Министерство науки и высшего образования Российской Федерации

Новосибирский государственный технический университет

Кафедра Вычислительной техники



Лабораторная работа №3

По дисциплине «Методы анализа данных»

По теме

«Применение деревьев классификации в решении задач интеллектуального анализа данных (ИАД) средствами интегрированной системы Statistica и языка R».

Факультет: АВТ

Группа: АВТ-819

Студент: Ванин К.Е.

Преподаватель: Пешков А.В.

Новосибирск

2022

Оглавление

[Цель лабораторной работы 3](#_Toc89104138)

[Задание к лабораторной работе 3](#_Toc89104139)

[Постановка задачи ИАД. Вариант 8. 6](#_Toc89104140)

[Исходные данные 6](#_Toc89104141)

[Ход работы 8](#_Toc89104142)

[Приложение 20](#_Toc89104143)

# Цель лабораторной работы

* изучить алгоритмы и методы деревьев классификации на примере решения задачи ИАД;
* исследовать эффективность использования алгоритмов и методов деревьев классификации для решения прикладной задачи;
* ознакомиться и получить практические навыки работы с модулями интегрированной системы *Statistica* и среды *R,* реализующими методы деревьев классификации.

# Задание к лабораторной работе

Постановка задачи ИАД и исходные данные для выполнения лабораторной работы такие же, как в л.р. 2., к данным необходимо добавить столбец с номером кластера (класса), определенного одним из методов кластеризации.

Постройте правило отнесения объектов к одному из заданных классов, используя методы деревьев классификации.

1. Постройте деревья классификации, используя разные комбинации признаков (показателей), заданных в задаче:

*1 вариант построения* – все качественные и количественные признаки;

*2 вариант построения –* только качественные признаки.

При построении дерева классификации задайте следующие опции:

- одинаковые цены ошибки классификации и априорные вероятности попадания объекта в тот или иной класс, оцененные по исходным данным;

- метод полного перебора деревьев с одномерным ветвлением по методу CART;

- меру Джини однородности вершины в качестве критерия согласия;

- правило останова – прямая остановка по методу FACT, доля неклассифицированных объектов – 0,05;

- параметры V – кратной кросс-проверки: V=3.

Постройте графическое представление деревьев (в узлах дерева задайте гистограммы). Оцените сложность и точность построенных деревьев классификации, по результатам анализа заполните табл. 1.

*Таблица 1.*

*Показатели качества деревьев решений*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **номер вар. дерева** | **признаки** | **правило останова** | **точность дерева** | | | | **сложность дерева** | | | |
| **Resub.cost\*** | **CV-cost+s.d.** | **Global CV-cost +s.d.** | **Test-cost+s.d.** | **число листьев** | **число вершин** | **максимальн. длина пути** | **длина внешнего пути** |
|  |  |  |  |  |  |  |  |  |  |  |

**\*Resub.cost – цена обучения**

**CV-cost+s.d. – цена кросс-проверки + стандартная ошибка цены**

**Global CV-cost+s.d. – цена глобальной кросс-проверки + стандартная ошибка цены**

**Test-cost – цена на тестовой выборке + стандартная ошибка цены**

**\*\* Для вариантов дерева 1 и 2 из показателей точности дерева оцениваются Resub.cost и CV-cost+s.d.**

**Для варианта дерева 3 оцениваются показатели: Resub.cost, CV-cost+s.d., Global CV-cost+s.d. Для варианта дерева 4 оцениваются все показатели точности.**

**Для варианта дерева 5 оцениваются показатели точности: Resub.cost, Test-cost.**

**Для варианта дерева 6 оцениваются показатели точности: Resub.cost, CV-cost.**

2. Постройте дерево классификации на основе использования качественных и количественных признаков. Укажите правило останова «отсечение по ошибке классификации»:

- задайте следующие параметры (*вариант* *построения дерева 3*): максимальное количество неверно классифицированных объектов – 1; стандартная ошибка правила – 0,1.

Остальные опции такие же, как в пункте 1. Полученные результаты внесите в табл. 1.

1. Для варианта 3. дерева классификации постройте:

- графической представление дерева (в узлах дерева задайте гистограммы);

- матрицу ошибок классификации;

- последовательность оптимально усеченных деревьев и их параметры качества;

- график цены обученияицены кросс проверки для цепочки последовательно усекаемых деревьев;

- выполните глобальную кросс-проверку (используйте параметры глобальной кросс проверки - по умолчанию), постройте матрицу ошибок классификации по результатам глобальной кросс проверки;

- сделайте вывод об эффективности работы процедуры «автоматического» выбора оптимального варианта дерева.

1. Постройте дерево классификации (*вариант построения дерева 4*) на основе использования качественных и количественных признаков с выделением тестовой выборки (в качестве тестовой выборки используйте 10 последних наблюдений). Все настройки задайте, как в пункте 1. Постройте графическое представление дерева (в узлах дерева задайте гистограммы). Оцените сложность и точность построенного дерева классификации, по результатам анализа заполните табл. 1.
2. Выполните построение дерева классификации на основе количественных и качественных признаков в среде R:

- с использованием тестовой выборки (*вариант построения дерева 5*);

- методом перекрестной кросс-валидации (*вариант построения дерева 6*).

Оцените сложность и точность построенного дерева классификации, по результатам анализа заполните табл. 1.

1. Выберите наилучший по совокупности показателей качества вариант построения дерева классификации.
2. По результатам проведенного исследования ответьте на следующие вопросы.

- На основе каких признаков (качественных, количественных) построено дерево классификации, обладающее лучшими показателями точности и сложности?

- Как повлияло изменение правила останова и параметров правила останова на результаты классификации объектов, на точность и сложность, построенного дерева классификации?

- Как повлияло изменение метода кросс-проверки на точность и сложность построенного дерева классификации?

1. Сделайте выводы по результатам проведенного исследования в произвольной форме.

# **Постановка задачи ИАД. Вариант 8.**

**Задача 8. (вариант 2, 5, 8, 11, 14)** Исследуются промышленныепредприятия г. Новосибирска. Необходимо разделить предприятия на группы (классы) в соответствии с рядом устойчивых показателей эффективности работы предприятия. Разделение предприятий на группы проводится с целью установления нормативных показателей для каждой выделенной группы. Также требуется построить правило отнесения предприятия к одной из выделенных групп (классов).

Исходные данные для проведения статистического анализа представлены в табл. 1. В статистическое исследование включены данные по 40 промышленным предприятиям г. Новосибирска. Зарегистрированы следующие показатели эффективности работы предприятия за 2018 г.

* объем выработки, в млн. руб.;
* фонд заработной платы, в млн. руб.;
* интегральный показатель использования рабочего времени;
* прибыль, в млн. руб.;
* уровень автоматизации производства (1 – ниже среднего; 2 – средний; 3- выше среднего);
* текучесть кадров (1 – ниже среднего уровня; 2 – средний уровень; 3 – выше среднего).

**вариант 8.**

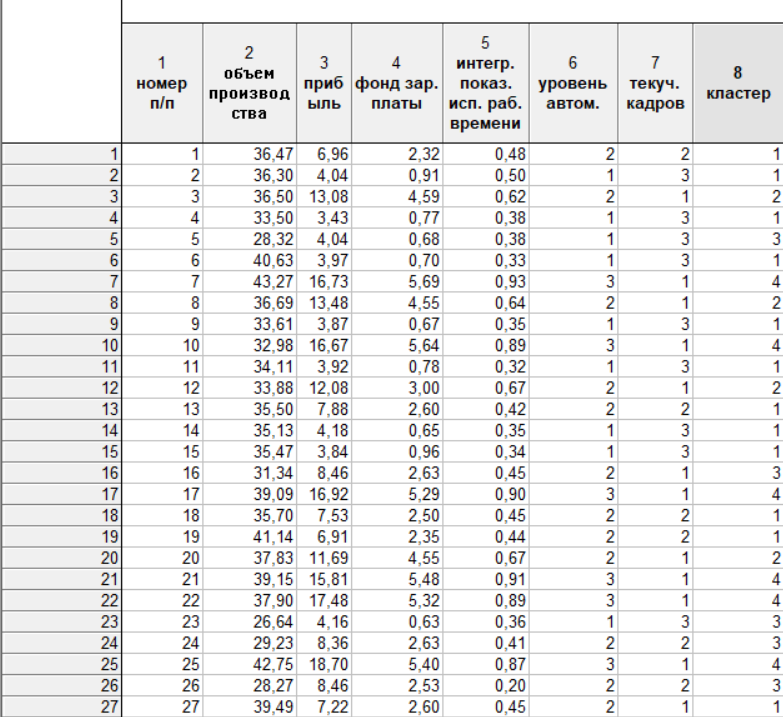
*Таблица 1.*

# Исходные данные

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **номер п/п** | **объем производства** | **прибыль** | **фонд зар. платы** | **интегр. показ.  исп. раб. времени** | **уровень  автом.** | **текуч. кадров** |
| 1 | 36,47 | 6,96 | 2,32 | 0,48 | 2 | 2 |
| 2 | 36,30 | 4,04 | 0,91 | 0,50 | 1 | 3 |
| 3 | 36,50 | 13,08 | 4,59 | 0,62 | 2 | 1 |
| 4 | 33,50 | 3,43 | 0,77 | 0,38 | 1 | 3 |
| 5 | 28,32 | 4,04 | 0,68 | 0,38 | 1 | 3 |
| 6 | 40,63 | 3,97 | 0,70 | 0,33 | 1 | 3 |
| 7 | 43,27 | 16,73 | 5,69 | 0,93 | 3 | 1 |
| 8 | 36,69 | 13,48 | 4,55 | 0,64 | 2 | 1 |
| 9 | 33,61 | 3,87 | 0,67 | 0,35 | 1 | 3 |
| 10 | 32,98 | 16,67 | 5,64 | 0,89 | 3 | 1 |
| 11 | 34,11 | 3,92 | 0,78 | 0,32 | 1 | 3 |
| 12 | 33,88 | 12,08 | 3,00 | 0,67 | 2 | 1 |
| 13 | 35,50 | 7,88 | 2,60 | 0,42 | 2 | 2 |
| 14 | 35,13 | 4,18 | 0,65 | 0,35 | 1 | 3 |
| 15 | 35,47 | 3,84 | 0,96 | 0,34 | 1 | 3 |
| 16 | 31,34 | 8,46 | 2,63 | 0,45 | 2 | 1 |
| 17 | 39,09 | 16,92 | 5,29 | 0,90 | 3 | 1 |
| 18 | 35,70 | 7,53 | 2,50 | 0,45 | 2 | 2 |
| 19 | 41,14 | 6,91 | 2,35 | 0,44 | 2 | 2 |
| 20 | 37,83 | 11,69 | 4,55 | 0,67 | 2 | 1 |
| 21 | 39,15 | 15,81 | 5,48 | 0,91 | 3 | 1 |
| 22 | 37,90 | 17,48 | 5,32 | 0,89 | 3 | 1 |
| 23 | 26,64 | 4,16 | 0,63 | 0,36 | 1 | 3 |
| 24 | 29,23 | 8,36 | 2,63 | 0,41 | 2 | 2 |
| 25 | 42,75 | 18,70 | 5,40 | 0,87 | 3 | 1 |
| 26 | 28,27 | 8,46 | 2,53 | 0,20 | 2 | 2 |
| 27 | 39,49 | 7,22 | 2,60 | 0,45 | 2 | 1 |
| 28 | 37,36 | 12,15 | 4,47 | 0,65 | 2 | 1 |
| 29 | 43,24 | 18,16 | 5,58 | 0,90 | 3 | 1 |
| 30 | 37,23 | 12,16 | 4,49 | 0,65 | 2 | 1 |
| 31 | 35,12 | 11,89 | 4,69 | 0,66 | 2 | 1 |
| 32 | 32,15 | 4,77 | 0,73 | 0,37 | 1 | 3 |
| 33 | 43,30 | 11,98 | 4,59 | 0,64 | 2 | 2 |
| 34 | 35,16 | 8,09 | 2,53 | 0,46 | 2 | 2 |
| 35 | 37,04 | 17,87 | 5,29 | 0,91 | 3 | 1 |
| 36 | 32,67 | 7,40 | 2,18 | 0,45 | 2 | 2 |
| 37 | 45,92 | 12,00 | 5,00 | 0,70 | 3 | 1 |
| 38 | 43,28 | 18,35 | 5,68 | 0,91 | 3 | 1 |
| 39 | 33,96 | 11,87 | 4,37 | 0,69 | 2 | 2 |
| 40 | 38,50 | 12,55 | 4,48 | 0,65 | 2 | 1 |
| 41 | 42,36 | 16,18 | 5,67 | 0,60 | 3 | 1 |
| 42 | 39,87 | 19,76 | 5,44 | 0,89 | 3 | 1 |
| 43 | 35,61 | 14,28 | 4,53 | 0,64 | 2 | 2 |
| 44 | 42,56 | 12,26 | 4,62 | 0,65 | 2 | 1 |
| 45 | 35,20 | 12,97 | 4,20 | 0,65 | 2 | 2 |
| 46 | 32,75 | 6,57 | 0,80 | 0,46 | 2 | 1 |
| 47 | 32,24 | 5,66 | 2,93 | 0,46 | 2 | 2 |
| 48 | 35,22 | 6,91 | 2,49 | 0,44 | 2 | 2 |
| 49 | 34,70 | 10,00 | 0,77 | 0,38 | 1 | 3 |
| 50 | 29,42 | 4,57 | 0,85 | 0,35 | 1 | 3 |

# Ход работы

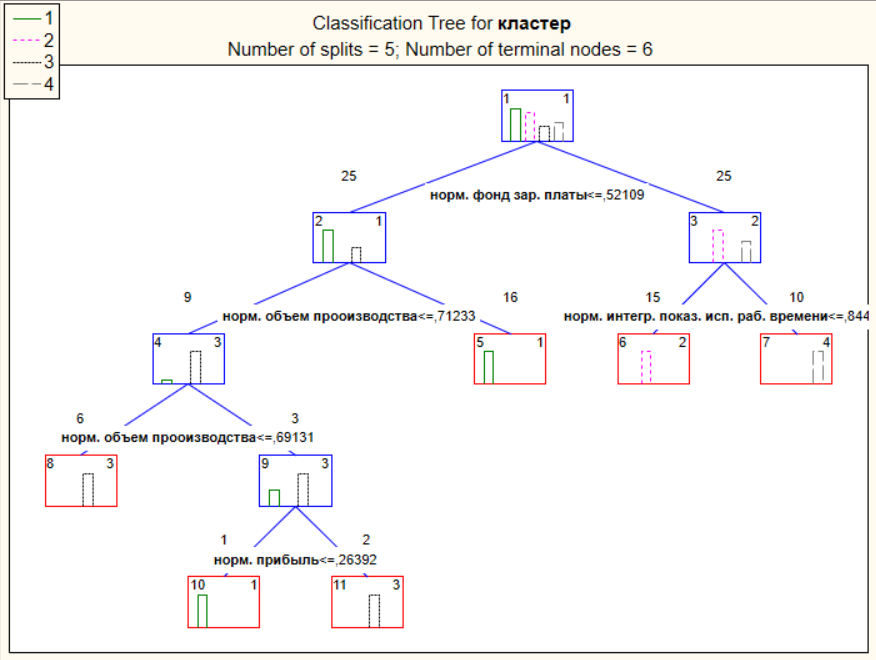
Загруженные данные с добавленным столбцом кластеров:

**

*Рисунок 1 – Данные варианта*

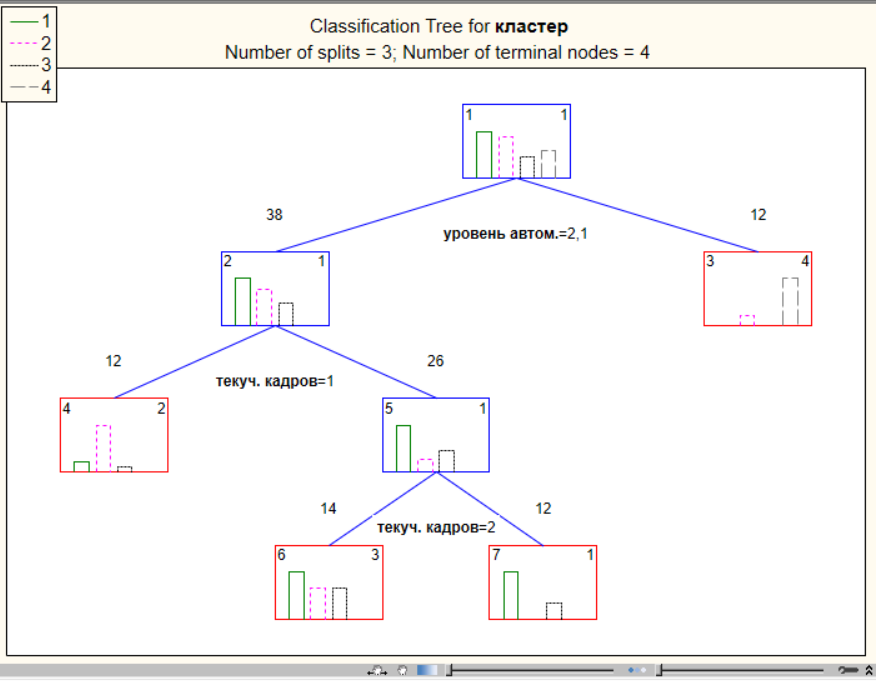
1. Построили деревья классификации, используя разные комбинации признаков (показателей), заданных в задаче:

*1 вариант построения* – все качественные и количественные признаки:

**

*Рисунок 2 - Все качественные и количественные признаки*

*2 вариант построения –* только качественные признаки:

**

*Рисунок 3 - Только качественные признаки*

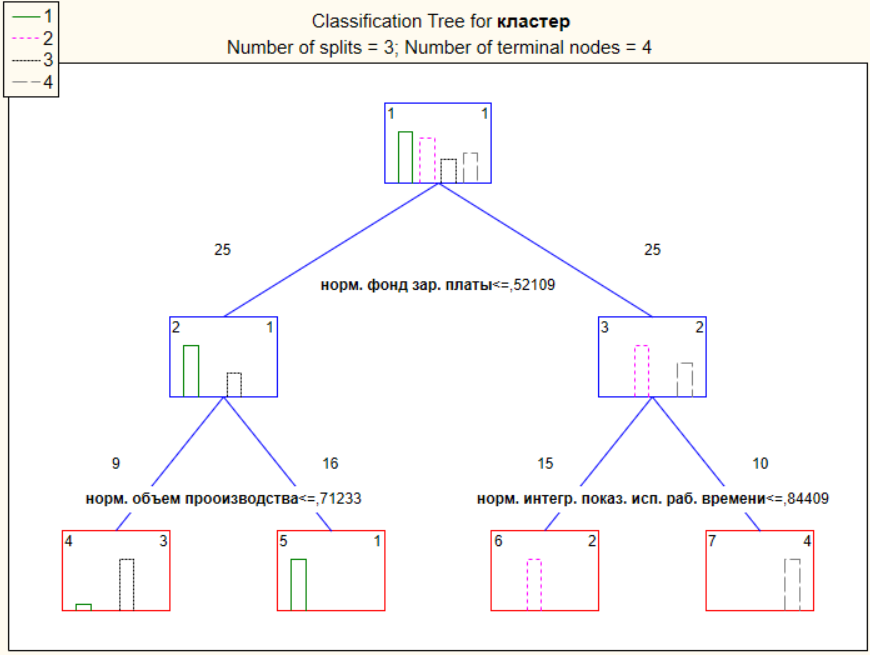
2. Построили дерево классификации на основе использования качественных и количественных признаков. Указали правило останова «отсечение по ошибке классификации»:

Задали следующие параметры (*вариант* *построения дерева 3*): максимальное количество неверно классифицированных объектов – 1; стандартная ошибка правила – 0,1.

Остальные опции такие же, как в пункте 1.

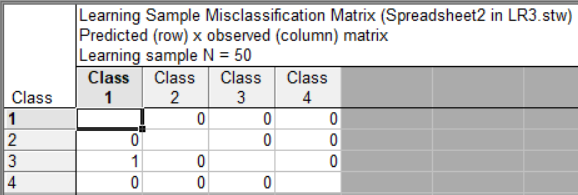
1. Для варианта 3. дерева классификации построили:

- графической представление дерева:



*Рисунок 4 - Качественные и количественные признаки (3 вар.)*

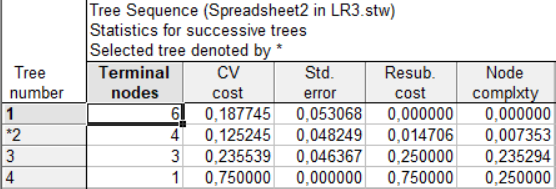
- матрицу ошибок классификации:

**

*Рисунок 5 - Матрица ошибок классификации*

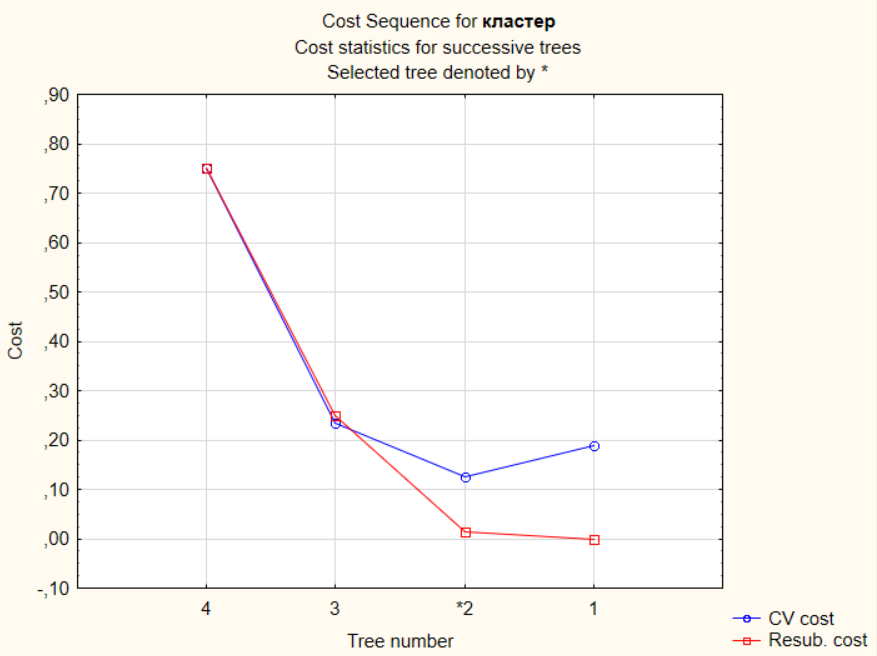
Из полученных данных видим, что один случай классифицирован неверно.

- последовательность оптимально усеченных деревьев и их параметры качества:

**

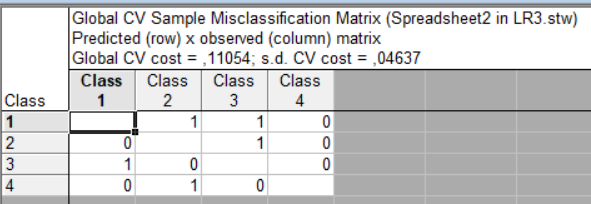
*Рисунок 6 – Последовательность оптимально усеченных деревьев и их параметры качества*

- график цены обученияицены кросс проверки для цепочки последовательно усекаемых деревьев:

**

*Рисунок 7 - График цены обучения и цены кросс проверки для цепочки последовательно усекаемых деревьев*

- выполнили глобальную кросс-проверку, построили матрицу ошибок классификации по результатам глобальной кросс проверки:

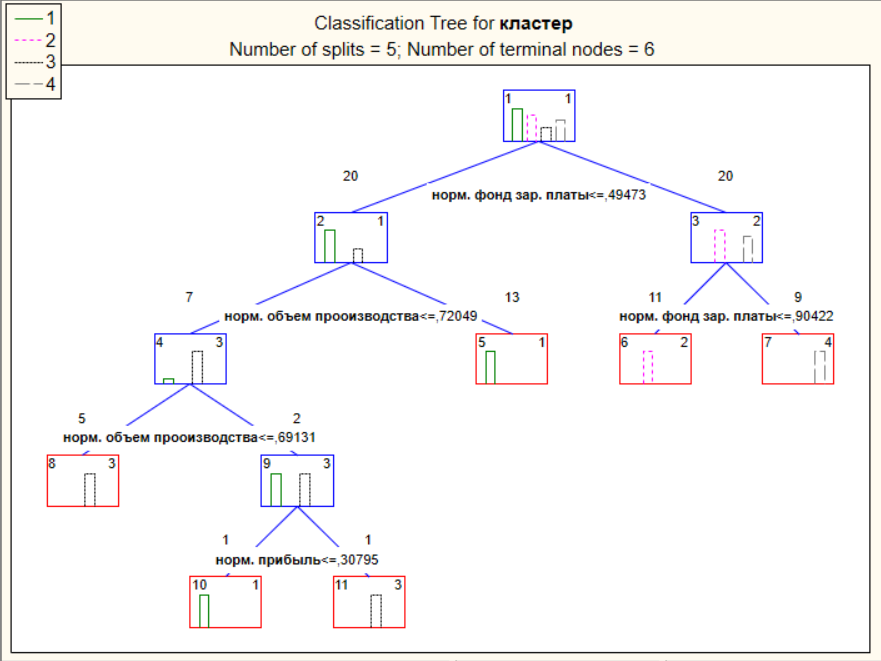
**

*Рисунок 8 - Матрица ошибок классификации по результатам глобальной кросс-проверки*

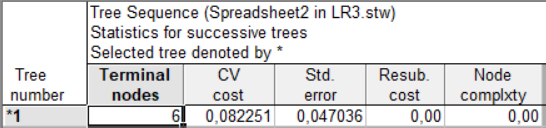
Вывод об эффективности работы:

Процедура «автоматического» выбора оптимального варианта дерева нужна для того, чтобы выбирать наименьшее по размеру дерево с близкой к минимальной цене кросс-проверки. Благодаря этому можно избегать потери точности нашего прогноза. У нас получилось 4 итерации. На четвертой итерации цена обучения приняла значение 0, цена кросс-проверки приняла значение близкое к 0,187.

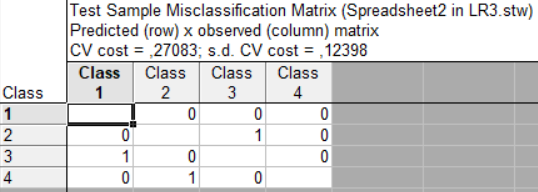
1. Построили дерево классификации (*вариант построения дерева 4*) на основе использования качественных и количественных признаков с выделением тестовой выборки (в качестве тестовой выборки были использованы 10 последних наблюдений). Все настройки заданы, как в пункте 1.



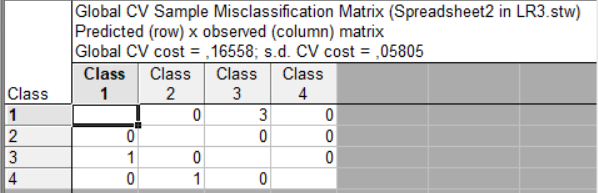
*Рисунок 9 – Дерево классификации, построенное при выделении тестовой выборки (вариант 4)*



*Рисунок 10 – Последовательность оптимально усеченных деревьев и их параметры качества*



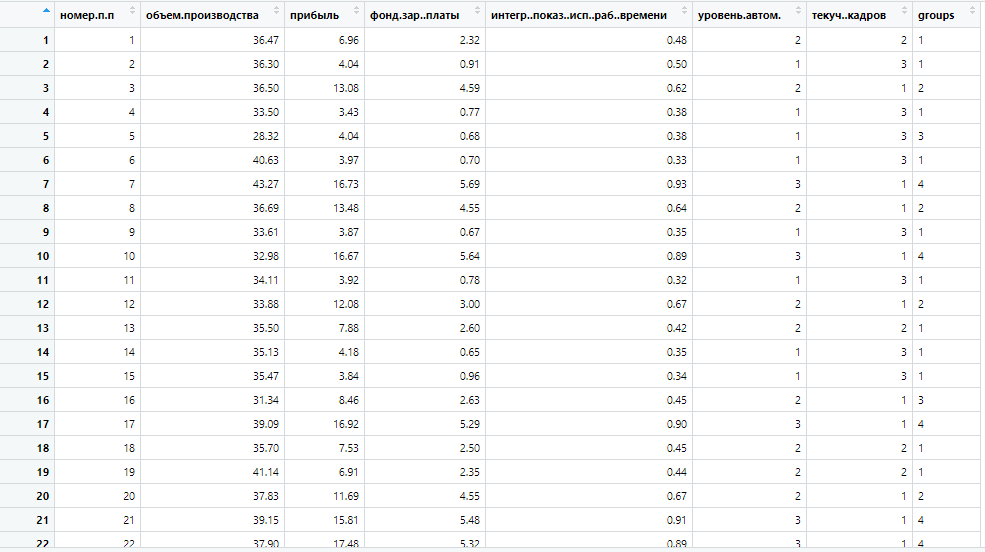
*Рисунок 11 – Результаты кросс проверки на тестовой выборке*



*Рисунок 12 - Матрица ошибок классификации по результатам глобальной кросс-проверки*

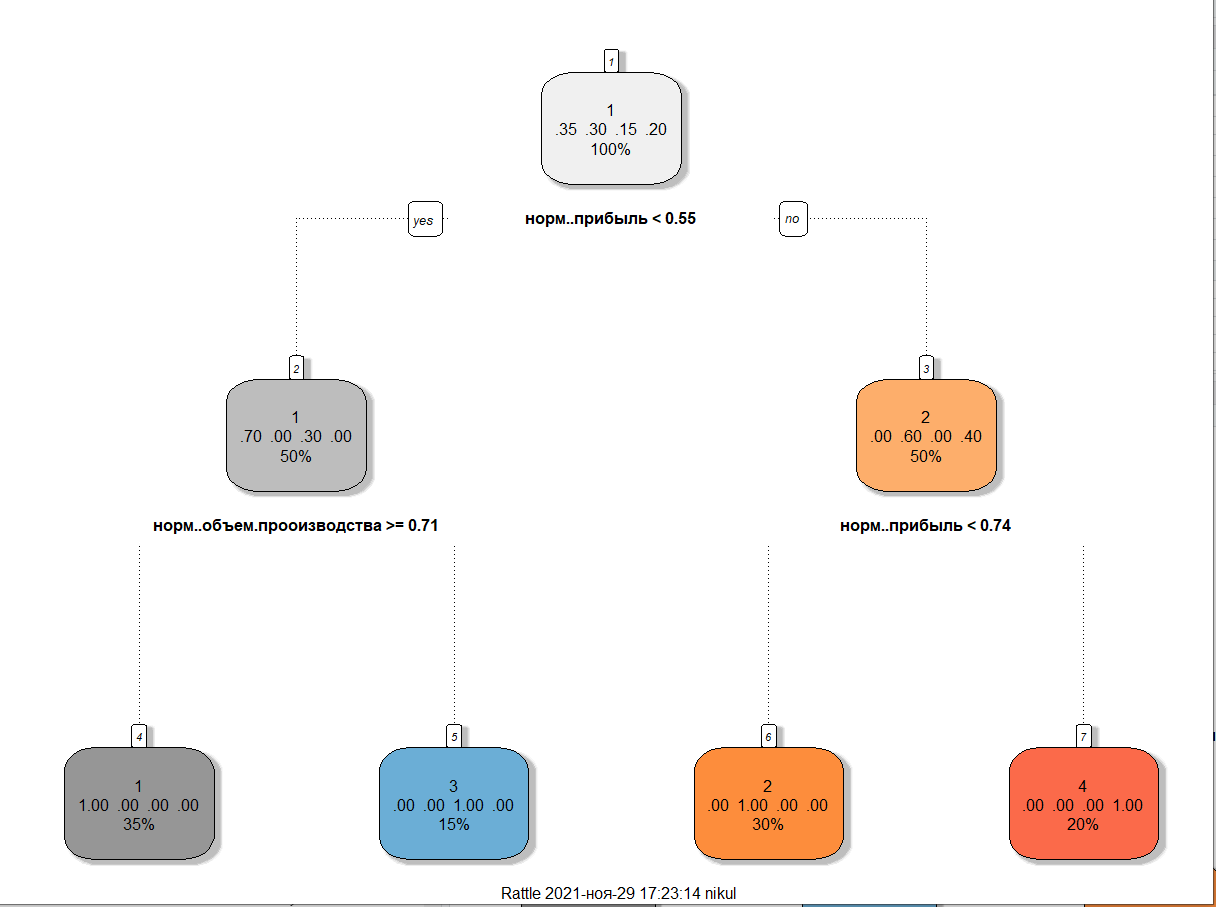
По итогам проведенного анализа, мы можем увидеть, что в каждом листе, полученного дерева (рисунок 9), находится объект только одного кластера. Таким образом, можно сказать, что дерево на обучающей выборке классифицирует данные с достаточно большой точностью.

1. Выполнили построение дерева классификации на основе количественных и качественных признаков в среде R:

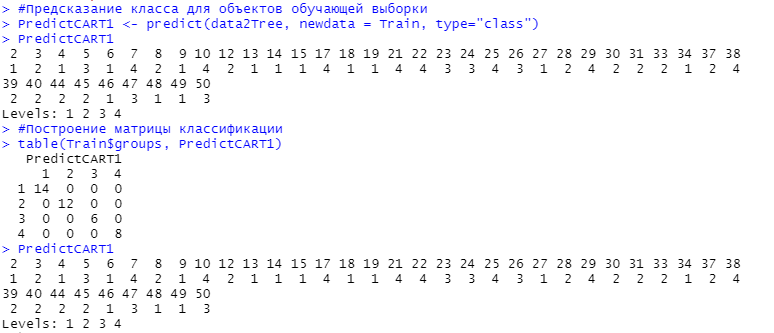


*Рисунок 13 – Загруженные данные варианта*

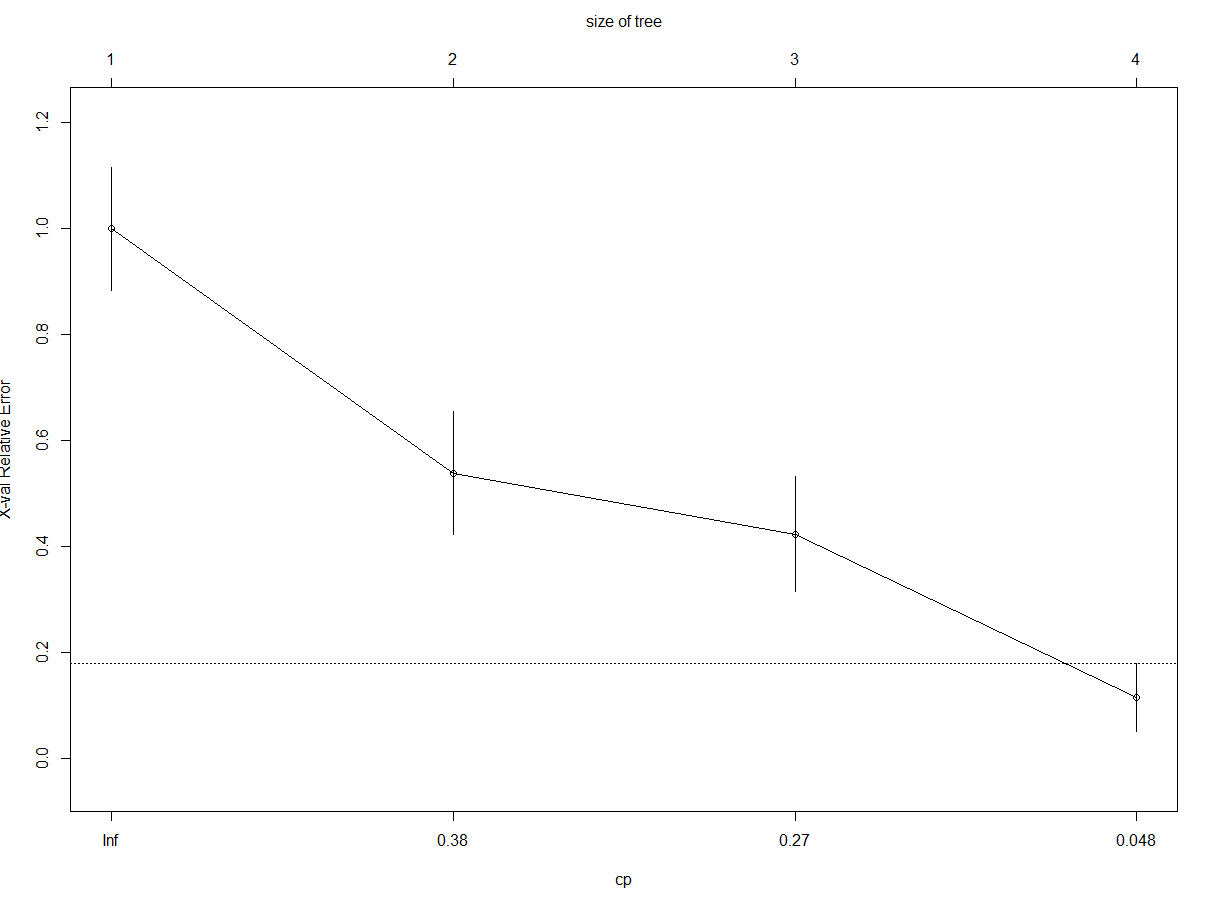
- с использованием тестовой выборки (*вариант построения дерева 5*);



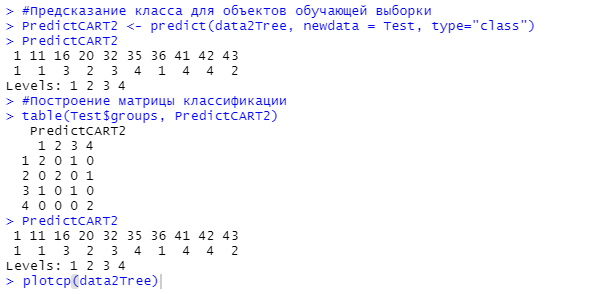
*Рисунок 14 – Дерево классификации на основе количественных и качественных признаков с выделением тестовой выборки (вариант 5)*



*Рисунок 15 – Предсказание класса для объектов обучающей выборки*



*Рисунок 16 – График изменения относительной точности дерева в зависимости от сложности модели и размера дерева*

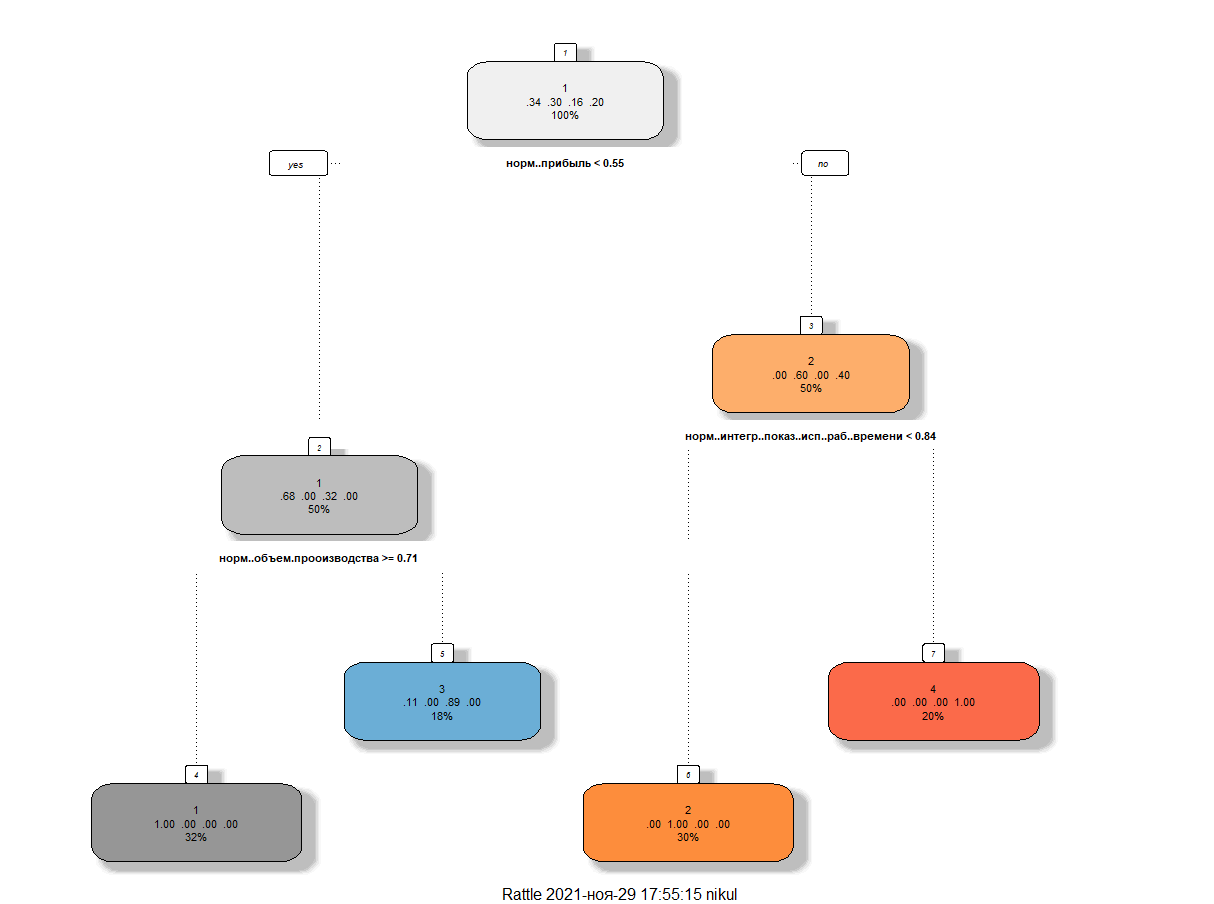


*Рисунок 17 – Предсказание класса для объектов обучающей выборки*

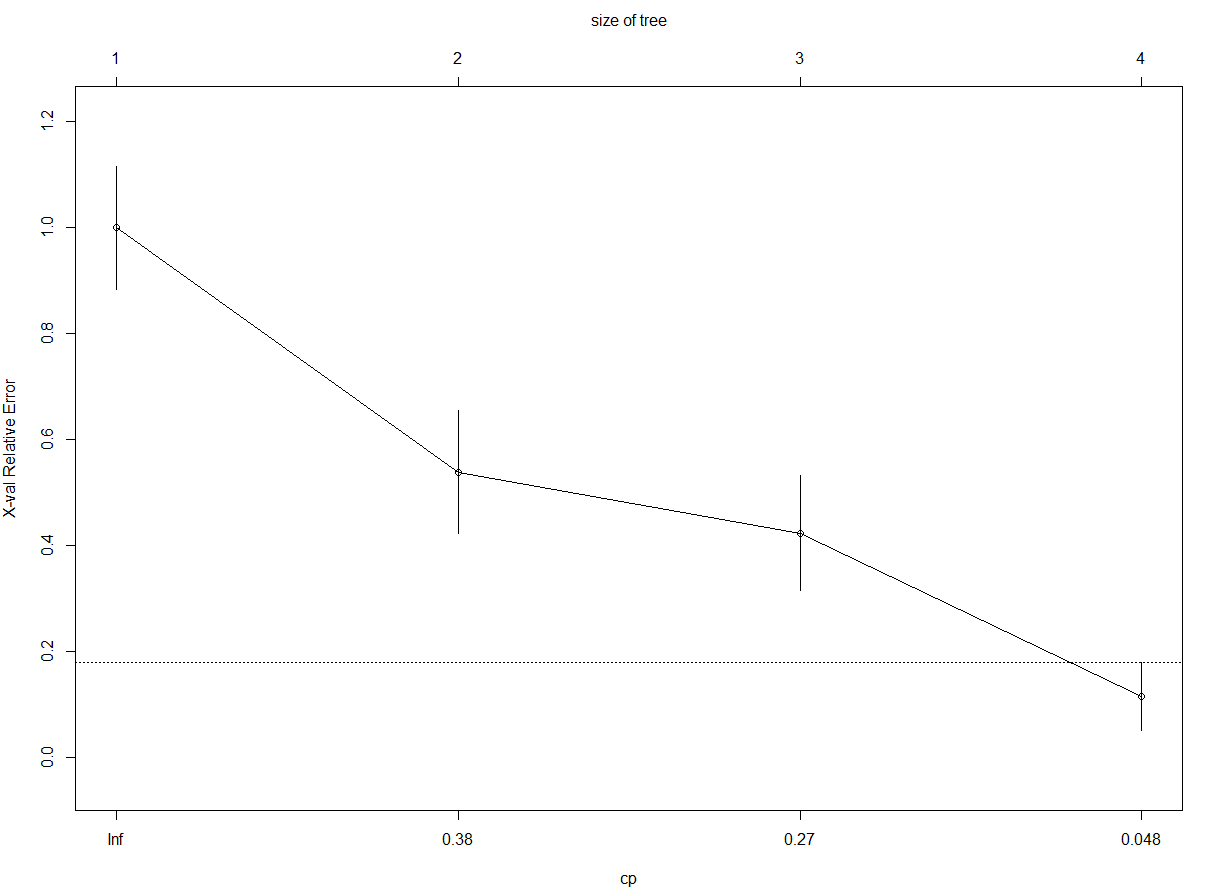
- методом перекрестной кросс-валидации (*вариант построения дерева 6*)



*Рисунок 18 – Вычисленное значение цены кросс-валидации при разных значениях параметра cp*

**

*Рисунок 19 – Дерево классификации с использованием кросс-валидации (вариант 6)*

**

*Рисунок 20 – График изменения относительной точности дерева в зависимости от сложности модели и размера дерева*

Анализ вариантов построения дерева классификации представлен в таблице 1.

*Таблица 1*- *Показатели качества деревьев решений*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **номер вар. дерева** | **признаки** | **правило останова** | **точность дерева** | | | | **сложность дерева** | | | |
| **Resub.cost\*** | **CV-cost+s.d.** | **Global CV-cost +s.d.** | **Test-cost+s.d.** | **число листьев** | **число вершин** | **максимальн. длина пути** | **длина внешнего пути** |
| **1** | все качественные и количественные признаки | прямая остановка по методу FACT | 0 | 0,187745098 | - | - | 6 | 11 | 4 | 12 |
| **2** | только качественные признаки | прямая остановка по методу FACT | 0,342647 | 0,375735294 | - | - | 4 | 7 | 3 | 8 |
| **3** | все качественные и количественные признаки | отсечение по ошибке классификации | 0 | 0,651445 | 0,15691 | - | 4 | 7 | 2 | 8 |
| **4** | все качественные и количественные признаки | с выделением тестовой выборки | 0 | 0,140301 | 0,22363 | 0,39481 | 6 | 11 | 4 | 12 |
| **5** | все качественные и количественные признаки | с выделением тестовой выборки | 0 | - | - | 0,18 | 4 | 7 | 2 | 8 |
| **6** | все качественные и количественные признаки | с выделением тестовой выборки | 0 | 0,18 | - | - | 4 | 7 | 2 | 8 |

1. Выберите наилучший по совокупности показателей качества вариант построения дерева классификации.

Наилучшие по совокупности показателей качества показали 4 и 6, это мы можем увидеть из таблицы 1. А наихудшие результаты показали варианты 2, в котором только качественные признаки, и 3.

1. По результатам проведенного исследования ответьте на следующие вопросы.

- На основе каких признаков (качественных, количественных) построено дерево классификации, обладающее лучшими показателями точности и сложности?

На основе всех качественных и количественных признаков

- Как повлияло изменение правила останова и параметров правила останова на результаты классификации объектов, на точность и сложность, построенного дерева классификации?

Сравним варианты 1 и 3, так как построение велось по одинаковым признакам, а также расчет происходил в одной программной среде Statistica. С изменением правила останова по методу FACT, на отсечение по ошибке классификации, для нашего варианта цена кросс-проверки увеличилась более чем в 3 раза, а дерево классификации уменьшилось.

При изменении правила останова на выделение тестовой выборки данный параметр уменьшился, в сравнении с обоими предшествующими вариантами.

- Как повлияло изменение метода кросс-проверки на точность и сложность построенного дерева классификации?

V-кратная кросс-проверка значительно отличается от глобальной кросс-проверка. Также можем отметить, что глобальной кросс-проверка показала наихудшую цену кросс-проверки.

1. Сделайте выводы по результатам проведенного исследования в произвольной форме.

В ходе выполнения лабораторной работы были изучены алгоритмы и методы деревьев классификации на примере решения задачи ИАД, исследованы эффективность использования алгоритмов и методов деревьев классификации для решения прикладной задачи, проведено ознакомление и получены практические навыки работы с модулями интегрированной системы Statistica и среды R, реализующими методы деревьев классификации.

Нами были опробованы все заданные варианты построения деревьев классификации. Отсюда можем сделать следующие выводы:

* Вариант 3 дал наихудшие результаты если сравнивать с остальными вариантами.
* Вариант 2 дал результаты лучше, чем 3, однако по сравнению с остальными тоже не очень хорошие.
* Наилучшими по совокупности показателей качества можем назвать 4 и 6. При этом вариант 4 стал наилучшим с показателем **CV-cost+s.d.=** 0,140301
* На основе всех качественных и количественных признаков построено дерево классификации, обладающее лучшими показателями точности и сложности.
* Изменение метода кросс-проверки ухудшило результаты точности построения дерева.

# Приложение

**Листинг**

#загрузка и просмотр данных варианта

data <- read.table("("C:\\Users\\рбт\\Desktop\\учеба\\7 семестр\\мад\\LR3.csv"", header=TRUE, sep=",")

View(data)

#удаление пропущенных значений

data1 <- na.omit(data)

#cтандартизация переменных(к минимуму 0 и максимуму 1)

maxs <- apply(data[,2:5], 2, max)

mins <- apply(data[,2:5], 2, min)

data1 <- scale(data[,2:5], center = mins, scale = maxs - mins)

View(data1)

#процедура кластерного анализа

dist.prog <- dist(data1[,1:4])

clust.prog <- hclust(dist.prog, method = "complete")

plot(clust.prog, hang = -1)

rect.hclust(clust.prog, k=4, border="red")

#Данные разделены на 4 группы (кластера), в векторе groups хранятся номера кластеров

groups <- cutree(clust.prog, k=4)

#Номер кластера внесен в таблицу с исходными данными

data2 <- cbind(data, groups)

#Удалены пустые столбцы из исходной таблицы, вектор groups объявлен как факторный

data2<-data2[,-c(12)]

View(data2)

str(data2)

data2$groups<-as.factor(data2$groups)

str(data2)

#Номер кластера (класса) хранится в переменной groups, которая объявлена как фактор:

data2$groups<-as.factor(data2$groups)

#Установлен пакет caTools для разделения данных на обучающую и тестовую выборки

install.packages("caTools")

library(caTools)

#Данные разделены на две выборки: обучающую и тестовую (20% - тестовая выборка и 80% - обучающая выборка).

split <- sample.split(data2$groups, SplitRatio = 0.8)

Train <- subset(data2, split == TRUE)

Test <- subset(data2, split == FALSE)

#Установлены пакеты для работы с деревьями решений (построение дерева и визуализация дерева)

install.packages("rpart")

install.packages("rpart.plot")

install.packages("rattle")

install.packages("RColorBrewer")

library(rpart)

library(rpart.plot)

library(rattle)

library(RColorBrewer)

data2Tree <- rpart(groups ~ уровень.автом. + текуч..кадров + норм..объем.прооизводства + норм..прибыль + норм..фонд.зар..платы + норм..интегр..показ..исп..раб..времени,

data = Train, method = "class", control=rpart.control(minbucket = 1))

#Вывод структуры дерева в текстовом виде

data2Tree

#Визуализация дерева

fancyRpartPlot(data2Tree, palettes=c("Greys", "Oranges","Blues", "Reds"))

#Предсказание класса для объектов обучающей выборки

PredictCART1 <- predict(data2Tree, newdata = Train, type="class")

PredictCART1

#Построение матрицы классификации

table(Train$groups, PredictCART1)

PredictCART1

plotcp(data2Tree)

#тест

#Предсказание класса для объектов обучающей выборки

PredictCART2 <- predict(data2Tree, newdata = Test, type="class")

PredictCART2

#Построение матрицы классификации

table(Test$groups, PredictCART2)

PredictCART2

plotcp(data2Tree)

#Построение дерева с помощью кросс-валидации

#Установка пакетов

install.packages("caret")

install.packages("e1071")

library(caret)

library(e1071)

#Использование метода кросс-валидации: method="cv" , выборка делится на 30 частей

number=30

fitControl <- trainControl(method="cv", number=30)

#Использование значения параметра сложности cp от 0.001 до 0.2

cartGrid <- expand.grid(.cp=(1:200)\*0.001)

train(groups ~ уровень.автом. + текуч..кадров + норм..объем.прооизводства + норм..прибыль + норм..фонд.зар..платы + норм..интегр..показ..исп..раб..времени, data=data2, method="rpart",

trControl=fitControl, tuneGrid=cartGrid)

CART

#Необходимо задать найденное значение cp при построении дерева

dataTree3 <- rpart(groups ~ уровень.автом. + текуч..кадров + норм..объем.прооизводства + норм..прибыль + норм..фонд.зар..платы + норм..интегр..показ..исп..раб..времени,

data = data2, method = "class",

control=rpart.control(cp = 0.187))

#Визуализировать дерево

fancyRpartPlot(dataTree3, palettes=c("Greys", "Oranges","Blues", "Reds"))

#Вычислить точность

PredictCART3 <- predict(dataTree3, newdata = data2, type="class")

table(data2$groups, PredictCART3)

plotcp(dataTree3)