Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Основи кластеризације

(материјали за предавања)

Садржај

1. Кластеризација
2. Кластеризација поделом
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Додатак
6. Извори и литература

Кластеризација

- кластер
  - група међусобно сличних (блиских) ентитета
    - клика, грозд, рој, јато
  - ентитети из истог кластера начелно су међусобно сличнији (блискији) него што су слични (блиски) ентитетима из других кластера
- кластеризација
  - поступак распоређивања посматраних ентитета у кластере
  - распоређивање ентитета на основу њихових карактеристика

Кластеризација

- дескриптивни поступак
- ненадгледано учење
- истражна анализа података
- кластеризација по врсти ентитета
  - кластеризација појава
    - формирање група појава на основу вредности обележја тих појава
  - кластеризација обележја

Скупови података коришћени у примерима

- скуп података **`abalon`**
  - скуп података *Abalone*
    - аутори *Warwick Nash*, *Tracy Sellers*, *Simon Talbot*, *Andrew Cawthorn* и *Wes Ford*
    - електронска локација (доступни скуп података и пратеће информације)
      - https://archive.ics.uci.edu/dataset/1/abalone
      - https://doi.org/10.24432/C55C7W
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 30. 11. 1995.
    - датотека *abalone.zip* (преузето 27. 3. 2024)
      - електронска локација
        - https://archive.ics.uci.edu/static/public/1/abalone.zip
      - подаци у датотеци *abalone.data*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Abalone* читан, обрађиван и анализиран језиком *R*
    - активности над скупом података и резултати представљени у наставку
    - датотеке визуализација генерисане за формат *TIFF* (15,5 x 15,5 cm или 20,5 x 15,5 cm, 300 *DPI*)

Скупови података коришћени у примерима

- скуп података **`abalon`**
  - подаци о абалонима с Тасманије (Аустралија)
  - 4177 записа
  - 9 обележја
  - пол, дужина, пречник, висина, различите тежине и број прстенова
  - коришћен скуп података *Abalone*
  - подаци из датотеке *abalone.data* након учитавања су даље обрађивани и анализирани

Скупови података коришћени у примерима

скуп података **abalon** – припрема

```
# install.packages("tidyverse")

library(readr)
library(dplyr)
library(magrittr)
```

УЛАЗ

Скупови података коришћени у примерима

скуп података **abalon** – припрема

```
abalon <- read_csv("abalone.data",
                   col_names=c("sex", "length", "diameter",
                               "height", "weight_whole",
                               "weight_shucked", "weight_viscera",
                               "weight_shell", "rings"),
                   col_types="fddddddddi")

abalon %<>%
  mutate(id=1:nrow(abalon), age=rings + 1.5) %>%
  select(id, everything())

set.seed(4)
```

УЛАЗ

Садржај

1. Кластеризација
2. **Кластеризација поделом**
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Додатак
6. Извори и литература

Кластеризација поделом

Метод средина

- метод *k* средина
  - енгл. *k-Means*
- формирање партиције скупа ентитета која обухвата *k* подскупова
  - сваки подскуп представља један кластер
    - сваки подскуп је представљен једном средином (центроидом)
    - ентитет је члан кластера чијој је средини најсличнији (најближи)
  - *k* је параметар чију вредност треба унапред задати
    - *k* одговара броју кластера

Кластеризација поделом

Метод средина

метод *k* средина

разноликост унутар кластера треба да буде што мања

унутаркластерска варијабилност за кластер C

$$W(C)=\frac{1}{|C|}\sum_{i,j\in C}\sum_{l=1}^{p}(x_{il}-x_{jl})^2$$

p је број карактеристика (број обележја у случају кластеризације појава)

укупна унутаркластерска варијабилност за скуп кластера K

$$\sum_{C\in K}\frac{1}{|C|}\sum_{i,j\in C}\sum_{l=1}^{p}(x_{il}-x_{jl})^2$$

Кластеризација поделом

Метод средина

- метод k средина – формирање кластера
  - насумично распоредити ентитете у k почетних кластера
  - побољшавати квалитет k кластера
    - за сваки кластер одредити средину (центроид)
    - ентитете распоредити у оне кластере чије су им средине најсличније (најближе)
      - еуклидска удаљеност као мера различитости
    - понављати одређивање средина и распоређивање ентитета докле год долази до смањења укупне унутаркластерске варијабилности за k кластера

Кластеризација поделом

Метод средина

- метод k средина – мањкавости
  - могућа конвергенција према локалном оптимуму
  - осетљивост на почетни распоред по кластерима
    - могуће размотрити већи број почетних стања
  - осетљивост на скале карактеристика
    - могуће спровести центрирање и скалирање
  - потребно одредити број кластера k
    - постоје помоћни поступци за испитивање погодности одређеног броја кластера
- могуће је увести кориговани критеријум завршетка
  - ограничити број итерација

Кластеризација поделом

Метод средина̂

пример

```
abalon.km <- abalon %>%
  select(length, height)

k <- 3

km <- abalon.km %>%
  select(length, height) %>%
    kmeans(centers=k)
```

УЛАЗ

Кластеризација поделом

Метод средина̂

пример

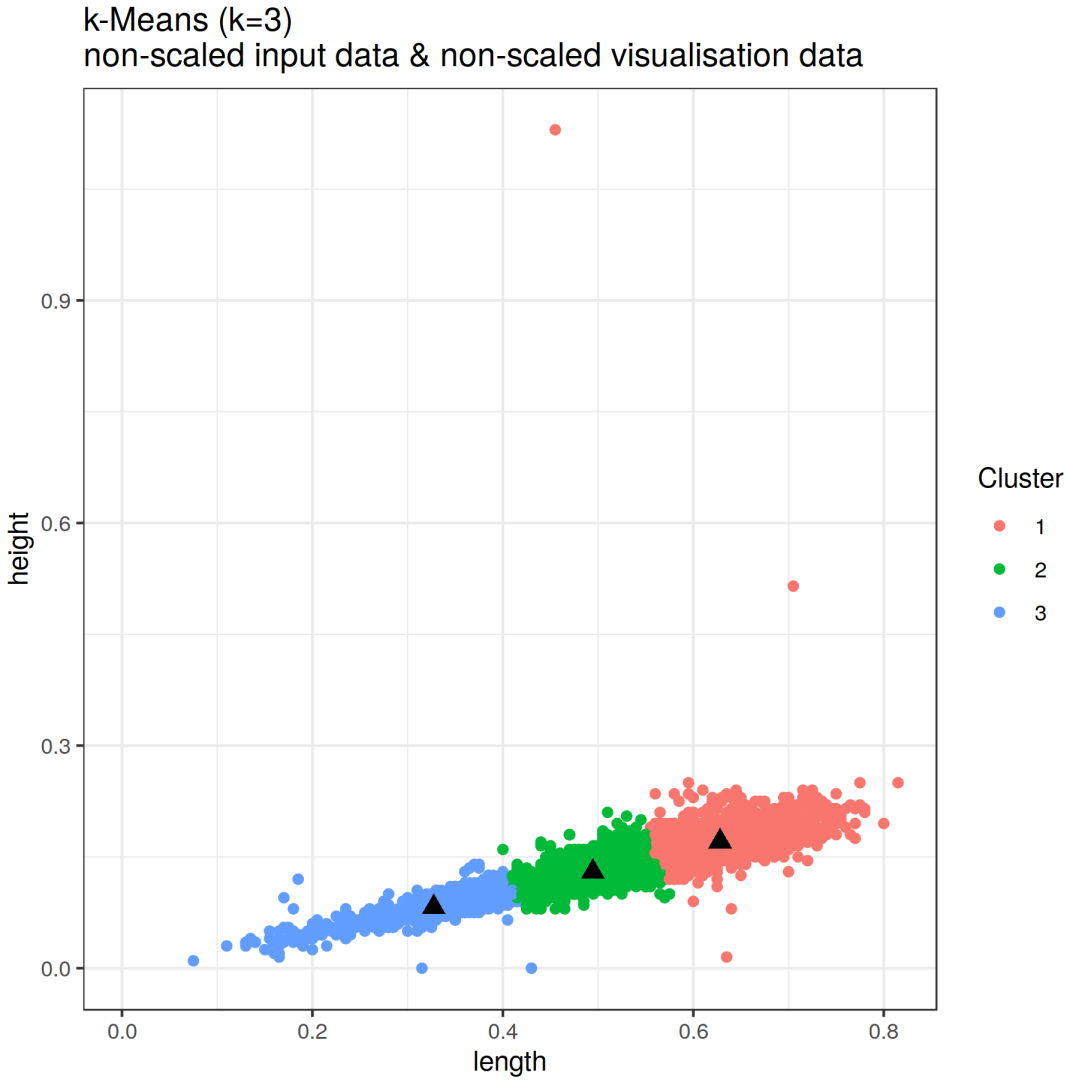
```
> tail(km$cluster, 20)
 [1] 2 2 2 1 1 3 3 3 3 2 2 2 2 2 1 1 1 1 1 1
> km$centers
     length     height
1 0.6281112 0.17082029
2 0.4945503 0.12985117
3 0.3274228 0.08214094
> km$withinss
[1] 5.960992 3.350221 3.563362
>
```

КОНЗОЛА

Кластеризација поделом

Метод средина̂

пример



Кластеризација поделом

Метод средина

пример

```
 1 abalon.km <- abalon %>%
 2   select(length, height)
 3
 4 k <- 3
 5
 6 km.skal <- abalon.km %>%
 7   select(length, height) %>%
 8   scale(center=T, scale=T) %>%
 9   kmeans(centers=k)
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Кластеризација поделом

Метод средина̂

пример

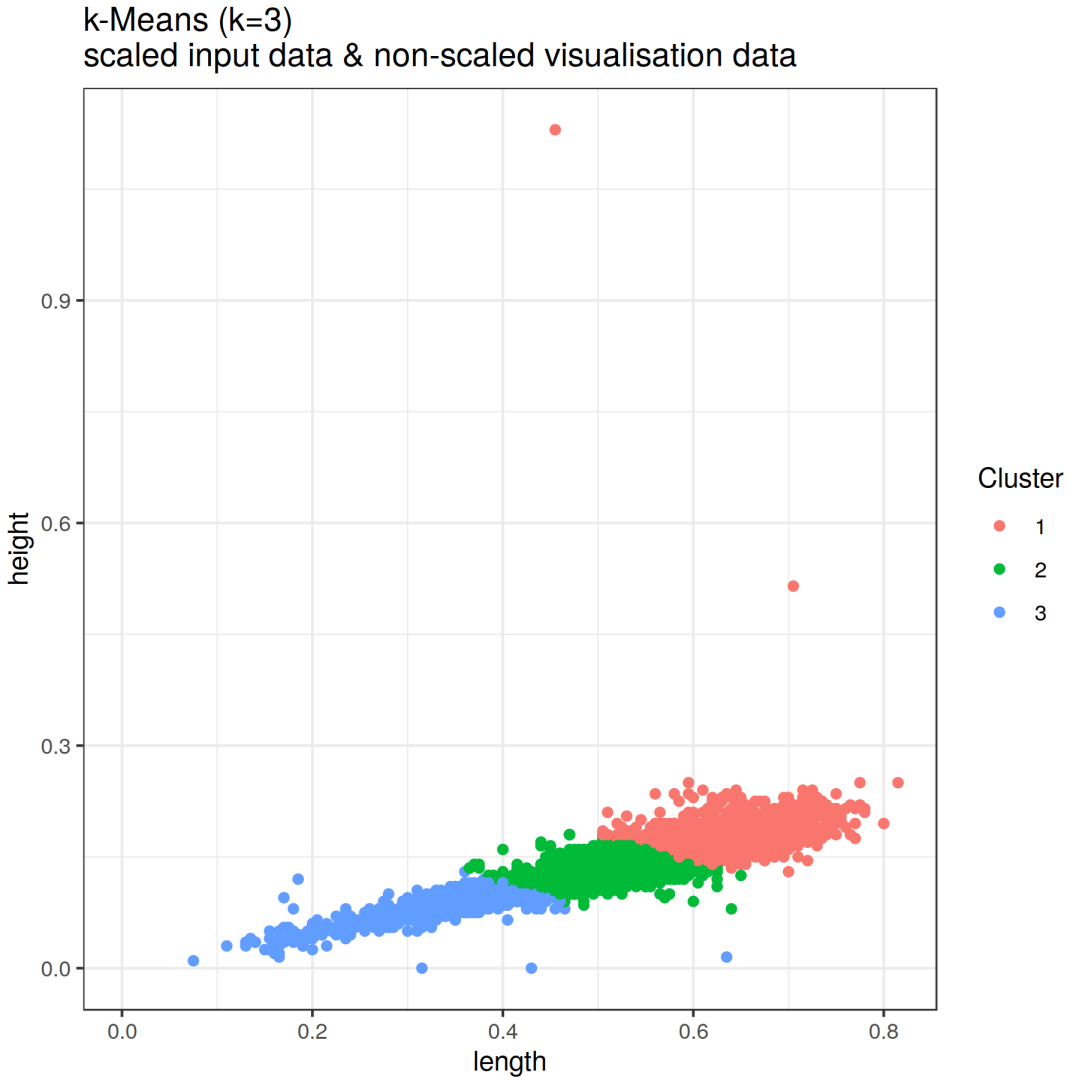
```
> tail(km.skal$cluster, 20)
 [1] 2 2 2 1 1 3 3 3 3 2 2 2 2 2 2 1 2 1 1 1
> km.skal$centers
      length     height
1  0.9000298  0.8700491
2 -0.1245050 -0.1909469
3 -1.5508725 -1.3529666
> km.skal$withinss
[1] 1209.4235  519.2548  464.3167
>
```

КОНЗОЛА

Кластеризација поделом

Метод средина̂

пример



Кластеризација поделом

Метод средина̂

пример

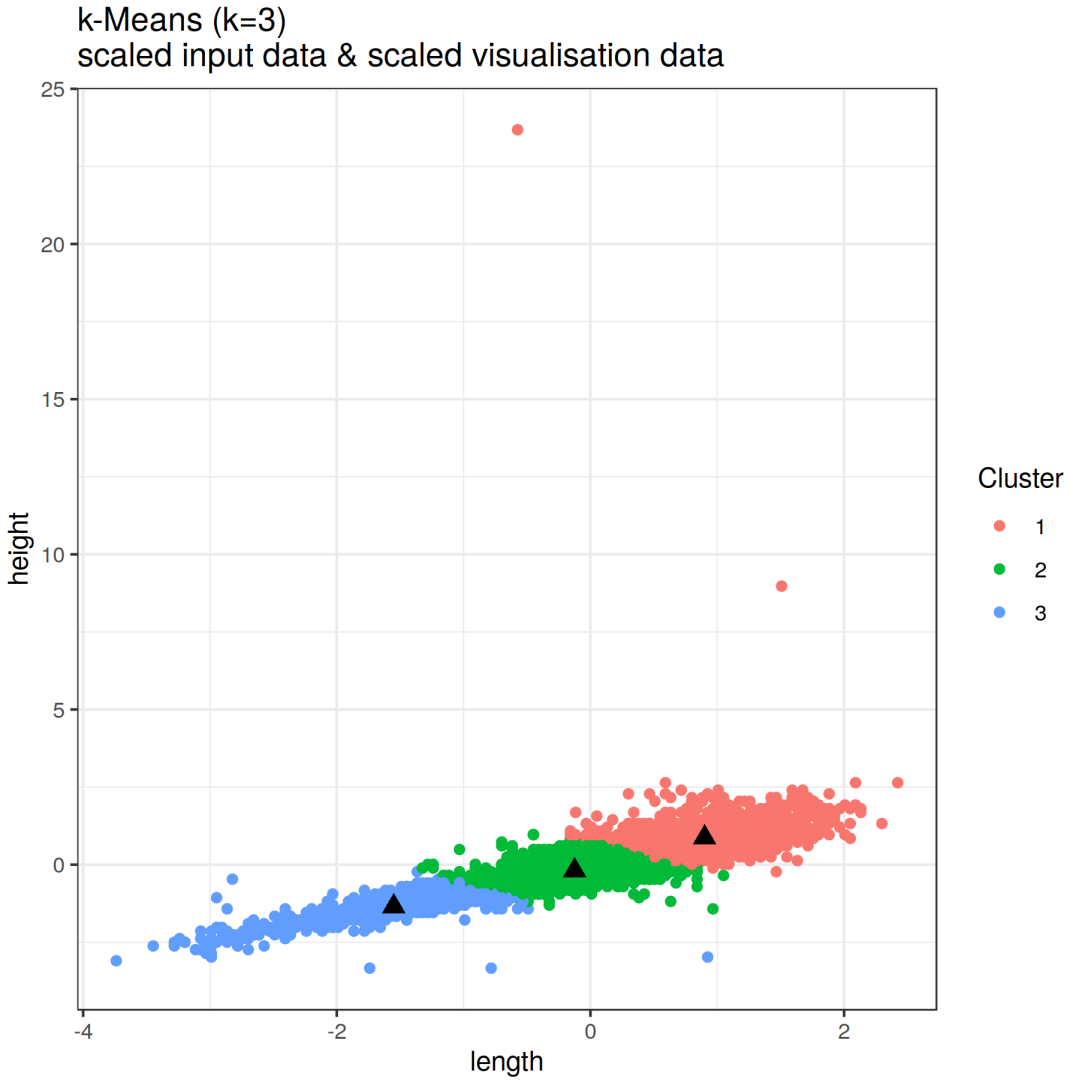
```
abalon.km.skal <- abalon.km %>%
  select(length, height) %>%
  scale(center=T, scale=T) %>%
  as.data.frame()
```

УЛАЗ

Кластеризација поделом

Метод средина̂

пример



Садржај

1. Кластеризација
2. Кластеризација поделом
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Додатак
6. Извори и литература

Хијерархијска кластеризација

- формирање кластера који су хијерархијски организовани
  - кластер на дну хијерархије обухвата један ентитет
  - кластер који није на дну хијерархије обухвата све ентитете својих подређених кластера
- из хијерархије начелно могуће издвајати различите бројеве кластера
  - за формирање хијерархије кластера није неопходно задати број кластера
- представа хијерархије кластера кроз дендограм
- две варијанте хијерархијске кластеризације
  - кроз поступак спајања (енгл. *agglomerative*)
    - *agnes* (*agglomerative nesting*)
  - кроз поступак дељења (енгл. *divisive*)
    - *diana (divisive analysis)*

Хијерархијска кластеризација

- хијерархијска кластеризација кроз спајање – формирање кластера
  - за сваки ентитет формирати засебан кластер
  - спојити два најсличнија (најближа) кластера
    - сличност (блискост) утврдити на основу изабране мере различитости
    - спајање најсличнијих кластера понављати док се не оствари постојање само једног кластера
      - на основу извршених спајања могуће је формирати одговарајуће стабло и представити га кроз дендограм
        - сваки чвор одговара једном кластеру формираном током кластеризације
          - корен одговара кластеру са свим ентитетима
          - лист одговара неком од могућих кластера са једним ентитетом

Хијерархијска кластеризација

- хијерархијска кластеризација кроз спајање – формирање кластера
  - повезаност (енгл. *linkage*)
    - мера различитости између две групе ентитета
    - често коришћене врсте повезаности
      - појединачна (енгл. *single*)
        - одговара најмањој вредности различитости уоченој између нека два ентитета из различитих група
      - средња (енгл. *average*)
        - одговара средњој вредности различитости уочених између ентитета из различитих група
      - потпуна (енгл. *complete*)
        - одговара највећој вредности различитости уоченој између нека два ентитета из различитих група

Хијерархијска кластеризација

- хијерархијска кластеризација кроз спајање – формирање кластера
  - мере различитости између ентитета
    - мере засноване на удаљености
      - еуклидска удаљеност
      - Менхетн удаљеност
      - Махаланобисова удаљеност
      - ...
    - мере засноване на корелацији

Хијерархијска кластеризација

пример

```
1  abalon.hc <- abalon %>%
2    filter(sex=="M") %>%
3    select(weight_whole, rings)
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Хијерархијска кластеризација

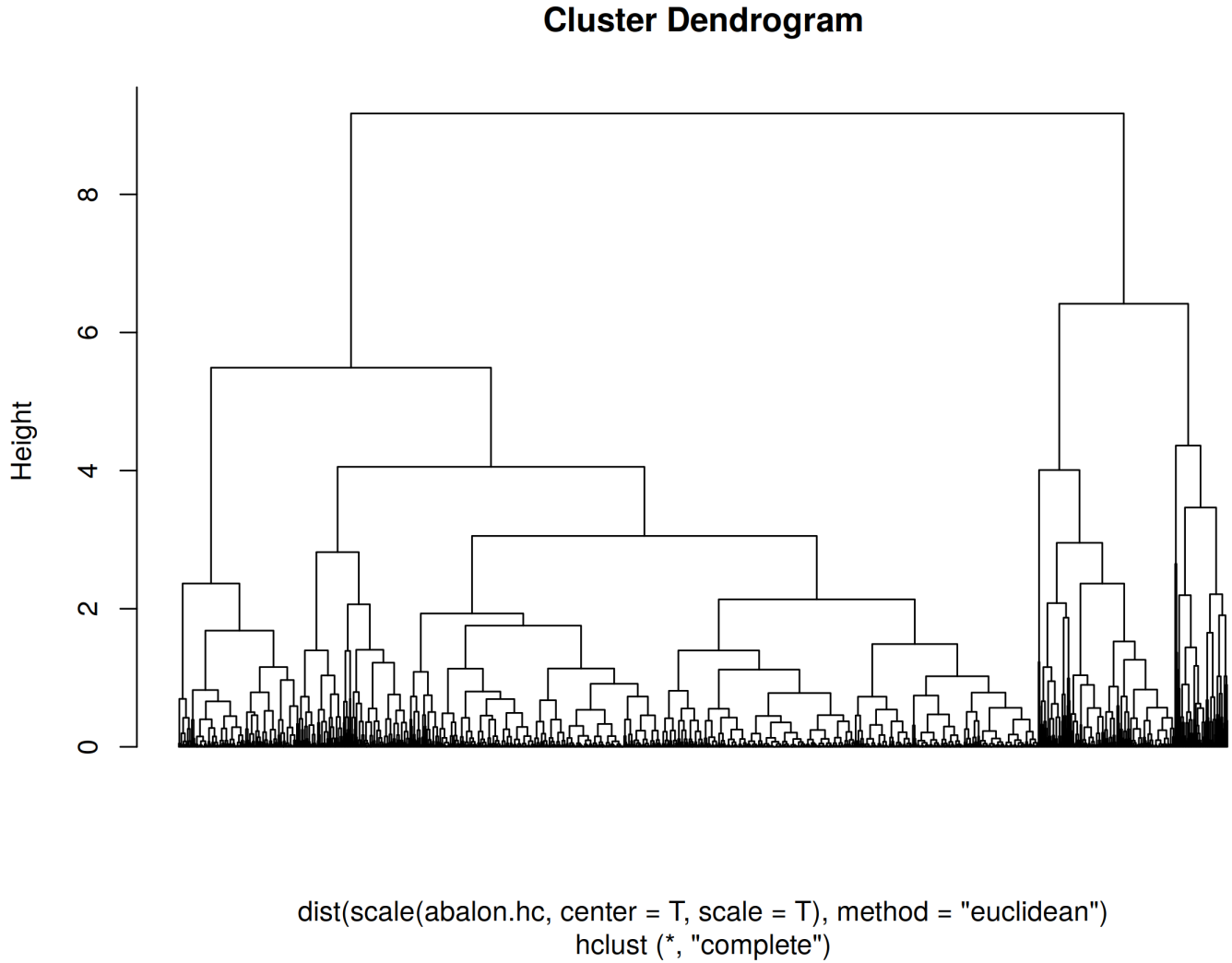
пример

```
1  hc.cmp.euc <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                       method="euclidean"), method="complete")
3
4  hc.cmp.euc.clust <- cutree(hc.cmp.euc, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

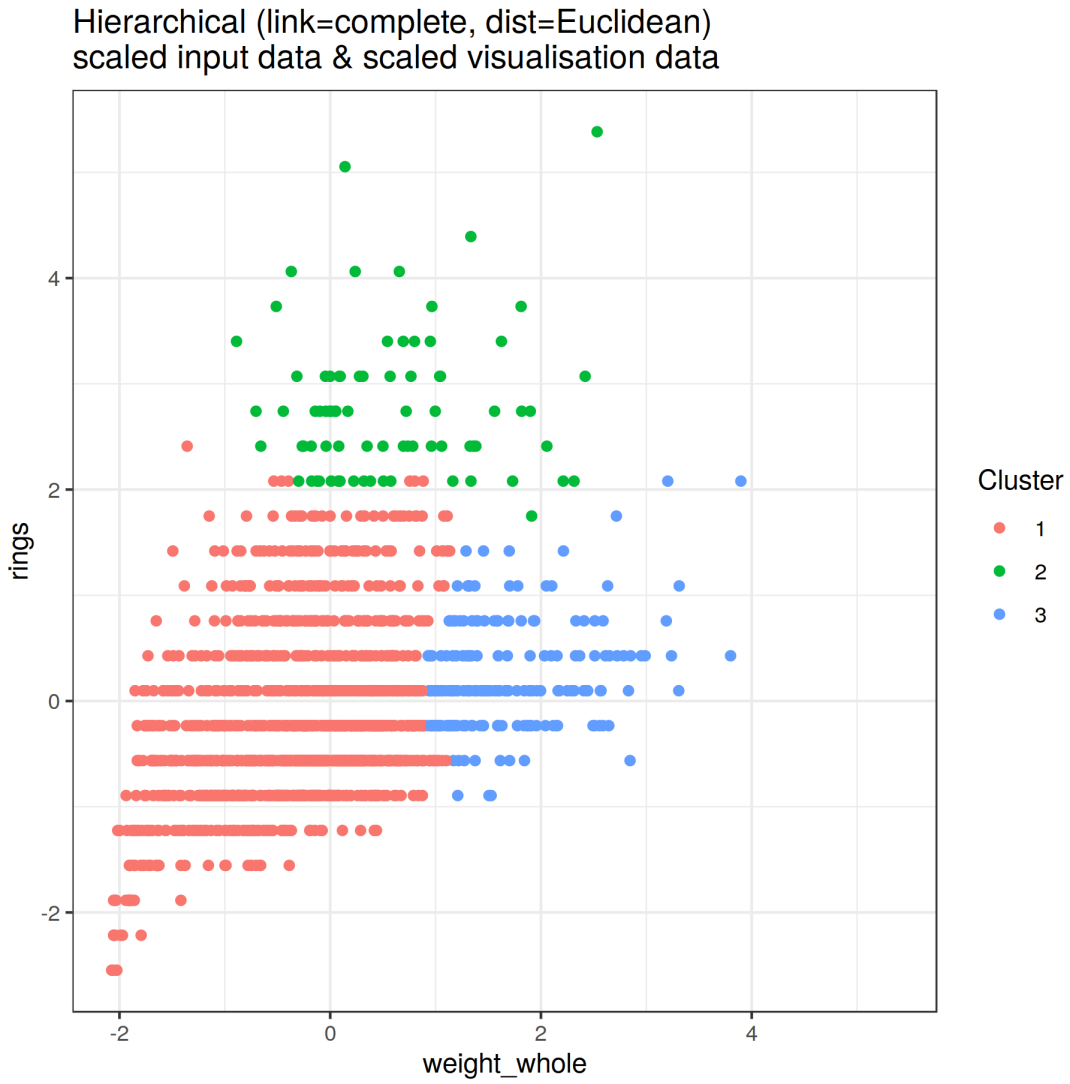
Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример



Хијерархијска кластеризација

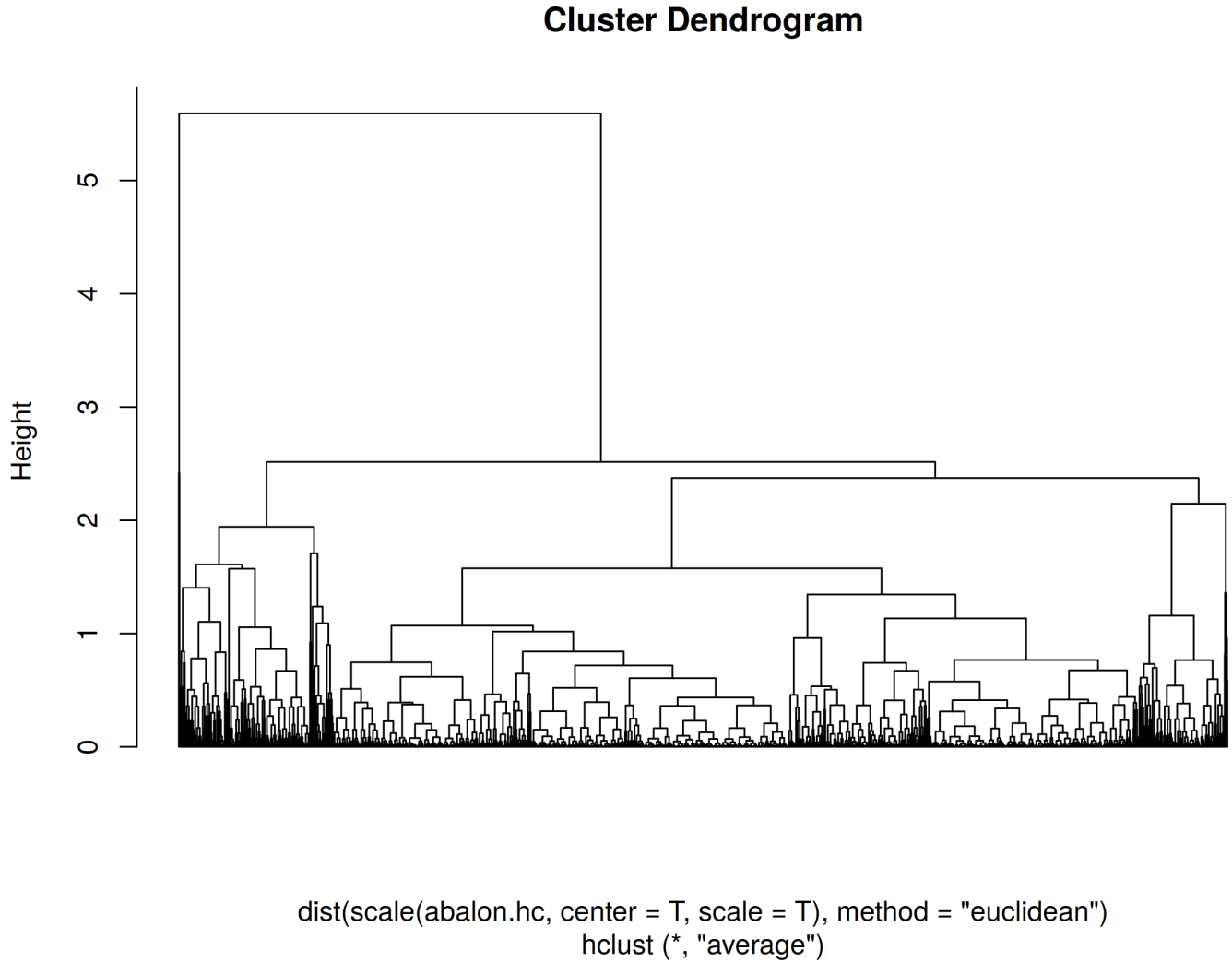
пример

```
1  hc.avg.euc <- hclust(dist(scale(abalon.hc, center=T, scale=T),
2                       method="euclidean"), method="average")
3
4  hc.avg.euc.clust <- cutree(hc.avg.euc, k=3)
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

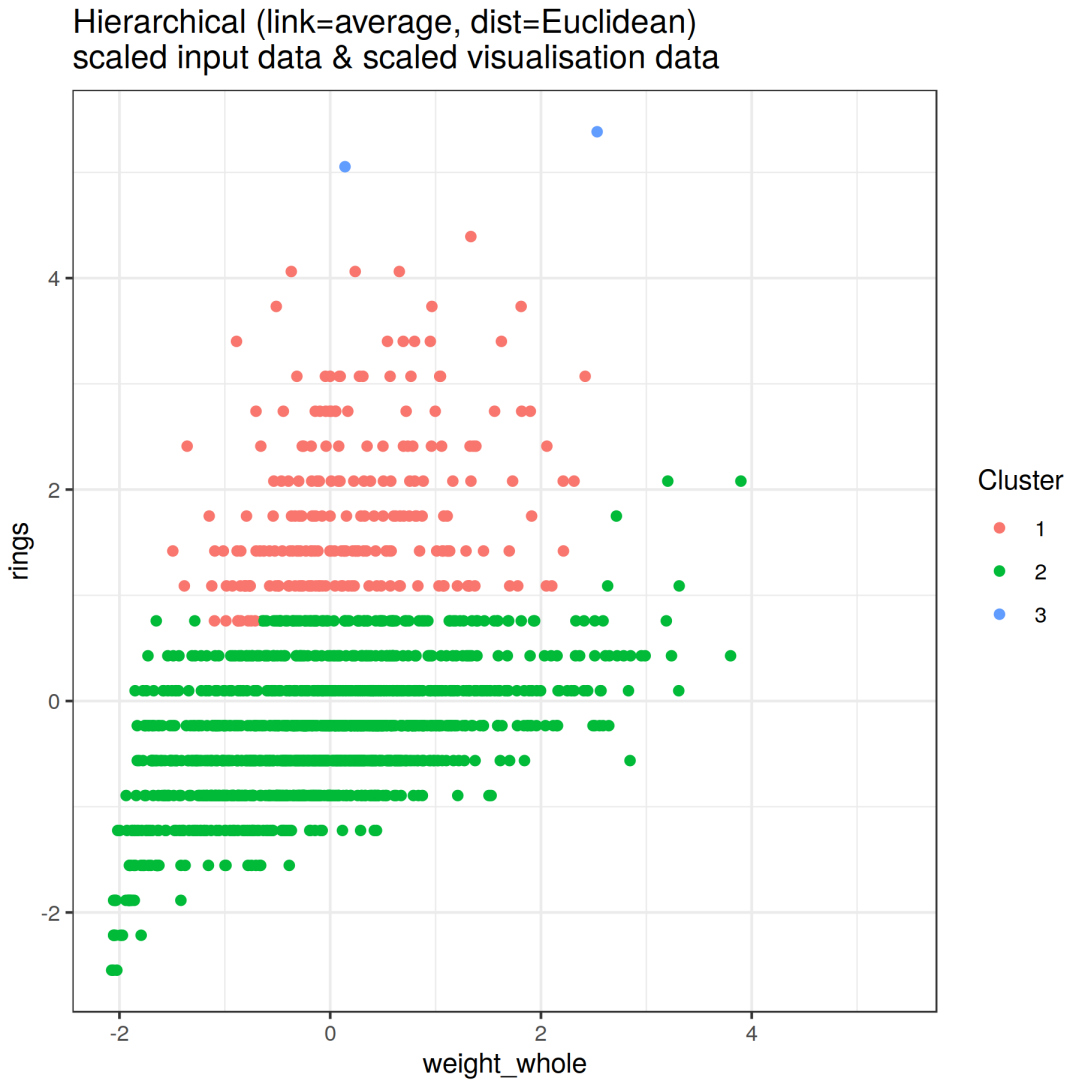
Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример

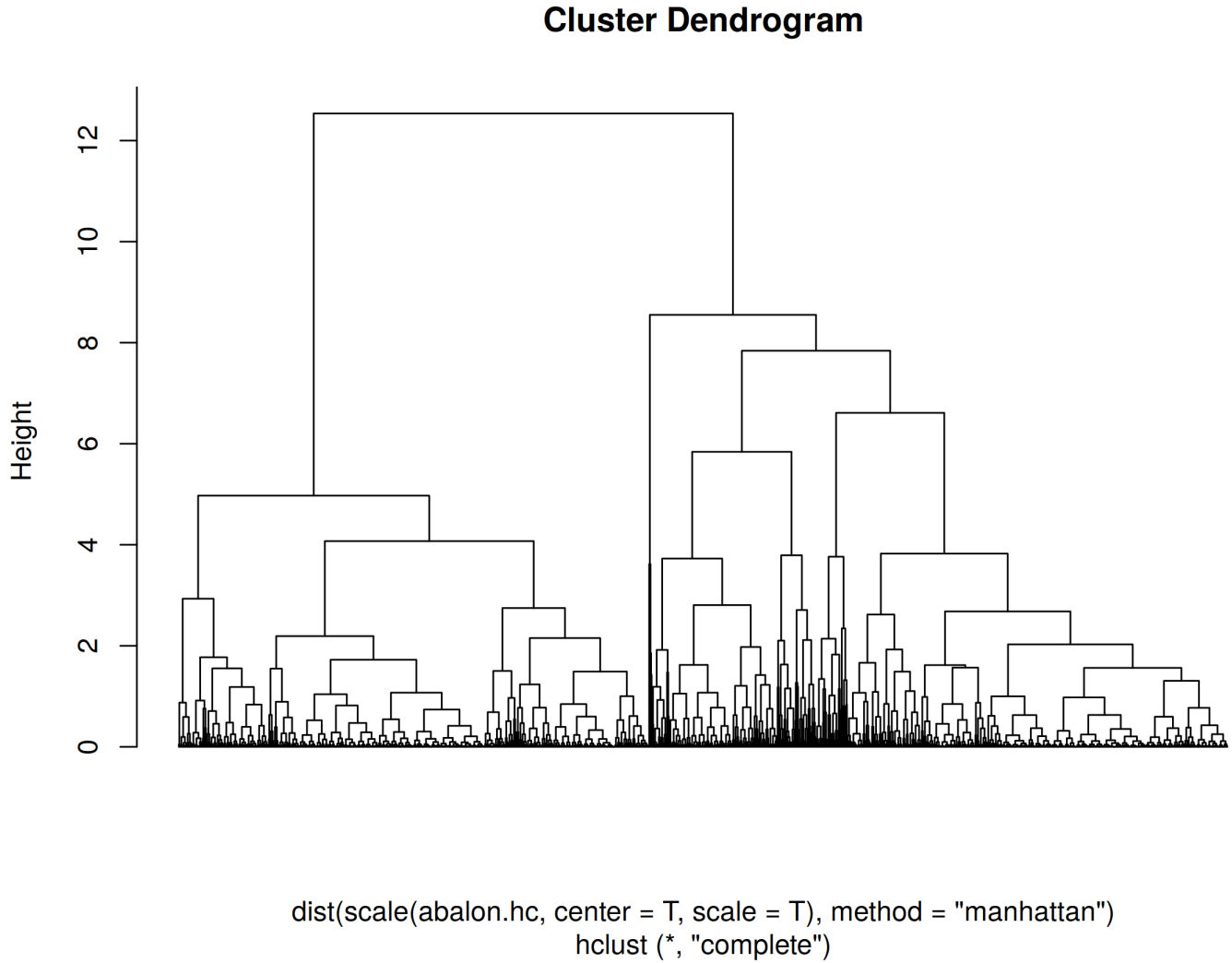
```
hc.cmp.man <- hclust(dist(scale(abalon.hc, center=T, scale=T),
                          method="manhattan"), method="complete")

hc.cmp.man.clust <- cutree(hc.cmp.man, k=3)
```

УЛАЗ

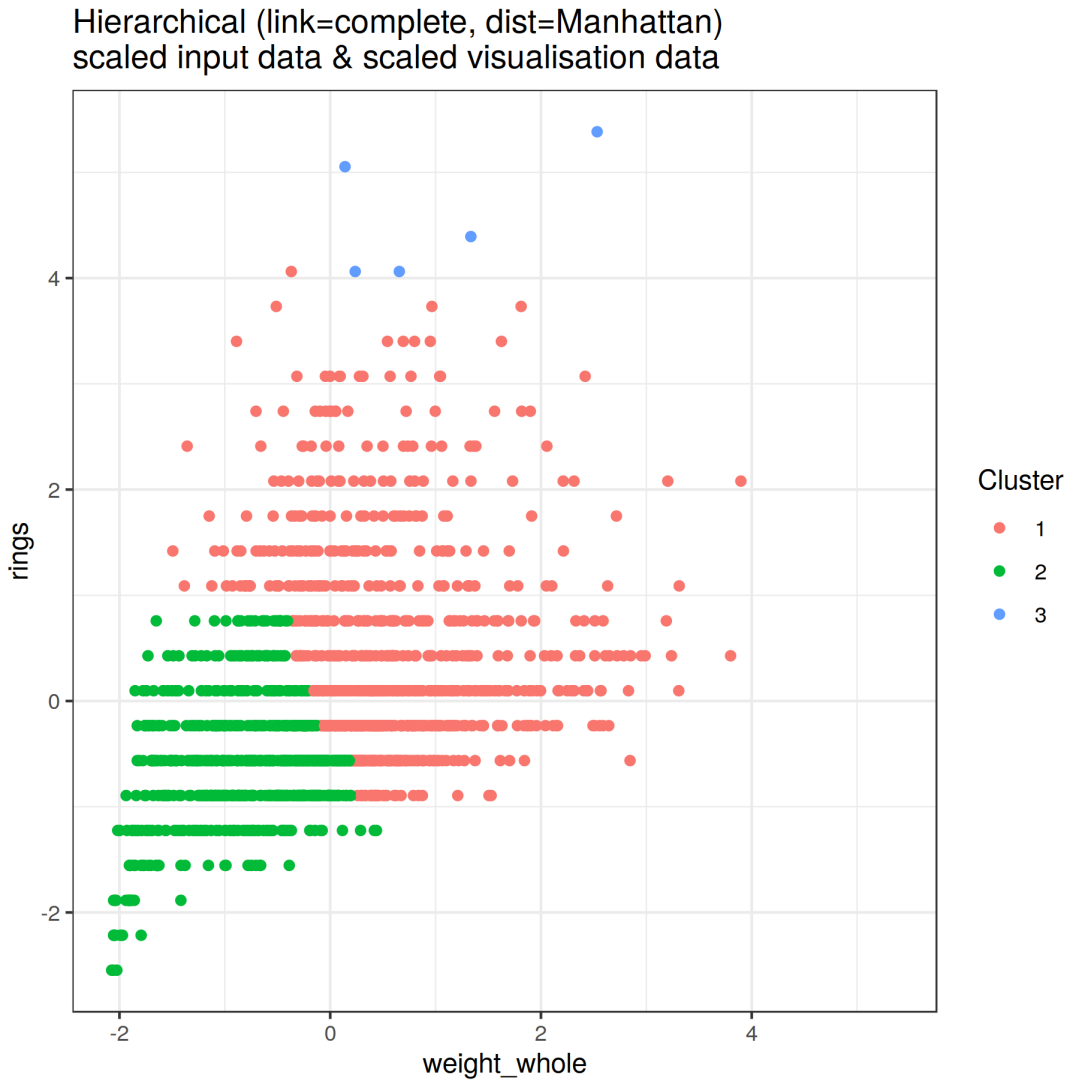
Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример

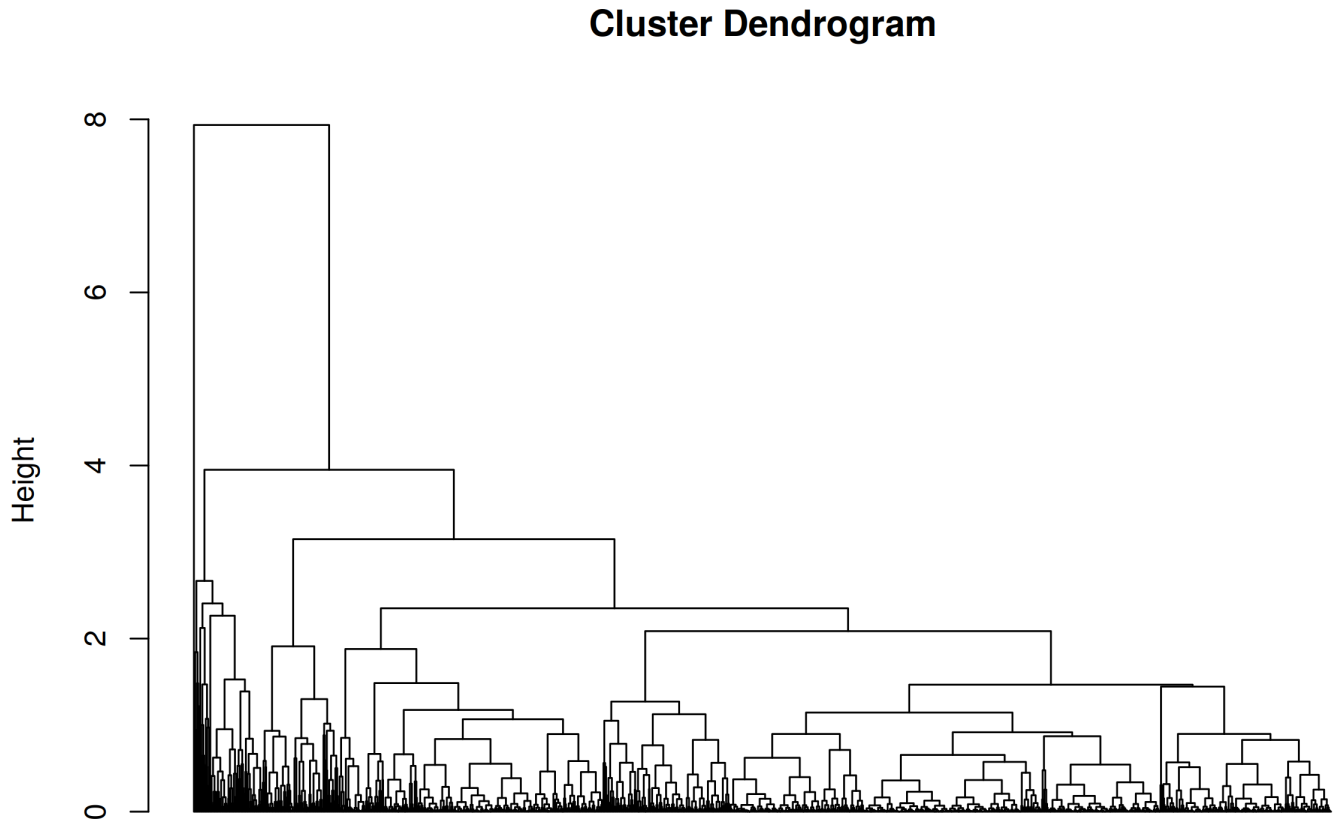
```
hc.avg.man <- hclust(dist(scale(abalon.hc, center=T, scale=T),
                          method="manhattan"), method="average")

hc.avg.man.clust <- cutree(hc.avg.man, k=3)
```

УЛАЗ

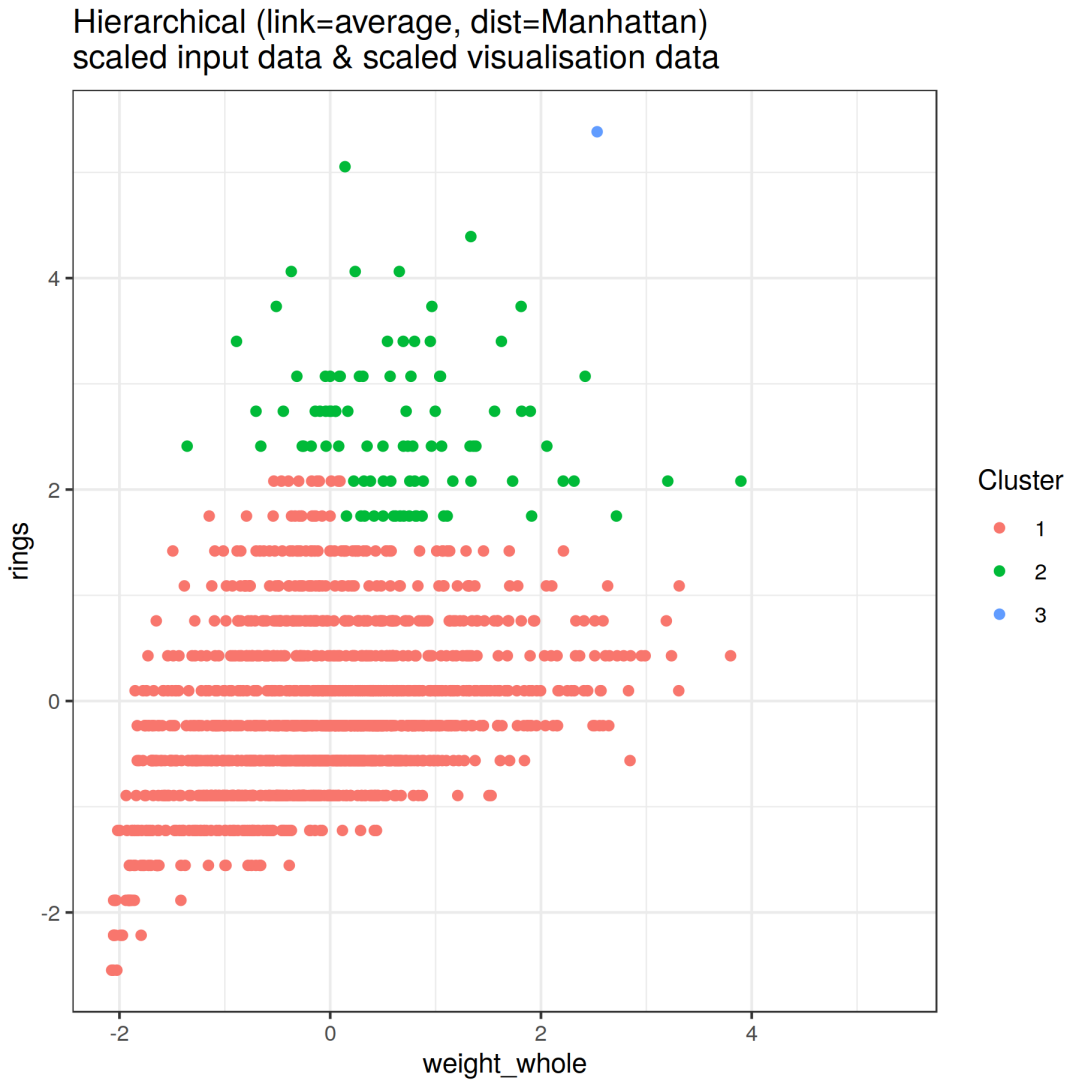
Хијерархијска кластеризација

пример



Хијерархијска кластеризација

пример



Садржај

1. Кластеризација
2. Кластеризација поделом
3. Хијерархијска кластеризација
4. **Густинска кластеризација**
5. Додатак
6. Извори и литература

Густинска кластеризација

- кластер као густо попуњен повезани део простора
- алгоритам *DBSCAN* (*Density-based spatial clustering of applications with noise*)
  - ентитети су у истом кластеру ако су међусобно густински достижни
    - густина се посматра у ε околини неког ентитета
    - битно је да ли у постоји одређени минимални број ентитета у ε околини посматраног ентитета
    - три врсте ентитета
      - основни ентитет
        - ентитет у чијој се ε околини налази барем минимални број ентитета
      - гранични ентитет
        - није основни ентитет али јесте у ε околини основног ентитета
      - шум
        - није ни основни ни гранични ентитет
  - параметри
    - ε као величина околине (позитиван реалан број)
    - минимални број ентитета за ε околину (позитиван цео број)
      - ентитет се увек налази у сопственој ε околини

Густинска кластеризација

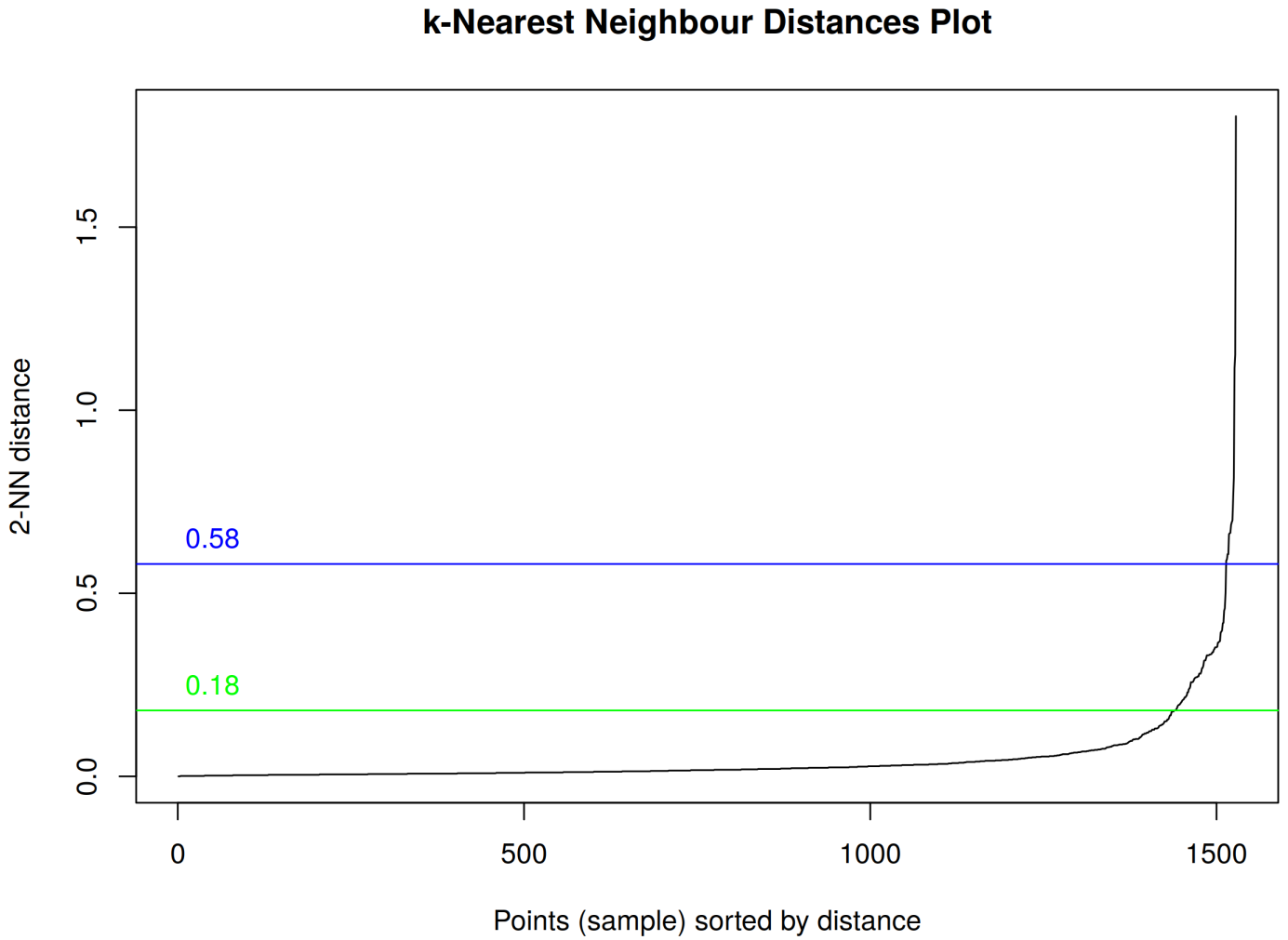
пример

```
 1 library(dbscan)
 2
 3 abalon.dbs <- abalon %>%
 4   filter(sex=="M") %>%
 5   select(weight_whole, rings)
 6
 7
 8
 9
10
11
12
13
14
15
16
17
18
19
20
```

УЛАЗ

Густинска кластеризација

пример



Густинска кластеризација

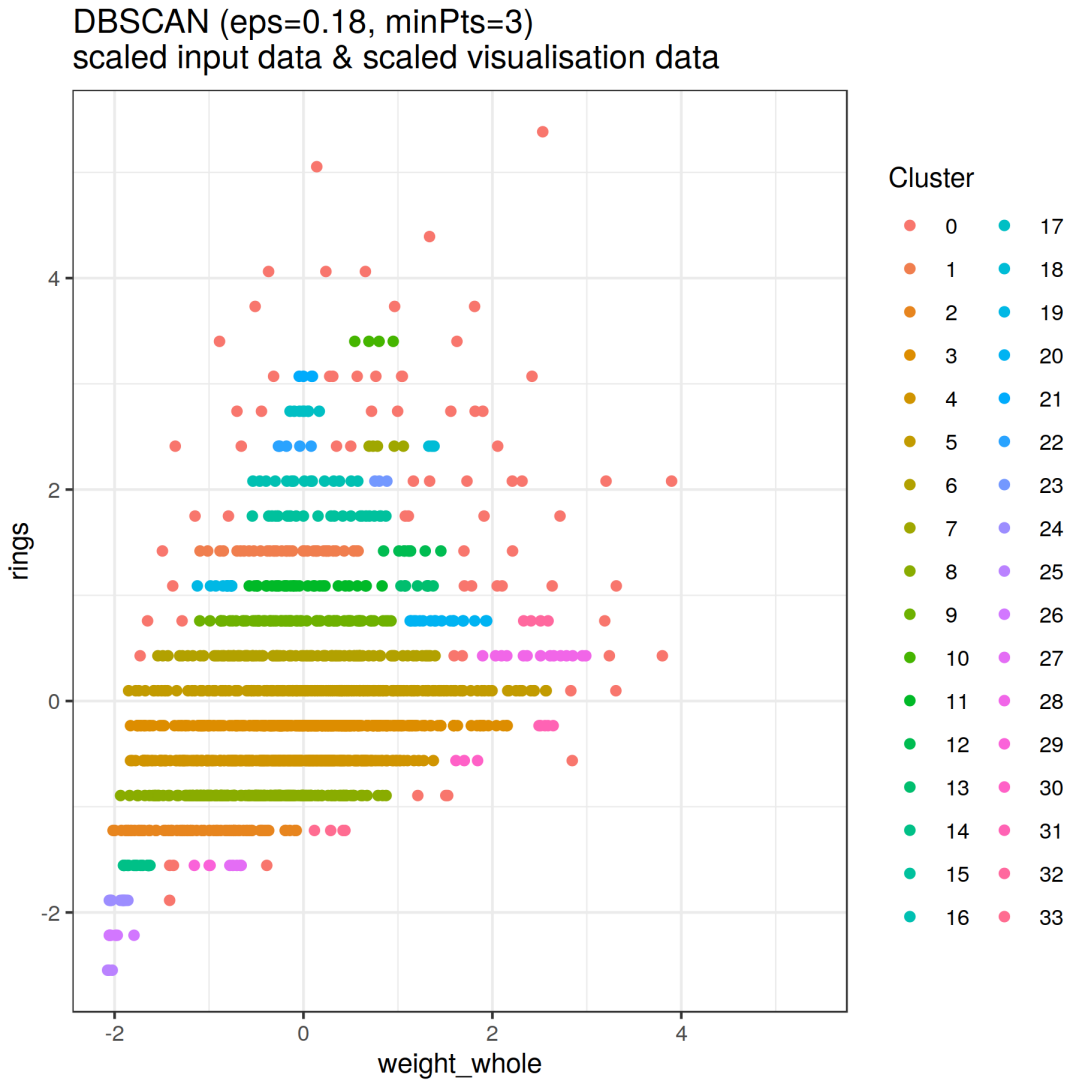
пример

```
dbs.low <- dbscan(scale(abalon.dbs, center=T, scale=T),
                  eps=0.18, minPts=3)
```

УЛАЗ

Густинска кластеризација

пример



Густинска кластеризација

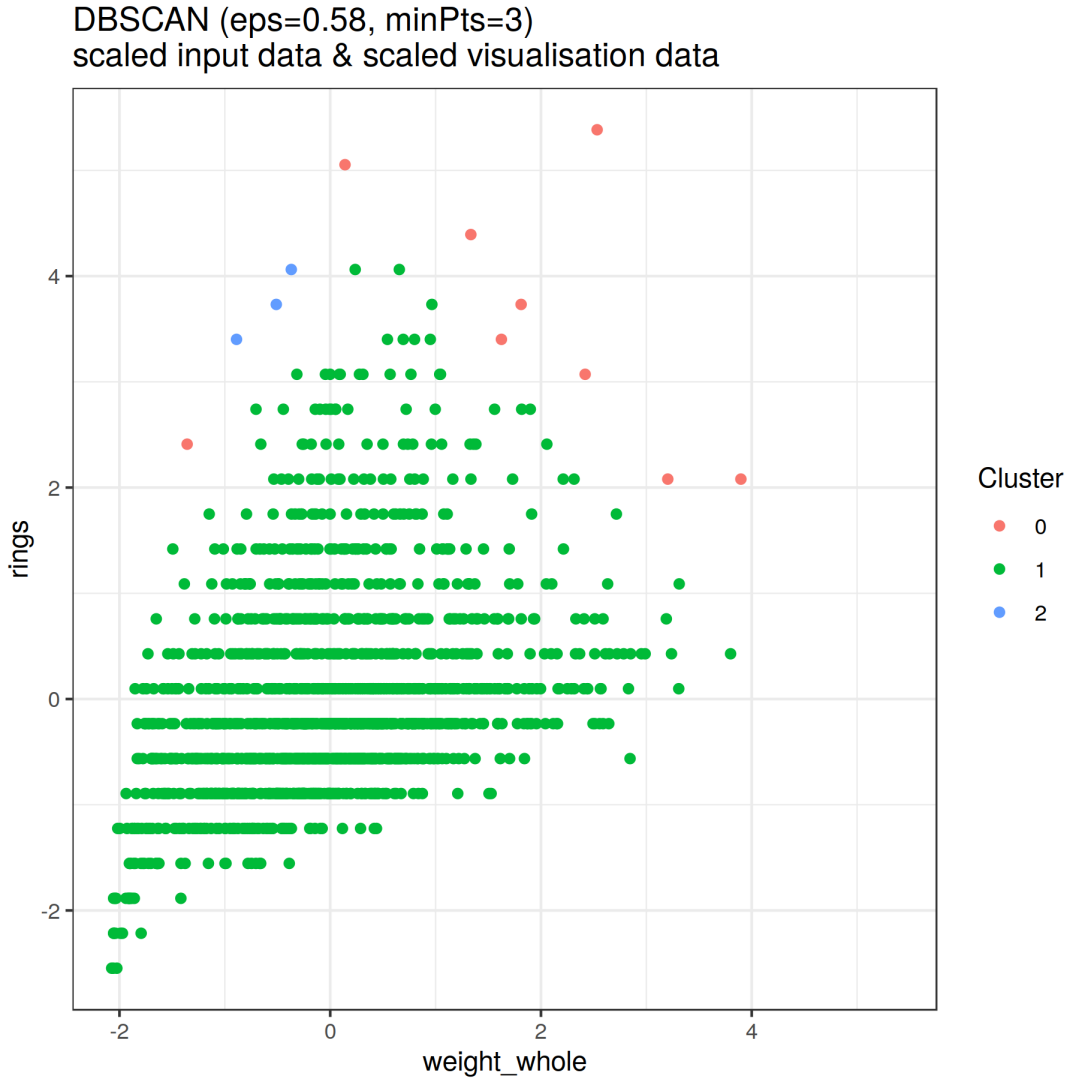
пример

```
dbs.high <- dbscan(scale(abalon.dbs, center=T, scale=T),
                   eps=0.58, minPts=3)
```

УЛАЗ

Густинска кластеризација

пример



Садржај

1. Кластеризација
2. Кластеризација поделом
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. **Додатак**
6. Извори и литература

Додатак

Функције удаљености

удаљеност Минковског

$$\left(\sum_{l=1}^{p} \left| x_{il} - x_{jl} \right|^{n} \right)^{1/n}$$

у случају $n = 1$, Менхетн удаљеност
у случају $n = 2$, еуклидска удаљеност

стандардизована еуклидска удаљеност

$$\left(\sum_{l=1}^{p} \left| \frac{x_{il} - x_{jl}}{\sigma_l} \right|^{2} \right)^{1/2}$$

Додатак

Временска сложеност

- метод k средина
  - $O(kst)$ – ниска
    - k је број кластера
    - s је број ентитета
    - t је број итерација
- алгоритам *DBSCAN*
  - $O(s \log s)$ – умерена
    - s је број ентитета

Додатак

Бисекциони метод срединâ

- бисекциони метод *k* срединâ
  - енгл. *bisecting k-Means*
  - хијерархијска кластеризација кроз дељење
    - хијерархијска кластеризација с кластеризацијом поделом (метод *k* срединâ) почиње се од дељења кластера у којем су садржане све појаве на два и такво дељење може бити даље примењивано рекурзивно на нове кластере, тако да буде *k* крајњих кластера и да је свако појединачно дељење извршено применом метода *k* срединâ

Бисекциони метод синâ

бисекциони метод *k* синâ

формирање кластера

- одабрати кластер који треба да буде подељен
  - нпр. кластер који је тренутно највећи или кластер који има најмању општу сличност
- формирати два подређена кластера применом основног метода *k* синâ (корак бисекције)
- понављати корак бисекције за одређени број итерација и одабрати ону поделу која резултује кластеризацијом највеће опште сличности
- понављати претходне кораке све док не буде формиран потребни број кластера

главни параметри

- *k* је потребни (циљни) број кластера
- *iter* је број понављања корака бисекције за један кластер

могућност паралелизације

- кораци бисекције за кластере који су на истом нивоу могу бити груписани

Садржај

1. Кластеризација
2. Кластеризација поделом
3. Хијерархијска кластеризација
4. Густинска кластеризација
5. Додатак
6. **Извори и литература**

Извори и литература

Основни извори и литература

- James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. Springer; 2013.
- Xu D, Tian Y. A comprehensive survey of clustering algorithms. Annals of Data Science. 2015;2(2); 165–193.
- R Project. R: A language and environment for statistical computing – Reference index – The R Core Team – Version 4.5.1 (2025-06-13). Internet: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf
- R Project. CRAN: Package dbscan. Internet: https://cran.r-project.org/web/packages/dbscan/index.html
- Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Technical report TR 00-034. Department of Computer Science and Engineering, University of Minnesota; 2000.
- Apache Spark. Clustering. Internet: https://spark.apache.org/docs/latest/ml-clustering.html

Извори и литература

Основни извори података

- скуп података **`abalon`**
  - скуп података *Abalone*
    - аутори *Warwick Nash*, *Tracy Sellers*, *Simon Talbot*, *Andrew Cawthorn* и *Wes Ford*
    - електронска локација (доступни скуп података и пратеће информације)
      - https://archive.ics.uci.edu/dataset/1/abalone
      - https://doi.org/10.24432/C55C7W
      - репозиторијум *UC Irvine Machine Learning Repository* на електронској локацији http://archive.ics.uci.edu/
        - скуп података дониран за репозиторијум 30. 11. 1995.
    - датотека *abalone.zip* (преузето 27. 3. 2024)
      - електронска локација
        - https://archive.ics.uci.edu/static/public/1/abalone.zip
      - подаци у датотеци *abalone.data*
    - лиценца *Creative Commons Attribution 4.0 International (CC BY 4.0)*
      - електронска локација
        - https://creativecommons.org/licenses/by/4.0/legalcode
  - скуп података *Abalone* читан, обрађиван и анализиран језиком *R*

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси у информационом инжењерингу

Основи кластеризације

(материјали за предавања)