (istotnych) odkrytych wzorców.
VII. Prezentacja odkrytej wiedzy – przedstawienie odkrytej wiedzy użytkownikowi za pomocą technik wizualizacji i reprezentacji danych.

Rysunek 1. Eksploracja danych jako jeden z kroków w procesie odkrywania wiedzy



Źródło: opracowanie własne na podstawie: Han J., Kamber M., *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Academic Press, 2001.

Wśród najbardziej znanych metod eksploracji danych możemy wyróżnić:

- Wyszukiwanie asocjacji
- Klasyfikacja
- Grupowanie
- Wykrywanie zmian i odchyleń.
- Odkrywanie wzorców sekwencji
- Eksploracja danych tekstowych

Wyszukiwanie asocjacji – umożliwia znajdowanie nieznanych zależności lub reguł (asocjacji) pomiędzy występującymi elementami w zbiorach danych.

Przykłady zastosowania metody wyszukiwania asocjacji:

- Analiza koszyka kupionych produktów przez klienta w celu planowania rozmieszczenia produktów w supermarketach. Odkrywane reguły mogą przykładowo wyglądać następująco:
  „Jeżeli klienci kupują chleb i mleko, to kupują również masło".
  „Jeżeli klienci kupują chipsy i paluszki, to kupują też napoje gazowane".
- Wycena ubezpieczenia. Przykłady reguł:
  „Jeżeli silnik samochodu jest dużej mocy, to ryzyko wypadku jest wysokie i cena ubezpieczenia duża".
  „Jeżeli dom jest nowy i nie posiada alarmu, to ryzyko włamania jest wysokie i cena ubezpieczenia duża".

Metoda klasyfikacji (ang. classification) - polega na tworzeniu modelu, który używany jest do klasyfikowania nowych obiektów bazy. Występuje tu tak zwany zbiór danych treningowych a odkryte modele klasyfikacji są później używane do klasyfikacji nowych obiektów o nieznanej klasyfikacji.

Przykład zastosowania metody klasyfikacji:

- Wykrywanie nadużyć i oszustw finansowych korzystając ze zbioru danych treningowych zawierającego przykłady nadużyć i przykłady operacji uczciwych.
- Diagnostyka chorób na podstawie wcześniejszej klasyfikacji schorzeń.

Metoda grupowania (ang. clustering) - polega na tworzeniu skończonych podzbiorów (klas, grup) obiektów posiadających podobne cechy. Liczba utworzonych podzbiorów nie jest ustalana początkowo i wynika z podobieństwa lub ze zróżnicowania grupowanych obiektów.

Przykłady zastosowania metody grupowania :

- Grupowanie klientów ze względu na podobieństwo zakupionych produktów lub ilość zrealizowanych transakcji;
- Tworzenie grup tematycznie powiązanych dokumentów (wyszukiwarki internetowe).

Wykrywanie zmian i odchyleń – analiza danych zmieniających się w przedziale czasu i znajdowanie różnic pomiędzy aktualnymi a oczekiwanymi wartościami danych.

Przykłady zastosowania metody wykrywania zmian i odchyleń:

- Sygnalizowanie awarii lub włamania do systemów sieciowych.
- Wykrywanie oszustw podatkowych lub wyłudzeń ubezpieczeniowych.

- Wyszukiwanie nadużyć w sieciach telekomunikacyjnych.

Odkrywanie wzorców sekwencji – odkrywanie wzorców zachowań, na podstawie analizy danych zmieniających się w czasie.

Przykłady zastosowania metody odkrywania wzorców sekwencji:

- Odkrywanie wzorców zachowań użytkowników korzystających z Internetu.
- Badanie notowań akcji i odkrywanie wzorców w celu ustalenia modelu decyzyjnego dla strategii inwestycyjnych.

Eksploracja danych tekstowych (text mining) **–** metody eksploracji danych służące analizie treści dokumentów tekstowych w celu znalezienia nowych informacji, które nie są dostępne bezpośrednio.

Przykłady zastosowania eksploracji danych tekstowych

- Porównywanie publikacji naukowych, prac dyplomowych w celu wykrycia plagiatów.
- Klasyfikacja dokumentów tekstowych np.: poczty internetowej - oddzielenie informacji ważnych od nieistotnych.
- Grupowanie danych tekstowych np. artykułów ze względu na tematy, autorów, treść opisywaną.
- Analiza treści zamieszczanych na portalach społecznościowych, w celu wyszukiwania nowych informacji lub oceny i weryfikacji osób wpisujących tam komentarze.

W praktyce można korzystać z wielu różnych metod eksploracji danych dla tego samego rozpatrywanego zagadnienia zależnie od tego, jaka wiedza jest potrzebna dla analityka. Stosowanie różnych metod eksploracji danych może okazać się korzystniejsze, gdyż zastosowanie jednej metody może nie być wystarczające do całościowego rozwiązania rozpatrywanego problemu.

Przykład

Do oceny klientów bankowych można użyć następujących metod eksploracji danych:

- Metody klasyfikacji:
  - Do analizy i oceny klientów dzieląc ich na klasy w zależności od cech klientów (np.: wiek, wielkość miesięcznego dochodu, rodzaj i wielkość zaciągniętego kredytu) – klient przyniesie zarobek lub stratę dla banku.
  - Do oceny kredytu dla klienta (dobry – zły) na podstawie informacji o zatrudnieniu (rodzaj stanowiska, kwalifikacje pracownika), częstotliwości otrzymywanego wynagrodzenia, wieku klienta, oraz informacji czy posiada kredyt.
- Metody grupowania:
  - Do analizy i oceny klientów bankowych poprzez podział ich na różniące się charakterystykami grupy.
- Metody wyszukiwania asocjacji:
  - W celu generowania reguł, pozwalających na ustalenie ryzyka pojedynczego wniosku kredytowego, przewidzenie zachowania klienta banku, lub ocenę zysku jaki może dać dany klient o danej charakterystyce.

Przykład reguły: „*Jeżeli wiek = stary, dochód = mały, depozyt = bardzo duży, kredyt = mały, to zysk = duży*"[1].

- Metodę wykrywania trendów i odchyleń:
  - o W celu wykrycia oszustw i wyłudzeń.

2. Programy do eksploracji danych

Data mining jest prężnie rozwijająca się dziedziną. Powstaje coraz więcej narzędzi wspomagających procesy decyzyjne korzystających z metod eksploracji danych. Na polskim rynku najbardziej znane oprogramowania komercyjne data mining to produkty firm zaprezentowanych w poniższej tabeli (tabela 1).

Tabela 1. Lista przykładowych firm sprzedających na polskim rynku oprogramowania do eksploracji danych[2]

Nazwa firmy	Adres strony
HP	http://www8.hp.com/pl/pl
IBM	http://www.ibm.com/pl/pl/
Microsoft	http://www.microsoft.com
Oracle	http://www.oracle.com/pl
SAS Institute	http://www.sas.com/pl
StatSoft	http://www.statsoft.pl/

Źródło: opracowanie własne na podstawie stron internetowych.

Firmy wymienione w powyższej tabeli (tabela 1) oferują rozwiązania analityczne z zastosowaniem metod eksploracji danych w najróżniejszych dziedzinach, zarówno w branżach produkcyjnych jak i branżach usługowych. Produkty tych firm są na ogół dedykowane, tzn. dostosowane do potrzeb danego odbiorcy. Na stronach internetowych tych firm podane są referencje, z których szczegółowo można dowiedzieć się jakie firmy i instytucje korzystają z ich produktów.

Oprócz programów komercyjnych dostępne są również oprogramowania niekomercyjne do eksploracji danych, zaprezentowane w poniższej tabeli (tabela 2).

[1] M. Lasek, *Data mining. Zastosowanie w analizach i ocenach klientów bankowych*, Zarządzanie i Finanse, Warszawa 2002.

[2] Kolejność wymienionych firm alfabetyczna, gdyż celem tego artykułu nie jest reklama ani ocena i porównywanie produktów.

Tabela 2. Lista przykładowych darmowych programów do eksploracji danych

Nazwa programu	Adres strony	Typ licencji
CMSR DATA Miner	http://www.roselladb.com/starprobe.htm	Licencja akademicka na trzy lata, wersja darmowa na 6 miesięcy
Databionic ESOM Tools	http://databionic-esom.sourceforge.net/	GNU GPL
ELKI	http://elki.dbs.ifi.lmu.de/	AGPL
KNIME	http://www.knime.org/	GNU GPL
Mloss	http://mloss.org/software/	GNU GPL
Mlpy	http://mlpy.sourceforge.net/	GNU GPL
Orange	http://orange.biolab.si/	GNU GPL
Projekt R	http://www.r-project.org/	GNU GPL
Rapid Miner	https://rapidminer.com/	AGPL/ Proprietary (prawnie zastrzeżone)
Rattle GUI	rattle.togaware.com	GNU GPL
SCaViS	http://jwork.org/scavis/	Mieszana: jądro silnika programu GPL, instalacja, dokumentacja, podzespoły darmowe ale nie do celów komercyjnych
SenticNEt API	http://sentic.net/api/	Darmowy z umieszczeniem informacji: Copyright © 2012 Yuri Malheiros
Weka 3	http://www.cs.waikato.ac.nz/ml/weka/	GNU GPL

Źródło: opracowanie własne na podstawie stron internetowych.

Jak widać z powyższej tabeli, darmowych programów do eksploracji danych jest bardzo wiele i nie zostały tutaj wymienione wszystkie. Licencja GNU GPL oznacza, że program można uruchamiać dowolną ilość razy. Program udostępniony jest z kodem źródłowym, który można ulepszać do własnych potrzeb oraz rozpowszechniać. Licencja AGPL jest natomiast licencją wolnego oprogramowania, które będzie uruchamiane przez sieć. Programy na tych licencjach, są więc rozbudowywane i dostosowywane do potrzeb odbiorców przez wielu informatyków jak to ma miejsce na przykładzie projektu R. Dlatego też, często programy darmowe są wykorzystywane do celów komercyjnych lub dydaktycznych. Co ciekawe, a propos projektu R, jest on podobno używany między innymi przez: Facebook, Forda, Google, Microsoft, Mozilla.[3] Natomiast firmy tworzące oprogramowania komercyjne do eksploracji danych np. SAS, SPSS, Statistica, oferują dedykowane mechanizmy zapewniające ich współpracę z R.[4][5][6]

[3] http://www.statsoft.pl/
[4] http://support.sas.com/rnd/app/studio/Rinterface2.html
[5] http://www.statsoft.com/Solutions/Cross-Industry/R-Integrations
[6] http://www-03.ibm.com/software/products/pl/spss-stats-developer

Dodatkową zaletą programów niekomercyjnych jest fakt, iż oprócz tego, że są darmowe spokojnie dorównują a czasem nawet przewyższają niektóre programy komercyjne. Interfejs graficzny w programach taki jak np. Orange, Weka sprawia, że są one łatwe w obsłudze i nie wymagają dużych umiejętności programistycznych.

Wadą programów darmowych jest natomiast to, że nie zawsze mamy zagwarantowaną pewność ich prawidłowego działania. Przykładem może być tu program SenticNEt API, w którym sam autor zaznacza, że program jest bez jakiejkolwiek gwarancji i nie ponosi on odpowiedzialności za szkody wynikające z używania tego programu.

Z powodu dużej listy programów nie omówię ich wszystkich szczegółowo ale chętnych namawiam do odwiedzenia stron internetowych, których adresy podałam w bibliografii.

Podsumowanie

Metody eksploracji danych (data mining) są narzędziem odkrywania nieznanej wiedzy, reguł, wzorców i zależności w bazach a raczej hurtowniach danych. Ich zastosowanie można wskazać we wszystkich dziedzinach, w których należy dokonywać analizy i oceny dużej ilości danych, których człowiek sam nie jest w stanie szybko przeanalizować. Od szybkości i poprawności odkrytej wiedzy w bazach danych oraz odpowiedniego jej zastosowania może zależeć sukces lub klęska analizowanego problemu a nawet całej firmy. Jednak pamiętać należy przy tym, aby wnioski otrzymane z metod eksploracji danych były formułowane w postaci domniemań, a nie w postaci kategorycznych stwierdzeń. Aby wiedza pozyskana z metod eksploracji danych była rozważnie wykorzystywana w procesach decyzyjnych. Nie każda bowiem odkryta reguła czy wzorzec będą przydatne. To człowiek musi dokonać ostatecznej oceny otrzymanej wiedzy.

Bibliografia

- Han J., Kamber M., *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Academic Press, 2001.
- Lasek M., *Data mining. Zastosowanie w analizach i ocenach klientów bankowych*, Zarządzanie i Finanse, Warszawa 2002.
- Larose D. T., *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa 2012.
- Morzy T., *Eksploracja danych. Metody i algorytmy*, Wydawnictwo Naukowe PWN, 2013.

Strony internetowe

- http://databionic-esom.sourceforge.net/
- http://elki.dbs.ifi.lmu.de/
- http://jwork.org/scavis/
- http://mloss.org/software/
- http://mlpy.sourceforge.net/
- http://orange.biolab.si/

- http://sentic.net/api/
- http://support.sas.com/rnd/app/studio/Rinterface2.html
- http://www.cs.waikato.ac.nz/ml/weka/
- http://www.ibm.com/pl/pl/
- http://www.knime.org/
- http://www.microsoft.com
- http://www.oracle.com/pl
- http://www.revolutionanalytics.com/companies-using-r
- http://www.r-project.org/
- http://www.sas.com/pl
- http://www.statsoft.com/Solutions/Cross-Industry/R-Integrations
- http://www.statsoft.pl/
- http://www-03.ibm.com/software/products/pl/spss-stats-developer
- http://www8.hp.com/pl/pl
- https://rapidminer.com/
- rattle.togaware.com

DATA MINING METHODS AND THEIR APPLICATIONS

Summary

Success in the financial market reach those companies that having fast access to data can it properly used. In modern databases and data warehouses are collected vast amounts of information, which man himself is not able to quickly analyze. For this purpose are used the data mining methods that enable the discovery of new knowledge, that is, rules, patterns and relationships in large databases. The aim of this article is to present the data mining methods and their applications. Article is divided into two parts. In the first part of the article explains the concept of data mining and data mining methods are discussed and provides examples of their applications. In the second part of the article presents the companies selling on the Polish market commercial data mining software and examples of free open-source data mining software are discussed.

Key words: data mining, data mining methods, examples of data mining methods applications, data mining software