

Содержание

Введение	2
1 SQL-запросы	2
2 Загрузка данных в R	3
3 Анализ данных	3
3.1 Выбор групп и периода для анализа	3
3.2 Кластеризация	4
3.2.1 Построение древовидной диаграммы кластеризации	4
3.2.2 Выбор дисциплин для визуализации k-means	7
3.2.3 Визуализация k-means	8
3.3 Статистический анализ	9

Введение

Требуется проанализировать успеваемость студентов групп бакалавриата специальности «Информационные системы и технологии» 2011-2013 года поступления за 2 и 3 семестр. То есть, группы ИС-Б11, Б12, Б13.

1 SQL-запросы

```
//Извлекаем все группы по шаблону бакалавры ИС
SELECT * FROM sgroup WHERE shname LIKE 'ИС-Б__';

//Берем только их GID
SELECT gid FROM sgroup WHERE shname LIKE 'ИС-Б__';

//Извлекаем всех студентов из выбранных групп (Бакалавры ИС)
SELECT * FROM student WHERE gid IN (SELECT gid FROM sgroup WHERE shname LIKE 'ИС-Б__');

//Извлекаем информацию об успеваемости конкретного студента за 2-3 семестр
SELECT stid, discid, semnum, sum(ktmark) FROM studmark
    WHERE stid = '3844' AND (semnum >=2 AND semnum <=3)
    GROUP BY stid, discid, semnum;

//Полная информация об успеваемости студентов выбранных групп за 2-3 семестр
SELECT stid, semnum, discid, SUM(ktmark) FROM studmark
    WHERE stid IN (SELECT stid FROM student
        WHERE gid IN (SELECT gid FROM sgroup
            WHERE shname LIKE 'ИС-Б__'))
    AND (semnum >=2 AND semnum <=3)
    GROUP BY stid, semnum, discid
    ORDER BY stid;

//Полная информация об успеваемости + имена групп и дисциплин
=====
SELECT g.shname, info.gid, info.stid, p.discname,
    info.discid, info.semnum, info.sum FROM (
SELECT stid, discid, semnum, sum(ktmark) FROM studmark
    WHERE stid IN (SELECT stid FROM student
        WHERE gid IN (SELECT gid FROM sgroup
            WHERE shname LIKE 'ИС-Б__'))
    AND (semnum >=2 AND semnum <=3)
    GROUP BY stid, discid, semnum
    ORDER BY stid
) info
JOIN predmet p ON p.discid = info.discid
JOIN (select gid, stid from student) s ON s.stid=info.stid
JOIN (select gid, shname from sgroup) g ON s.gid=g.gid;
```

2 Загрузка данных в R

Исходный код скрипта загрузки данных, с последующей их записью в CSV файл

```
library(RPostgreSQL)
drv = dbDriver("PostgreSQL")
con = dbConnect(drv, dbname = "test", host = "127.0.0.1",
                user = "test", password = "123")

extractQuery <- "SELECT g.shname, info.gid, info.stid, p.discname, info.discid,
info.semnum, info.sum FROM (
SELECT stid, discid, semnum, sum(ktmark) FROM studmark
  WHERE stid IN (SELECT stid FROM student
    WHERE gid IN (SELECT gid FROM sgroup
      WHERE shname LIKE 'ИС-Б__'))
  AND (semnum >=2 AND semnum <=3)
  GROUP BY stid, discid, semnum
  ORDER BY stid
) info
JOIN predmet p ON p.discid = info.discid
JOIN (select gid, stid from student) s ON s.stid=info.stid
JOIN (select gid, shname from sgroup) g ON s.gid=g.gid;"

data <- dbGetQuery(con, extractQuery)
write.csv(data, file="/extractDataPR.csv", na="")

dbDisconnect(con)
dbUnloadDriver(drv)
```

Восстановление данных производится аналогично путем загрузки данных из CSV файла в dataframe.

3 Анализ данных

3.1 Выбор групп и периода для анализа

В качестве периода для анализа успеваемости был выбран 2 семестр, так как нет сведений оценок ИС-Б13 за 3 семестр. Первоначальным этапом анализа являлась кластеризация студентов по оценкам 2 семестра методом k-средних. Преобразование в широкую таблицу:

```
isB13sem2 <- dcast(cutGroup(333,2), stid ~ discid)
```

содержит N/A значение, более того в этой группе меньше всего студентов. Поэтому группа ИС-Б13 не кластеризуется.

```

library(dplyr)
library(ggplot2)
library(reshape2)

adata <- tbl_df(read.csv(file = "~/extractDataPR.csv")[,2:8])

cutGroup <- function(agid, asem) {
  g <- filter(adata, gid == agid & semnum == asem)
  cg <- g[,c("stid", "discid", "sum")]
  return(cg)
}

B11 <- cutGroup(222,2)
B12 <- cutGroup(278,2)

```

3.2 Кластеризация

В качестве метода кластеризации использовался метод k-средних(k-means). Данный метод применялся для выявления классов студентов со схожей успеваемостью по евклидовой метрике оценок.

3.2.1 Построение древовидной диаграммы кластеризации

Была построена древовидная диаграмма кластеров отдельно для каждой из 2 исследуемых групп. Таким образом, из их построения видно, что выбранное число кластеров обоснованно и последующий алгоритм процедуры kmeans может разделить данные на 3 кластера. Красным пунктиром обозначена возможная линия разделения на кластеры.

```

isB11sem2 <- dcast(cutGroup(222,2), stid ~ discid)
isB12sem2 <- dcast(cutGroup(278,2), stid ~ discid)

n_clusters <- 3
k1 <- kmeans(isB11sem2[,2:ncol(isB11sem2)],n_clusters)
k2 <- kmeans(isB12sem2[,2:ncol(isB12sem2)],n_clusters)

isB11sem2$cluster <- factor(k1$cluster)
isB12sem2$cluster <- factor(k2$cluster)

```

```
plot(hclust(dist(isB11sem2[, 2:ncol(isB11sem2)])), xlab="", ylab= "distance",
      sub="", main=NULL)
abline(h=60, col="red", lwd=2, lty=3)
```

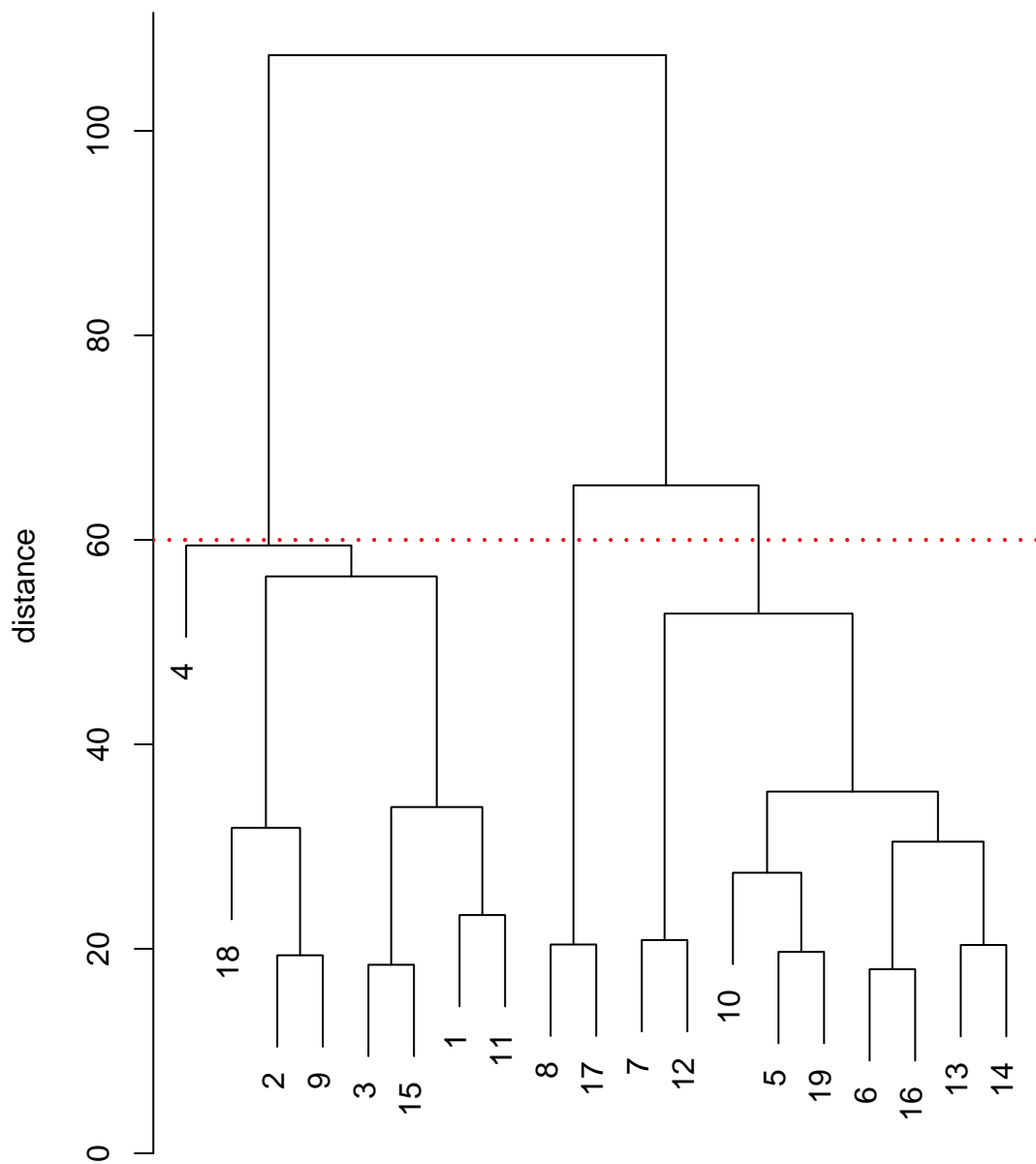


Рис. 1: Древоидная диаграмма кластеризации ИС-Б11

```

plot(hclust(dist(isB12sem2[, 2:ncol(isB11sem2)])), xlab="", ylab= "distance",
      sub="", main=NULL)
abline(h=45, col="red", lwd=2, lty=3)
abline(h=50, col="red", lwd=2, lty=3)

```

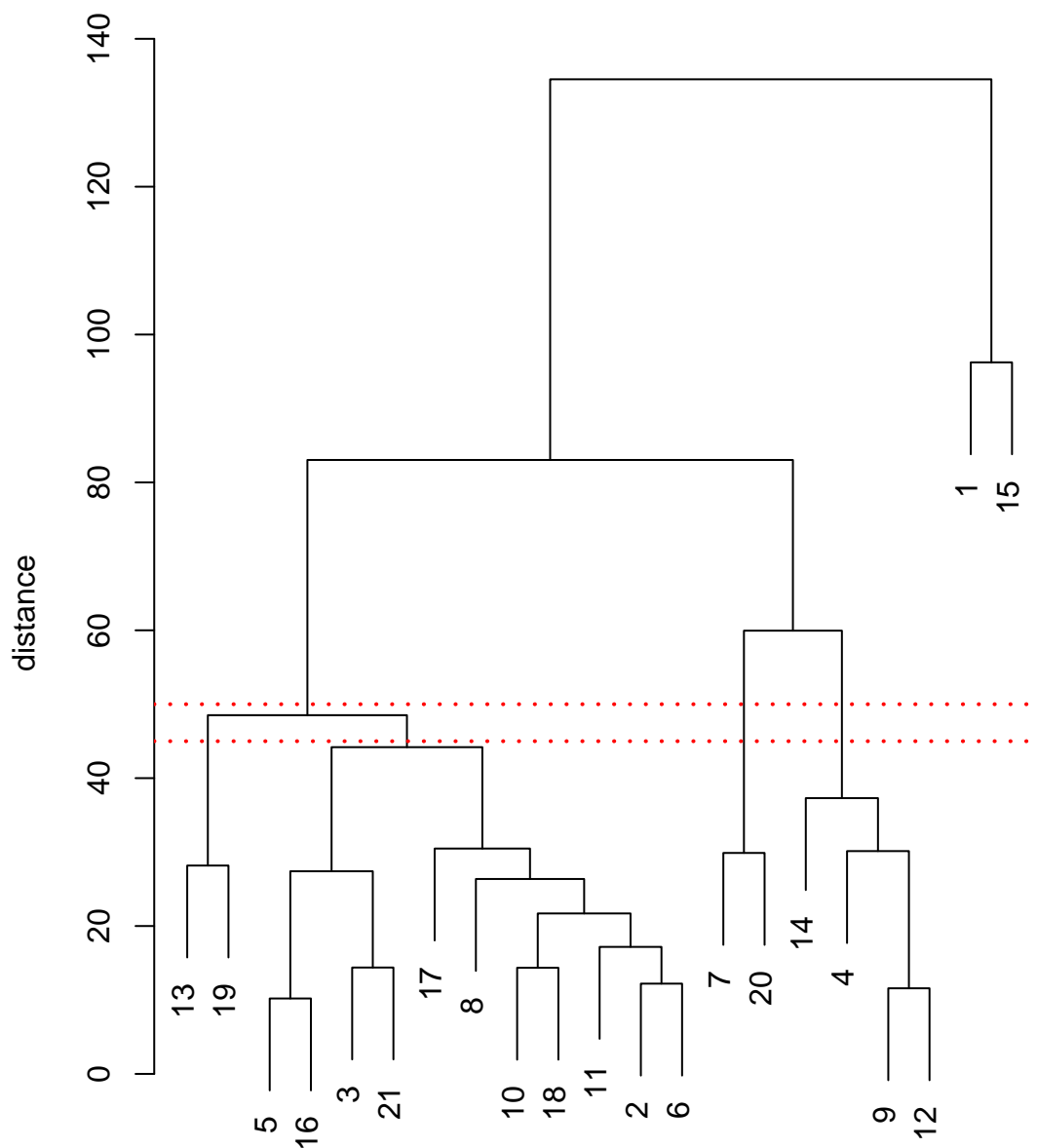


Рис. 2: Древоидная диаграмма кластеризации ИС-В12

3.2.2 Выбор дисциплин для визуализации k-means

На основании достаточно большого межквартильного размаха (IQR) были выбраны соответствующие дисциплины для визуализации результатов кластеризации по алгоритму kmeans. Таким образом, в качестве осей выбирались дисциплины с достаточно большим разбросом оценок. Дополнительно к IQR была вычислена медиана для более конкретного представления о разбросе оценок студентов.

```
discB11 <- B11 %>% group_by(discid) %>% summarise(n(), IQR(sum))
discB12 <- B12 %>% group_by(discid) %>% summarise(n(), IQR(sum))
studB11 <- B11 %>% group_by(stid) %>% summarise(n(), median(sum))
studB12 <- B12 %>% group_by(stid) %>% summarise(n(), median(sum))

discB11

## # A tibble: 10 x 3
##   discid `n()` `IQR(sum)`
##   <int> <int>     <dbl>
## 1     2    19      17.5
## 2     3    19       3.5
## 3     6    19      22.5
## 4    12    19      15.0
## 5    14    19      12.5
## 6    23    19      20.5
## 7   171    19       0.0
## 8   339    19       0.0
## 9   913    19      10.0
## 10  1114    19      17.5

discB12

## # A tibble: 10 x 3
##   discid `n()` `IQR(sum)`
##   <int> <int>     <dbl>
## 1     2    21       15
## 2     3    21        5
## 3     4    21       10
## 4     6    21       20
## 5    12    21       13
## 6    14    21       11
## 7    21    21       20
## 8    23    21       19
## 9   171    21        0
## 10  339    21        0
```

Идентификаторы выбранных дисциплин: 6 и 23 для ИС-Б11 и 6 и 21 для ИС-Б12.

3.2.3 Визуализация k-means

Имена дисциплин загружаются в подпись графика динамически через функцию `getDiscName(id)`

```
getDiscName <- function(id) {  
  name <- filter(adata, discid == id)  
  return(pull(name, discname)[1])  
}
```

```
plot(isB11sem2[,c("6")], isB11sem2[,c("23")], col=k1$cluster,  
     xlab=getDiscName(6), ylab=getDiscName(23))
```

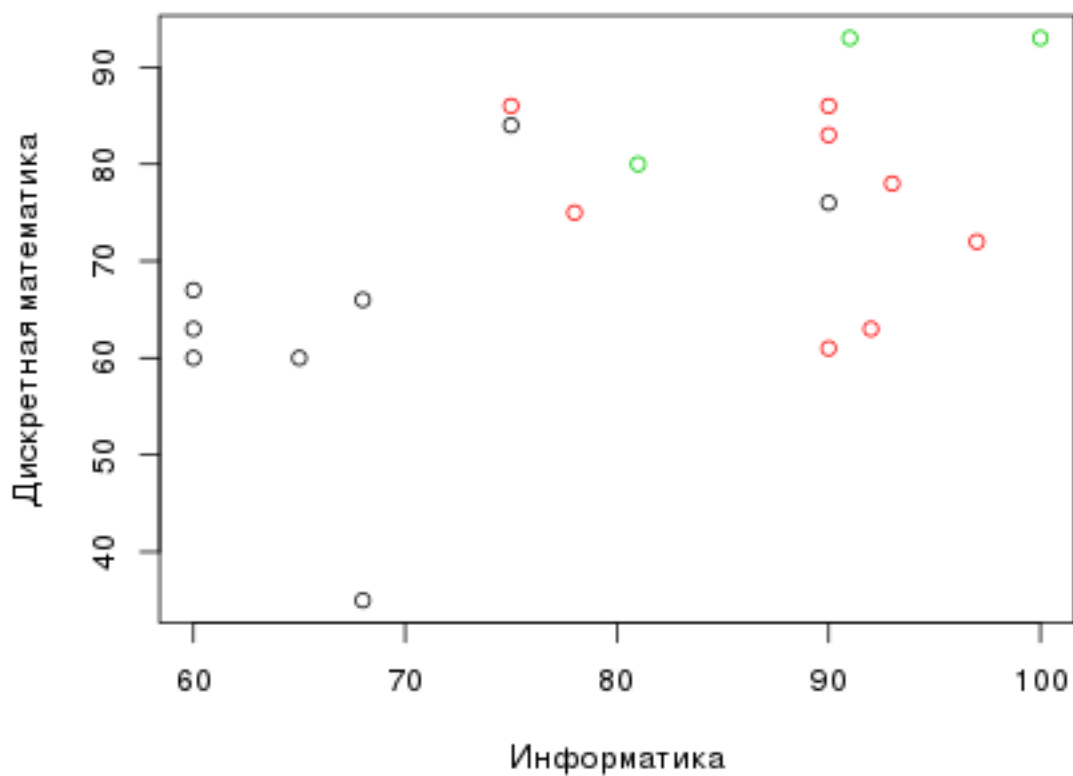


Рис. 3: Результат кластеризации kmeans для ИС-Б11


```
plot(isB12sem2[,c("6")], isB12sem2[,c("21")], col=k2$cluster,
     xlab=getDiscName(6), ylab=getDiscName(21))
```

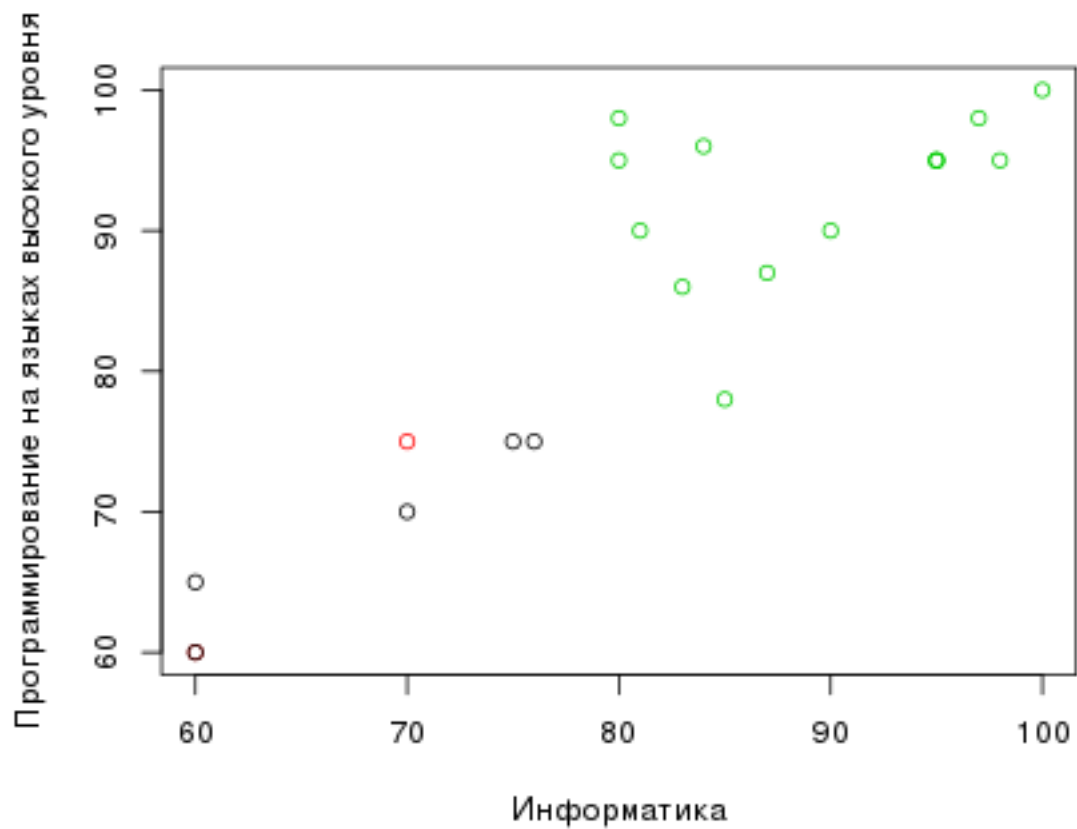


Рис. 4: Результат кластеризации kmeans для ИС-Б12

3.3 Статистический анализ