Eksploracja danych

true

2019-03-12

Spis treści

Wstęp

O książce

Niniejsza książka powstała na bazie doświadczeń autora, a głównym jej celem jest przybliżenie czytelnikowi podstaw z dziedziny *Data mining* studentom kierunku *Matematyka* Politechniki Lubelskiej. Będzie łączyć w sobie zarówno treści teoretyczne związane z przedstawianymi etapami eksploracji danych i budową modeli, jak i praktyczne wskazówki dotczące budowy modeli w środowisku **R** (R Core Team, 2018). Podane zostaną również wskazówki, jak raportować wyniki analiz i jak dokonać właściwych ilustracji wyników. Bardzo użyteczny w napisaniu książki były pakiety programu R: **bookdown** (Xie, 2018a), **knitr** (Xie, 2018b) oraz pakiet **rmarkdown** (Allaire et al., 2018).

Zakres przedmiotu

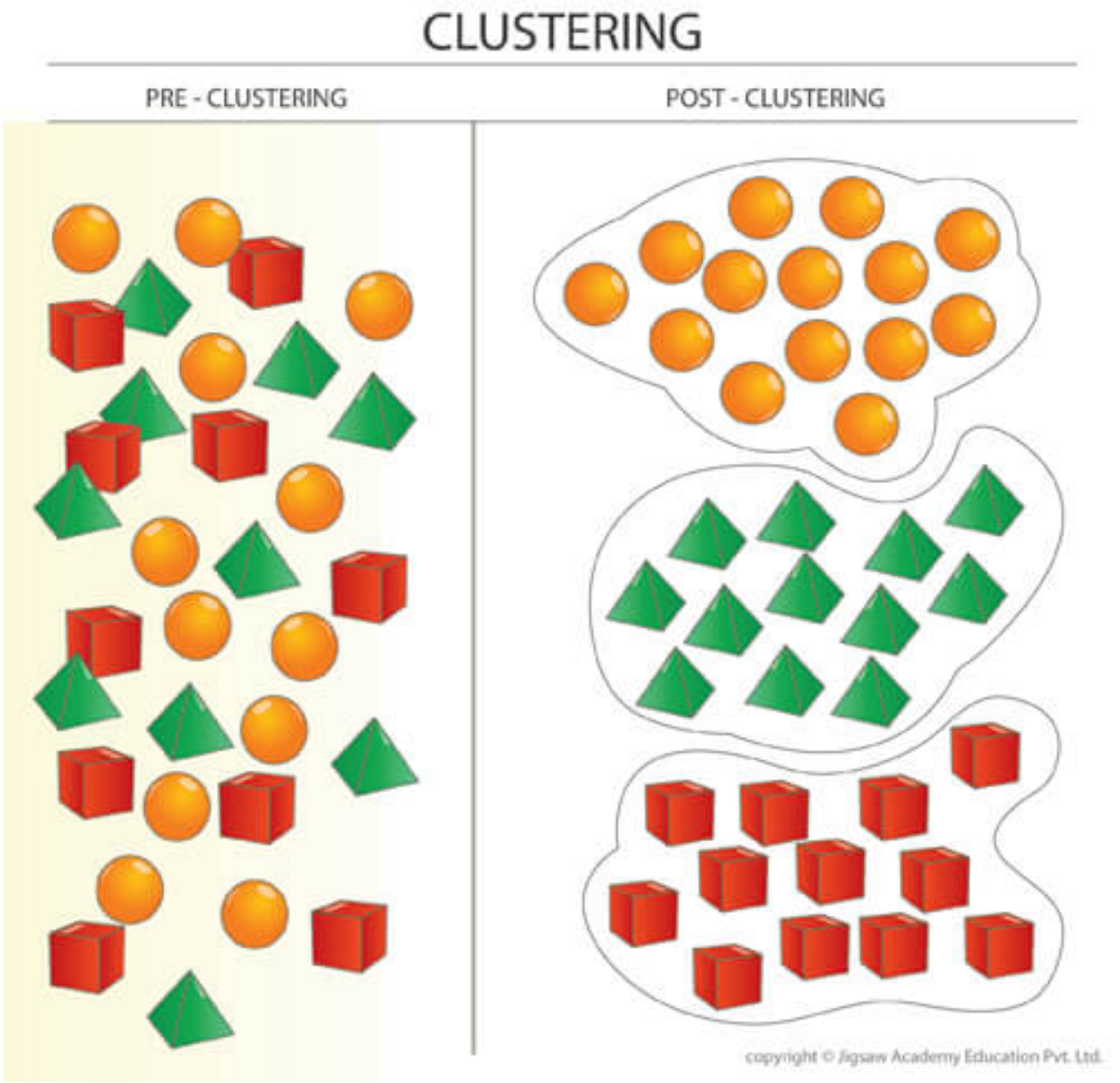
Przedmiot *Eksploracja danych* będzie obejmował swoim zakresem eksplorację i wizualizację danych oraz uczenie maszynowe. Eksploracja danych ma na celu pozyskiwanie i systematyzację wiedzy pochodzącej z danych. Odbywa się ona głównie przy użyciu technik statystycznych, rachunku prawdopodobieństwa i metod z zakresu baz danych. Natomiast uczenie maszynowe, to gałąź nauki (obejmuje nie tylko statystykę, choć to na niej się głównie opiera) dotyczącej budowy modeli zdolnych do rozpoznawania wzorców, przewidywania wartości i klasyfikacji obiektów. Data mining to szybko rosnaca grupa metod analizy danych rozwijana nie tylko przez statystyków ale również przez biologów, genetyków, cybernetyków, informatyków, ekonomistów, osoby pracujace nad rozpoznawaniem obrazów i wiele innych grup zawodowych. W dzisiejszych czasch trudno sobie wyobrazić życie bez sztucznej inteligencji. Towarzyszy ona nam w codziennym, życiu kiedy korzystamy z telefonów komórkowych, wyszukiwarek internetowych, robotów sprzątających, automatycznych samochodów, nawigacji czy gier komputerowych. Lista ta jest niepełna i stale się wydłuża.

href="https://twitter.com/i/status/1091069356367200256">January 31, 2019

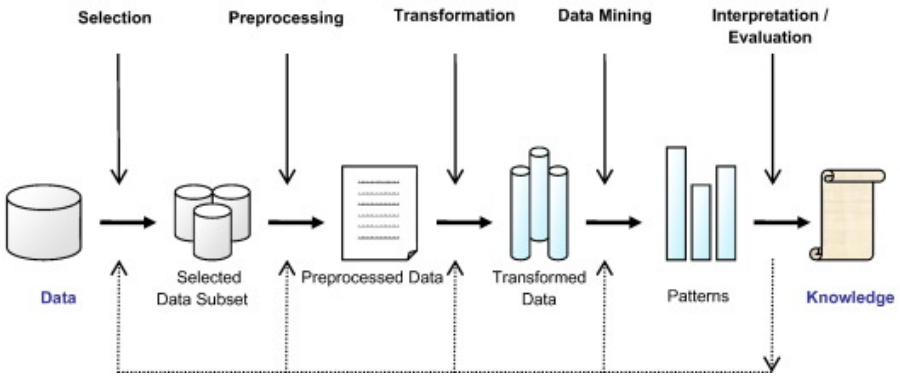
Zakres technik stosowanych w data mining

- statystyka opisowa
- wielowymiarowa analiza danych
- analiza szeregów czasowych
- analiza danych przestrzennych
- reguły asocjacji
- uczenie maszynowe[1], w tym:
  - klasyfikacja
  - predykcja
  - analiza skupień
  - *text mining*
- i wiele innych

[1] ang. *machine learning*



Rysunek 1: Przykład nienadzorowanego uczenia maszynowego. *Źródło:*https://analyticstraining.com/cluster-analysis-for-business/



Rysunek 2: Etapy eksploracji danych (Kavakiotis et al., 2017)

href="https://twitter.com/i/status/1097199751072690176">Ferbruary 17, 2019

Etapy eksploracji danych

1. Czyszczenie danych - polega na usuwaniu braków danych, usuwaniu stałych zmiennych, imputacji braków danych oraz przygotowaniu danych do dalszych analiz.
2. Integracja danych - łączenie danych pochodzących z różnych źródeł.
3. Selekcja danych - wybór z bazy tych danych, które są potrzebne do dalszych analiz.
4. Transformacja danych - przekształcenie i konsolidacja danych do postaci przydatnej do eksploracji.
5. Eksploracja danych - zastosowanie technik wymienionych wcześniej w celu odnalezienia wzorców[2] i zależności.
6. Ewaluacja modeli - ocena poprawności modeli oraz wzorców z nich uzyskanych.
7. Wizualizacja wyników - graficzne przedstawienie odkrytych wzorców.
8. Wdrażanie modeli - zastosowanie wyznaczonych wzorców.

[2]ang. *patterns*

Rozdział 1

Import danych

Placeholder

1.1 Przykład

Rozdział 2

Przygotowanie danych

Placeholder

2.1 Korekta zbioru danych

2.1.1 Identyfikacja braków danych

2.1.2 Zastępowanie braków danych

2.2 Przykład

Rozdział 3

Podział metod data mining

Placeholder

3.1 Rodzaje wnioskowania

3.1.1 Dziedzina

3.1.2 Obserwacja

3.1.3 Atrybuty obserwacji

3.1.4 Zbiór uczący

3.1.5 Zbiór testowy

3.1.6 Model

3.1.7 Jakość dopasowania modelu

3.2 Modele regresyjne

3.3 Modele klasyfikacyjne

3.4 Modele grupujące

Rozdział 4

Drzewa decyzyjne

Placeholder

4.1 Węzły i gałęzie

4.2 Rodzaje reguł podziału

4.2.1 Podziały dla atrybutów ze skali nominalnej

4.2.2 Podziały dla atrybutów ze skali ciągłej

4.2.3 Podziały dla atrybutów ze skali porządkowej

4.3 Algorytm budowy drzewa

4.4 Kryteria zatrzymania

4.5 Reguły podziału

4.6 Przycinanie drzewa decyzyjnego

4.6.1 Przycinanie redukujące błąd

4.6.2 Przycinanie minimalizujące błąd

4.6.3 Przycinanie ze względu na współczynnik złożoności drzewa

4.7 Obsługa braków danych

4.8 Zalety i wady

4.8.1 Zalety

4.8.2 Wady

4.9 Przykład

4.9.1 Podział zbioru na próbę uczącą i testową

4.9.2 Budowa drzewa

4.9.3 Przycinanie drzewa

4.9.4 Ocena dopasowania modelu

4.10 Inne algorytmy budowy drzew decyzyjnych implementowane w R

4.11 Przykład

4.11.1 `ctree`

4.11.2 `J48`

4.11.3 `C50`

Rozdział 5

Pochodne drzew decyzyjnych

Placeholder

5.1 Bagging

5.1.1 Przykład

5.2 Lasy losowe

5.2.1 Przykład

5.3 Boosting

Bibliografia

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2018). *rmarkdown: Dynamic Documents for R*. R package version 1.11.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104 – 116.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Xie, Y. (2018a). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.

Xie, Y. (2018b). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.21.