DOI 10.37882/2223–2966.2022.09.14

КЛАССИФИКАЦИЯ ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА R

CLASSIFICATION OF MEDICAL RESEARCH DATA WITH R

S. Kasyuk
G. Didenko
O. Stepanova

Summary. The article considers modern technologies of data classification in medical research. Support vector machines, classification of numerical and categorical data in neural networks, and classification trees are described. Appropriate functions of R language are considered. Examples of data classification for breast cancer data sets are given.

Keywords: medical research, data classification, support vector machine, neural networks, classification trees, R language.

Касюк Сергей Тимурович
К.т.н., доцент, ФГБОУ ВО «Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации (г. Челябинск)
sergey.kasyk@gmail.com
Диденко Галина Александровна
К.п.н., доцент, ФГБОУ ВО «Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации (г. Челябинск)
pga80@mail.ru
Степанова Оксана Александровна
К.п.н., доцент, ФГБОУ ВО «Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации (г. Челябинск)
okalst@mail.ru

Аннотация. В статье рассматриваются современные технологии классификации данных медицинских исследований. Описываются метод опорных векторов, классификация числовых и категориальных данных в нейронных сетях и деревья решений. Даются соответствующие функции языка R. Приводятся примеры классификации пациенток с раком молочной железы, решенные средствами языка R.

Ключевые слова: медицинские исследования, классификация данных, метод опорных векторов, нейронные сети, деревья решений, язык R.

1. Введение

Данная статья продолжает публикацию [1], посвященную анализу данных пациенток с раком молочной железы, размещенных в онлайн репозитории машинного обучения The UCI Machine Learning Repository.

Цель статьи

Дать обзор актуальных методов классификации данных с использованием статистического языка программирования R.

Классификация данных медицинских исследований является важной задачей диагностики, в ходе которой по результатам измерения различных параметров пациента принимается решение о необходимом лечении.

В статье приводятся такие технологии классификации, относящиеся к машинному обучению, как метод опорных векторов, нейронные сети и деревья решений с алгоритмом CART.

Для всех приведенных примеров данные были разбиты на *обучающую* (*train*) и *тестовую* (*test*) *выборки* в пропорциях 80% и 20% соответственно. Классификационные модели строились на *обучающих выборках*, а кросс-проверка — на *тестовых*. При обработке на языке R данные предварительно очищались от пропусков с помощью функции *na.omit*.

2. Классификация данных с использованием метода опорных векторов

Метод опорных векторов (*Support Vector Machine*, *SVM*), как технология машинного обучения,

использует набор контролируемых методов обучения. В основе SVM лежит нахождение гиперплоскости в n-мерном пространстве признаков, разделяющую наблюдения на классы. Алгоритм SVM максимизирует *отступ* между гиперплоскостью и объектами классов, которые находятся ближе всего к гиперплоскости. Такие объекты называют *опорными векторами.*

Для построения SVM используют следующую математическую модель *двойственной задачи* квадратичного программирования [2]:

$$\begin{cases} L(\lambda) = -\sum_{i=1}^{l}\lambda_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y_i y_j K(x_i, x_j) \to min_\lambda, \\ \sum_{i=1}^{l}\lambda_i y_i = 0, \quad 0 \le \lambda_i \le C, \end{cases}$$

где $\lambda = (\lambda i, \ldots, \lambda_l)$ — вектор множителей Лангранжа; $x = (x_1, \ldots, x_n)$ — признаковое описание объекта; $y_i \in \{-1, 1\}$ — классы объектов; $K(x_i, x_j)$ — нелинейная функция ядра; C — коэффициент регуляризации.

Решая двойственную задачу, получим *алгоритм* SVM [2]:

$$\begin{cases} a(x) = sign\left(\sum_{i=1}^{l}\lambda_i y_i K(x, x_j) - w_0)\right), \\ w = \sum_{i=1}^{l}\lambda_i y_i x_i, \ w_0 = \langle w, x_i\rangle - y_i, \ \lambda_i > 0, \end{cases}$$

где вектор $w = (w_1, \ldots, w_n) \in R^n$ и скалярный порог $w_0 \in R^n$ — параметры алгоритма. Классификатор $a(x)$ зависит только от *опорных векторов*.

Для создания *нелинейного классификатора* используют полиномиальную, сигмоидальную и радиальную базисную функции ядра. Например, ядро для радиальной базисной функции следующее:

$$K(x, x') = \exp\left(-\gamma\|x - x'\|^2\right),$$

где γ — параметр, регулирующий ширину ядра.

Пример классификации пациенток с раком молочной железы [3]

Университетский госпитальный центр г. Коимбры (Португалия) предоставил данные о 64 пациентках с раком молочной железы и 52 здоровых женщинах. В файле dataR2.csv* содержатся следующие параметры:

1. *Age* — возраст;
2. *BMI* — индекс массы тела, кг/м^2;
3. *Glucose* — содержание сахара в крови, мг/дл;
4. *Insulin* — инсулин, мЕд/л;
5. *HOMA* — индекс HOMA;
6. *Leptin* — лептин, нг/мл;
7. *Adiponectin* — адипонектин, мг/мл;
8. *Resistin* — резистин, нг/мл;
9. *MCP*.1 — моноцитарный хемоаттрактантный белок 1, пг/дл;
10. *Classification* — классификационные метки (1 — здоровые; 2 — онкобольные).

Предварительно произведем разбиение данных на *обучающую* и *тестовую* выборки. Затем, используя алгоритм *SVM*, построим на данных *обучающей выборки* нелинейный классификатор с *радиальной базисной функцией*. Для функции *svm* из пакета *e*1071 зададим следующие параметры: коэффициент регуляции *cost*, равный –1; коэффициент *gamma* для функции ядра, равный 0.5; коэффициент *epsilon* для нечувствительной функции потерь, равный 0.1. После этого, используя функцию *predict*, определим классы наблюдений из *тестовой выборки*, построим матрицу ошибок и рассчитаем статистики. Результаты классификации визуализируем с использованием ROC-кривой.

Решение задачи на языке R

```
> BCancer <- read.table("C:/Data/dataR2.csv",
header = TRUE, sep = ",")
> set.seed(1000)
> indexes <- createDataPartition(BCancer$Classifi
cation,
p = 0.8, list = FALSE)
> train <- BCancer[indexes,]
> test <- BCancer[–indexes,]
> library(e1071)
> classifier <- svm(Classification ~ ., data = train,
type = 'C-classification',
kernel = 'radial', cost = 1,
gamma = 0.5, epsilon = 0.1)
> predictor <- predict(classifier, newdata = test[–10])
> cm <- confusionMatrix(as.factor(predictor),
as.factor(test$Classification))
> print(cm)
> library(pROC)
>pROC_obj <-roc(test$Classification,as.numeric(y_
pred),
smoothed = TRUE, ci = TRUE, ci.alpha = 0.95,
stratified = FALSE, plot = TRUE,
auc.polygon = TRUE, max.auc.polygon = TRUE,
grid = TRUE, print.auc = TRUE,
show.thres = TRUE)
> sens.ci <- ci.se(pROC_obj)
> plot(sens.ci, type = "shape", col = "lightblue")
> plot(sens.ci, type = "bars")
Confusion Matrix and Statistics
```

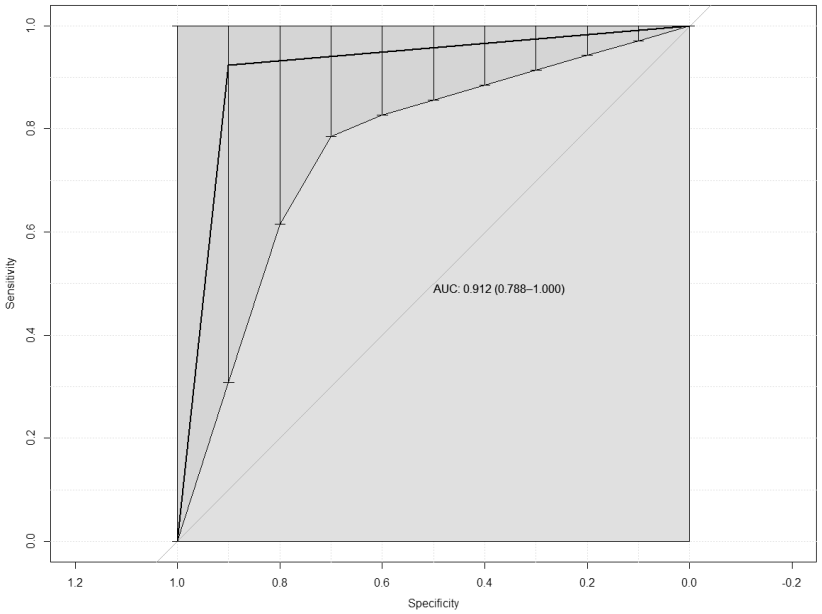

Рис. 1. ROC-кривая для данных тестовой выборки

Reference
Prediction 1 2
1 9 1
2 1 12
Accuracy: 0.913
95% CI: (0.7196, 0.9893)
No Information Rate: 0.5652
P-Value [Acc > NIR]: 0.0003367
Kappa: 0.8231
Mcnemar's Test P-Value: 1.0000000
Sensitivity: 0.9000
Specificity: 0.9231
Pos Pred Value: 0.9000
Neg Pred Value: 0.9231
Prevalence: 0.4348
Detection Rate: 0.3913
Detection Prevalence: 0.4348
Balanced Accuracy: 0.9115
'Positive' Class: 1

Построенная ROC-кривая, оценивающая качество бинарной классификации, представлена на рис. 1. Расчетное значение AUC равно 0,912; 95% доверительный интервал для AUC — [0,778; 1,000].

Алгоритм SVM показал точность классификации 91,3% на *тестовой выборке* из 23 наблюдений. Количество ошибок классификации — 2. Расчетное значение чувствительности составляет 0,9. Проверка по критерию МакНемара *наблюдаемых* и *предсказанных частот* дала уровень значимости p = 1.0000, что подтверждает *нулевую гипотезу* и позволяет прийти к выводу о несущественности различий в частотах между наблюдаемыми и классифицированными данными.

Для сравнения, португальские исследователи М. Патрисио, Дж. Перейра, Дж. Крисостомо и др. в своей работе по классификации рассматриваемых данных с использованием алгоритма SVM [4] приводят информацию о полученных значениях чувствительности от 0,82 до 0,88 и 95% доверительном интервале для AUC, составляющим [0,87; 0,91].

3. Классификация данных с использованием нейронных сетей

Классификация данных является типовой задачей, решаемой в нейронных сетях. В общем случае обученная нейронная сеть рассчитывает вероятность принадлежности наблюдений к одному из классов [5].

Классификация осуществляется в сетях различного типа, однако базовой архитектурой здесь является многослойный персептрон (Multilayer Perceptron, MLP) с сигмоидной функцией активации нейронов. На входные нейроны MLP поступают значения вектора признаков объекта, затем эти значения распространяются на нейроны первого скрытого слоя, и тем самым изменяется размерность задачи. Дальнейшие слои делят объекты на классы в пространстве признаков более высокой размерности, чем исходное. Таким образом, подобрав количество нейронов на скрытых слоях и их функции активации, а затем настроив веса нейронов путем обучения, можно выполнить качественную классификацию данных.

В языке R нейронные сети реализованы в различных пакетах. Например, пакет *neuralnet* [6] содержит сети, работающие с номинальными выходными переменными.

Пример классификации пациенток с раком молочной железы [7]

Клинический научный центр университета Висконсина (США) предоставил данные пациенток с раком молочной железы. В файле breast-cancer-wisconsin.data* содержатся результаты биопсии с 9 ранговыми характеристиками новообразований для 1251 пациентки:

1. *ID number* — идентификационный номер;
2. *Clump Thickness* — размер образований (1–10);
3. *Uniformity of Cell Size* — однородность размера клетки (1–10);
4. *Uniformity of Cell Shape* — однородность формы клетки (1–10);
5. *Marginal Adhesion* — межклеточная мембранная адгезия (1–10);
6. *Single Epithelial Cell Size* — размер эпителиальной клетки (1–10);
7. *Bare Nuclei* — ядро клетки (1–10);
8. *Bland Chromatin* — деконденсированный хроматин (1–10);
9. *Normal Nucleoli* — нормальные ядра (1–10);
10. *Mitoses* — динамика митоза (1–10);
11. *Class* — диагноз для доброкачественной или злокачественной опухоли (2 — «Benign tumor»; 4 — «Malignant tumor»).

Предварительно удалим пропуски из данных, зададим имена переменных, перекодируем числовые значения переменной *Class* в «Benign tumor» или «Malignant tumor», удалим идентификационный номер и разобьём данные на *обучающую* и *тестовую выборки*. Затем на данных *обучающей выборки*, используя функцию *neuralnet*, произведем обучение MPL с пятью нейронами на скрытом слое. После этого, применяя функцию *predict*, классифицируем наблюдения *тестовой выборки*, построим матрицу ошибок и рассчитаем статистики. С помощью функции *plot* визуализируем архитектуру нейронной сети.

Решение задачи на языке R

```
> BCancer <- read.table("C:/Data/breast-cancer-wisconsin.data", header = TRUE, sep = ",")
> BCancer <- na.omit(BCancer)
> colnames(BCancer)<-c("Sample_code_number",
"Clump_Thickness",
"Uniformity_of_Cell_Size",
"Uniformity_of_Cell_Shape",
"Marginal_Adhesion",
"Single_Epithelial_Cell_Size",
"Bare_Nuclei",
"Bland_Chromatin",
"Normal_Nucleoli",
"Mitoses",
"Class")
> BCancer["Class"][BCancer["Class"] == 2] <-
"Benign tumor"
> BCancer["Class"][BCancer["Class"] == 4] <-
"Malignant tumor"
> BCancer <- BCancer[,—1]
> set.seed(1000)
> indexes <- createDataPartition(BCancer$Class, p = 0.8,
list = FALSE)
> train <- BCancer[indexes,]
> test <- BCancer[–indexes,]
> library(neuralnet)
> model <- neuralnet(Class ~., train, hidden = c(5), linear.output = FALSE)
> ypred = neuralnet:: compute(model, test[,—10])
> yhat = ypred$net.result
> yhat <- data.frame("yhat" = ifelse(max.col(yhat[,1:2]) == 1, "Benign tumor", "Malignant tumor"))
> cm <- confusionMatrix(as.factor(test[, 10]), as.factor(yhat$yhat))
> print(cm)
Confusion Matrix and Statistics
Reference
Prediction Benign tumor Malignant tumor
Benign tumor 84  4
Malignant tumor 2  45
Accuracy: 0.9556
95% CI: (0.9058, 0.9835)
No Information Rate: 0.637
P-Value [Acc > NIR]: <2e-16
Kappa: 0.903
Mcnemar's Test P-Value: 0.6831
Sensitivity: 0.9767
Specificity: 0.9184
Pos Pred Value: 0.9545
Neg Pred Value: 0.9574
Prevalence: 0.6370
Detection Rate: 0.6222
Detection Prevalence: 0.6519
Balanced Accuracy: 0.9476
'Positive' Class: Benign tumor
```

После удаления пропусков количество наблюдений уменьшилось до 683. Нейронная сеть, обученная на выборке из 547 наблюдений, имеет следующую архитектуру: 9 входных нейронов, 5 нейронов на скрытом слое и 2 выходных нейрона (рис. 2). Выходные значения MLP — апостериорные вероятности принадлежности наблюдений к классам.

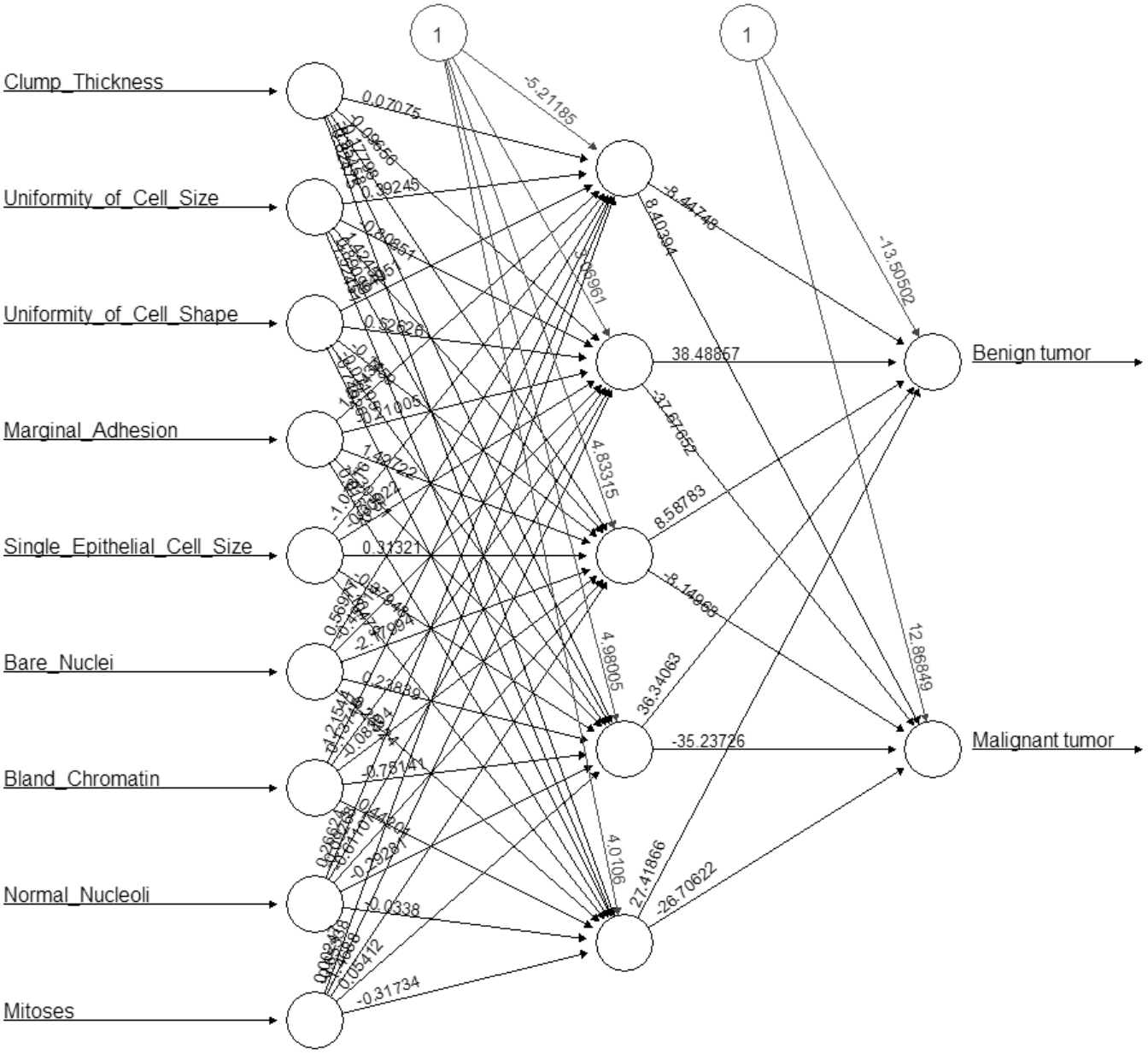

Рис. 2. Архитектура нейронной сети

Точность классификации на *тестовой выборке* из 135 наблюдений составила 95,56%. Среди пациенток с *доброкачественной опухолью* было 2 ошибки классификации (чувствительность 97,67%), а среди пациенток со *злокачественной опухолью* — 4 ошибки (специфичность 91,84%). Сравнение по критерию МакНемара наблюдаемых и прогнозируемых частот дало уровень значимости $p = 0.6831$, что позволяет принять *нулевую гипотезу* и сделать вывод о несущественности различий между частотами наблюдаемых и классифицированных данных.

Для сравнения, австралийский исследователь Н. Аббасс решил эту задачу классификации с использованием эволюционного многоцелевого подхода к нейронным сетям, основанном на алгоритме дифференциальной эволюции Парето. Точность классификации составила 98,1% [8].

4. Классификация данных с использованием деревьев решений

Деревья решений являются древовидными иерархическими структурами, строящимися на основе машинного обучения. Классификация объектов здесь осуществляется с помощью решающих правил формата «если…то…». Эти проверки реализуются в узлах

деревьев, а листья содержит объекты, относящиеся только к определённым классам. Популярные алгоритмы построения деревьев решений — CART, C4.5, CHAID и другие. Например, алгоритм CART [9], предложенный профессорами Л. Брейманом, Дж. Фридманом, Ч. Стоуном и Р. Олшеном в 1984 г., реализован в языке R с помощью функции *rpart* [10].

Пример кластеризации пациенток с раком молочной железы [11]

Национальный институт биомедицинской инженерии в г. Порто (Португалия) предоставил данные пациенток с раком молочной железы. В файле BreastTissue.xls* содержатся результаты обследования 106 пациенток, включающие 9 характеристик электрического импеданса образцов тканей молочной железы:

1. *Class* — классы («car» — карцинома; «fad» — фиброаденома; «mas» —местопатия; «gla» — железистая ткань; «con» — соединительная ткань; «adi» — жировая ткань);
2. *I*0 — инпеданс на нулевой частоте;
3. *PA*500 — фазовый угол на частоте 500 кГц;
4. *HFS* — высокочастотный наклон (крутизна) фазового угла;
5. *DA* — расстояние импеданса между спектральными концами;
6. *AREA* — площадь области под спектром;
7. *A.DA* — площадь, отнесенная к величине *DA*;
8. *MAX.IP* — максимум спектра;
9. *DR* — расстояние между *I*0 и реальной частью точки максимальной частоты;
10. *P* — длина спектральной кривой.

Предварительно исходный файл BreastTissue.xls преобразуем в текстовый с расширением «txt». Разобьём данные на *обучающую* и *тестовую* выборки. Затем, используя функцию *rpart* с алгоритмом CART, построим бинарное дерево решений на данных *обучающей выборки*. Глубину дерева *maxdepth* зададим, равную 5; меру сложности *cp* зададим, равную –1, что является гарантией полного выращивания дерева. Визуализируем дерево с помощью функции *rpart.plot*. Характеристики построенного дерева выведем при помощи функции *printcp*. График кросс-валидации выведем при помощи функции *plotcp*. В конце, используя функцию *predict*, классифицируем наблюдения *тестовой выборки*, построим матрицу ошибок и выведем статистики.

Решение задачи на языке R

```
> BTissue <- read.table("C:/Data/BreastTissue.txt",
header = TRUE, sep = "\t")
> library(caret)
> set.seed(1000)
> indexes <- createDataPartition(BTissue$Class, p =
0.8,
list = FALSE)
> train <- BTissue[indexes,]
> test <- BTissue[–indexes,]
> library(rpart)
> library(rpart.plot)
> tree <- rpart(Class ~ ., data = train,
method = "class", maxdepth = 5,
minsplit = 2, minbucket = 1, cp = –1)
> rpart.plot(tree, fallen.leaves = FALSE, cex = 0.7)
> printcp(tree)
> plotcp(tree)
> BTissuePredict <- predict(tree, test, type = 'class')
> cm <- confusionMatrix(as.factor(BTissuePredict),
as.factor(test$Class))
> print(cm)
Variables actually used in tree construction:
[1] A.DA Area DA I0 Max.IP P PA500
Root node error: 69/87 = 0.7931
n = 87
CP nsplit rel error xerror xstd
1 0.246377 0 1.00000 1.02899 0.052370
2 0.173913 1 0.75362 0.81159 0.064739
3 0.115942 3 0.40580 0.68116 0.067371
4 0.028986 4 0.28986 0.50725 0.066287
5 0.014493 6 0.23188 0.52174 0.066578
6 0.000000 7 0.21739 0.50725 0.066287
7–1.000000 8 0.21739 0.50725 0.066287
Confusion Matrix and Statistics
Reference
Prediction adi car con fad gla mas
adi 3 0 0 0 0 0
car 0 4 0 0 0 0
con 1 0 2 0 0 0
fad 0 0 0 3 0 2
gla 0 0 0 0 3 1
mas 0 0 0 0 0 0
Overall Statistics
Accuracy: 0.7895
95% CI: (0.5443, 0.9395)
No Information Rate: 0.2105
P-Value [Acc > NIR]: 1.139e-07
Kappa: 0.7467
Mcnemar's Test P-Value: NA
```

Дерево решений было построено на *обучающей выборке* из 87 наблюдений с использованием решающих правил для переменных *A.DA, Area, DA, I*0, *Max.IP, P и PA*500. В выведенной на экран таблице представлены такие характеристики дерева, как *мера сложности CP*, *количество ветвлений nsplit*, *ошибка обучающей выборки rel error* и *ошибка кросс-валидации xerror*. Ошибка кросс-валидации достигла

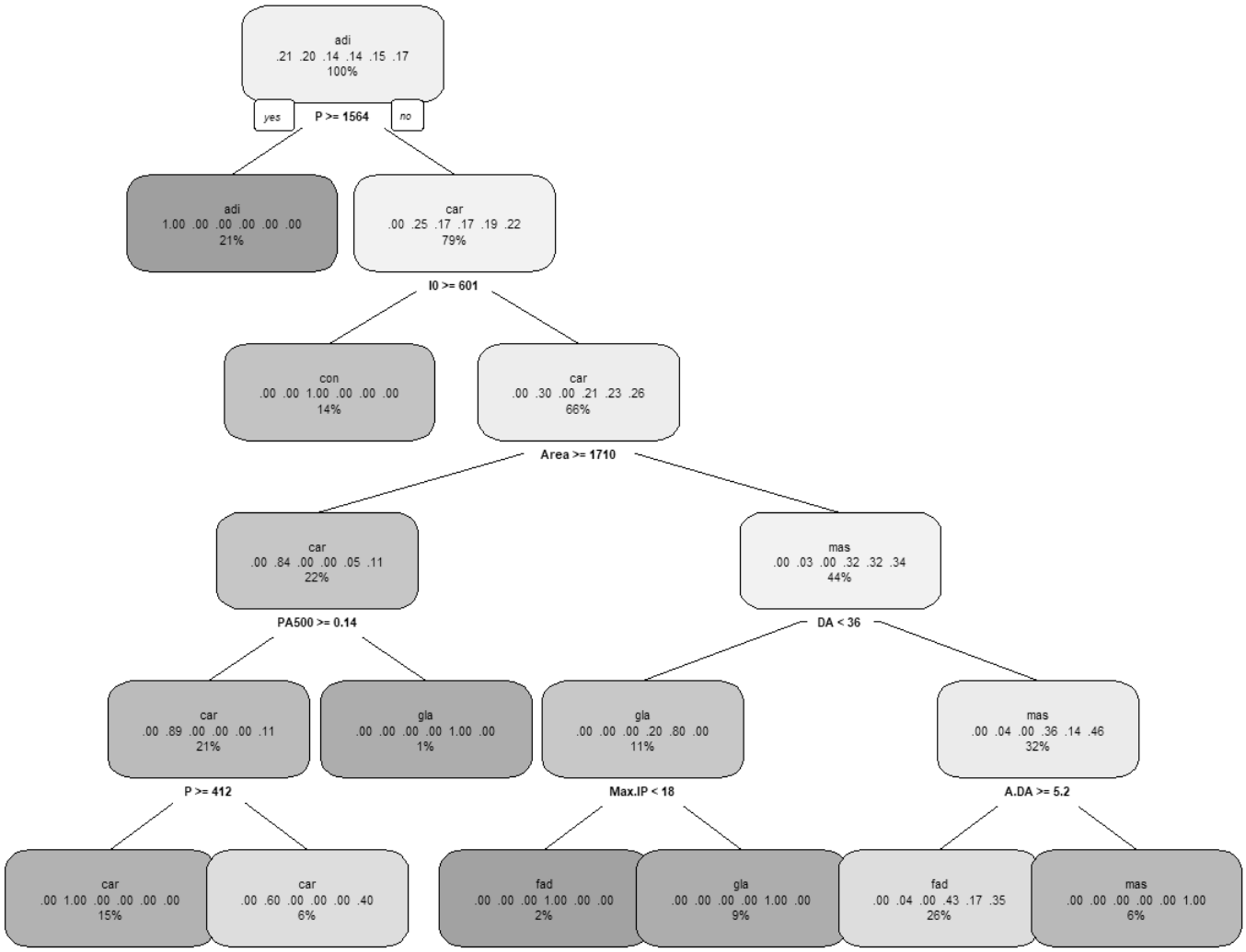

Рис. 3. Построенное бинарное дерево решений

минимума для величин *nsplit* и *CP*, равных соответственно 4 и 0,028986.

Визуализация полученного бинарного дерева решений с глубиной 5, количеством узлов 8 и количеством листов 9 представлена на рис. 3. График зависимости ошибки кросс-валидации от размера дерева и меры сложности *CP* представлен на рис. 4. Согласно этому графику оптимальная глубина дерева составляет 5.

Построенное дерево решений показало точность классификации 78,95% на *тестовой выборке* из 19 наблюдений. Классы «car», «con», «fad» и «gla» были определены без ошибок (точность 100%); класс «adi» — с *одной ошибкой* (точность 75%); однако все *три наблюдения* класса «mas» были определены неверно (точность 0%). Критерий МакНемара в данном случае не рассчитывается.

Для сравнения, китайские исследователи Л. Чанг, Ч. Тяньтянь, Л. Чансин, применяя для классификации анализируемых данных *метод опорных векторов*, получили точность 80,625% на 32% тестовой выборке [12].

5. Классификация категориальных данных с использованием нейронных сетей

Нейронные сети могут осуществлять классификацию категориальных данных при соответствующем кодировании этих данных в числовые значения. Номинальные переменные, являющиеся бинарными, кодируются как 0 и 1 (неактивное и активное состояния), и под такие переменные отводится *один нейрон* входного слоя *сети*. Для переменных, имеющих *n* уровней, каждый конкретный уровень кодируется как 0 и 1, и под эти переменные отводится *n нейронов входного слоя* [5].

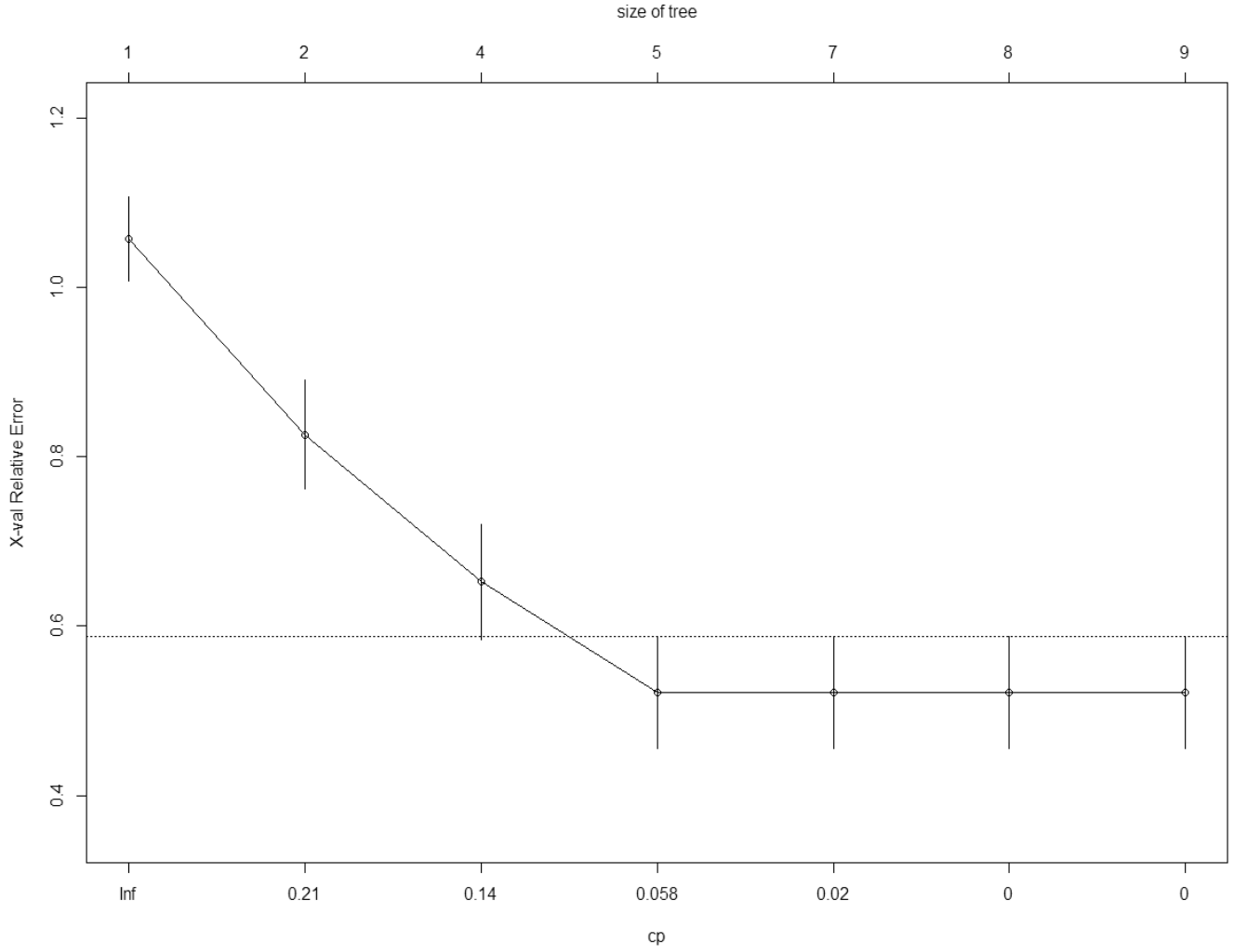

Рис. 4. График зависимости ошибки кросс-валидации от размеров дерева

Так, в языке R классификацию категориальных данных осуществляет функция *nnet* [13], которая производит обучение MLP с одним скрытым слоем.

Пример кластеризации пациенток с раком молочной железы [14]

Институт онкологии г. Любляна (Югославия) предоставил данные о 286 пациентках с раком молочной железы. В файле breast-cancer.data* содержатся следующие категориальные данные:

1. *Class* — классы («no-recurrence-events» — без повторения; «recurrence-events» — повторяющееся событие);
2. *age* — возрастные группы (10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99);
3. *menopause* — предклимактерический или климактерический период («lt40», «ge40» или «premeno»);
4. *tumorsize* — размер новообразования (0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59);
5. *invnodes* — количество подмышечных лимфатических узлов, содержащих метастатический рак молочной железы, видимых при гистологическом исследовании (0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39);
6. *nodecaps* — метастазы рака в лимфатические узлы («*yes*» — да; «*no*» — нет);
7. *degmalig* — степень злокачественности (1, 2, 3);
8. *breast* — грудь («left» — левая; «right» — правая);
9. *breastquad* — зоны груди («left-up» — слева вверху; «left-low» — слева внизу; «right-up» — справа вверху; «right-low» — справа внизу; «central» — по центру);
10. *irradiat* — проведение лучевой терапии («yes»— да; «no» — нет).

Предварительно зададим имена для всех переменных и преобразуем эти переменные к категориальному типу с помощью функции *as.factor*. Затем произведем разбиение данных на *обучающую* и *тестовую выборки*. Обучим нейронную сеть *nnet* на данных *обучающей выборки*, используя следующие параметры: количество нейронов на скрытом слое 8; границы для диапазона начальных значений весов [*rang*; *–rang*], равные 10^{-6}, коэффициент ослабления весов *decay*, равный 10^{-4}. Поскольку классифицирующая переменная является бинарной, MLP будет содержать *один* выходной нейрон. Затем, используя функцию *predict*, классифицируем наблюдения *тестовой выборки*, построим матрицу ошибок и рассчитаем статистики. Визуализируем архитектуру нейронной сети и построим диаграмму важности входных переменных с помощью функций *plotnet* и *olden* соответственно.

Решение задачи на языке R

```
> BCancer <- read.table("C:/DataSK/breast-cancer.data",
header = FALSE, sep = ",")
> BCancer <- na.omit(BCancer)
> colnames(BCancer)<-c("Class", "age", "menopause",
"tumorsize", "invnodes",
"nodecaps", "degmalig", "breast",
"breastquad", "irradiat")
> BCancer$Class <- as.factor(BCancer$Class)
> BCancer$age <- as.factor(BCancer$age)
> BCancer$menopause <- as.factor(BCancer$menopause)
> BCancer$tumorsize <- as.factor(BCancer$tumorsize)
>BCancer$invnodes<-as.factor(BCancer$invnodes)
> BCancer$nodecaps <- as.factor(BCancer$nodecaps)
> BCancer$degmalig <- as.factor(BCancer$degmalig)
> BCancer$breast <- as.factor(BCancer$breast)
> BCancer$breastquad <- as.factor(BCancer$breastquad)
> BCancer$irradiat <- as.factor(BCancer$irradiat)
> library(caret)
> set.seed(5000)
> indexes <- createDataPartition(BCancer$Class, p = 0.8,
list = FALSE)
> train <- BCancer[indexes,]
> test <- BCancer[–indexes,]
> library(nnet)
> newNet <- nnet(Class ~ ., data = train, size = 8,
rang = 1.0e-06, decay = 1.0e-04, maxit = 2000)
> library(NeuralNetTools)
> plotnet(newNet)
> olden(newNet)
> myPrediction <- predict(newNet, newdata = test,
type = "class")
> cm <- confusionMatrix(as.factor(myPrediction),
as.factor(test$Class))
> print(cm)
Confusion Matrix and Statistics
Reference
Prediction no-recurrence-events recurrence-events
no-recurrence-events 32  5
recurrence-events 7  11
Accuracy: 0.7818
95% CI: (0.6499, 0.8819)
No Information Rate: 0.7091
P-Value [Acc > NIR]: 0.1488
Kappa: 0.49
Mcnemar's Test P-Value: 0.7728
Sensitivity: 0.8205
Specificity: 0.6875
Pos Pred Value: 0.8649
Neg Pred Value: 0.6111
Prevalence: 0.7091
Detection Rate: 0.5818
Detection Prevalence: 0.6727
Balanced Accuracy: 0.7540
'Positive' Class: no-recurrence-events
```

После очистки от пропусков количество наблюдений уменьшилось до 277. Нейронная сеть, обученная на выборке из 222 наблюдений, имеет архитектуру, представленную на рис. 5: 32 входных нейронов, 8 нейронов на скрытом слое, 1 нейрон на выходном слое и 273 весов межнейронных связей.

Построенная диаграмма *относительной важности* входных переменных нейронной сети, описанная в работе [15], показана на рис. 6. Полученные по этой диаграмме ранжированные значения категориальных переменных, приведены в табл. 1. Наиболее важные значения входных переменных здесь имеют ранги от 1 до 13.

Обученная нейронная сеть показала точность классификации 78,18% на тестовой выборке из 55 наблюдений. Класс «no-recurrence-events» определен с *семью ошибками* (чувствительность 82,05%), а класс «recurrence-events» — с пятью ошибками (специфичность 68,75%). Проверка по критерию МакНемара *наблюдаемых* и *предсказанных частот* для двух классов дала уровень значимости $p = 0.7728$, что позволяет принять *нулевую гипотезу* и прийти к выводу о несущественности различий в частотах между наблюдаемыми и классифицированными данными.

Для сравнения, в работе американских исследователей Р. Мичалски, И. Мозэтика, Дж. Нонга, Н. Лавраса

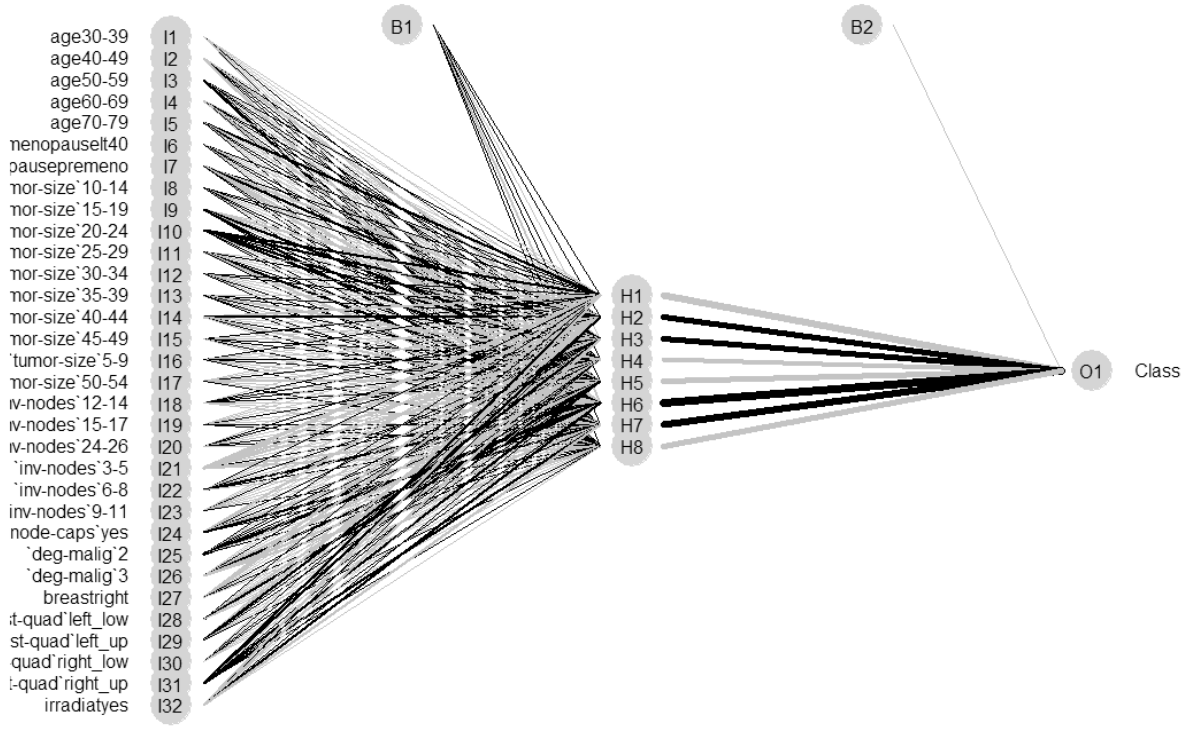

Рис. 5. Архитектура нейронной сети

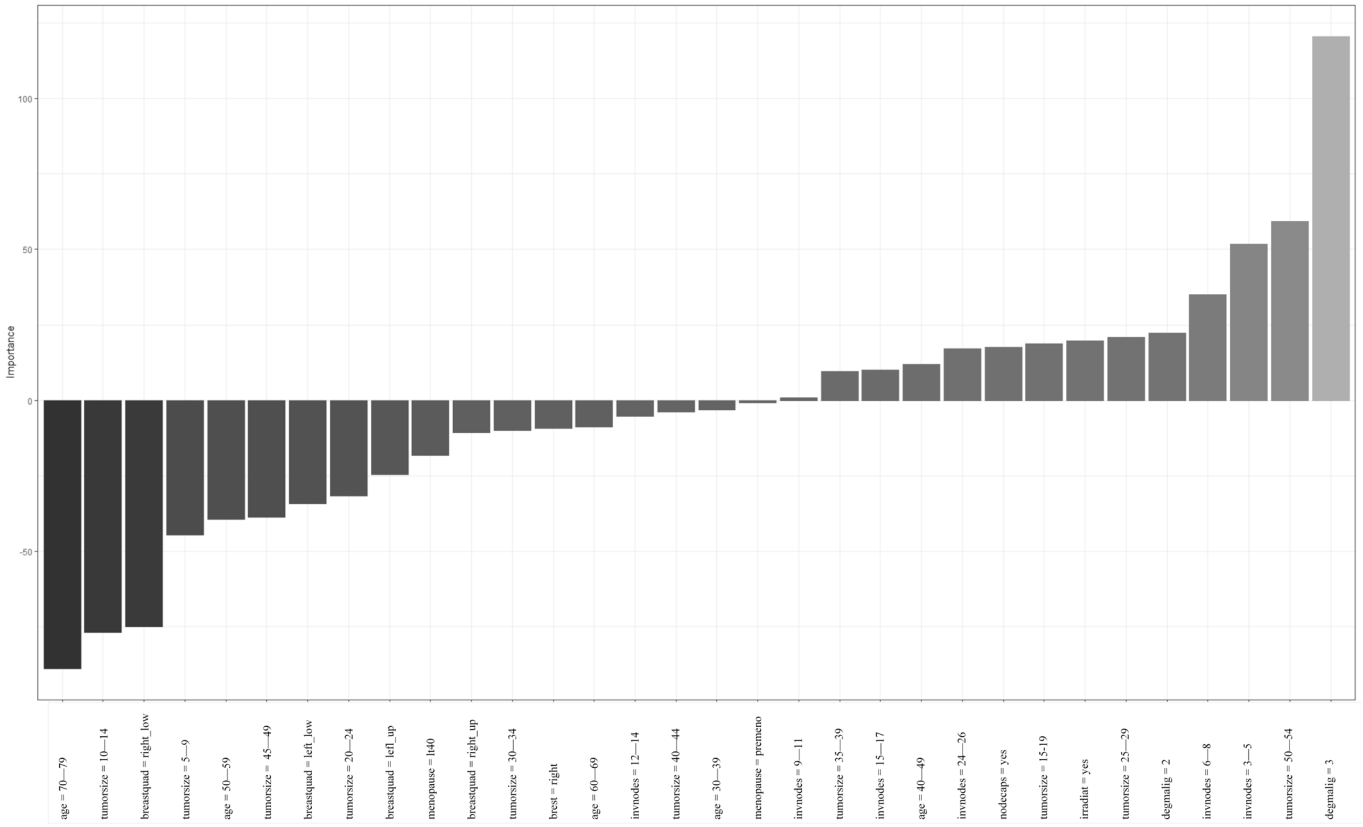

Рис. 6. Диаграмма относительной важности входных переменных

Таблица 1. Ранжированные по важности значения категориальных переменных для нейронной сети

Ранг	Категориальная переменная	Значение переменной	Ранг	Категориальная переменная	Значение переменной
1	degmalig	3	17	tumorsize	40–44
2	tumorsize	50–54	18	invnodes	12–14
3	invnodes	3–5	19	age	60–69
4	invnodes	6–8	20	brest	right
5	degmalig	2	21	tumorsize	30–34
6	tumorsize	25–29	22	breastquad	right_up
7	irradiat	yes	23	menopause	*lt*40
8	tumorsize	15–19	24	breastquad	lefl_up
9	nodecaps	yes	25	tumorsize	20–24
10	invnodes	24–26	26	breastquad	left_low
11	age	40–49	27	tumorsize	45–49
12	invnodes	15–17	28	age	50–59
13	tumorsize	35–39	29	tumorsize	5–9
14	invnodes	9–11	30	breastquad	right_low
15	menopause	premeno	31	tumorsize	10–14
16	age	30–39	32	age	70–79

[16] приводится информация о точности 64% при классификации рассматриваемых данных с применением *алгоритма AQ*.

Заключение

Язык R является эффективным средством классификации данных медицинских исследований. Авторами были решены задачи классификации данных рака молочной железы, размещенные в репозитории UCI Machine Learning Repository. Точность классификации на тестовых выборках, полученная с использованием метода опорных векторов, нейронных сетей и деревьев решений, находится на уровне или превосходит результаты опубликованных зарубежных исследований.

ЛИТЕРАТУРА

1. Касюк, С.Т. Кластерный анализ данных медицинских исследований с использованием языка R / С.Т. Касюк, Г.А. Диденко, О.А. Степанова // Современная наука: актуальные проблемы теории и практики. серия: естественные и технические науки. — 2021. — № 4–2. — С. 23–32.
2. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 15.06.2022)
3. Breast Cancer Coimbra Data Set [Электронный ресурс]. — Режим доступа: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra (дата обращения: 15.06.2021).
4. Patrício, M., Pereira, J., Crisóstomo, J. et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 18, 29 (2018). https://doi.org/10.1186/s12885–017–3877–1
5. Нейронные сети. STATISTICA Neural Networks: Методология и технологии современного анализа данных / Под редакцией В.П. Боровикова. — 2-е изд., перераб. и доп. — М.: Горячая линия — Телеком, 2008. — 392 с.
6. Package «neuralnet», February 7, 2019, Version 1.44.2 [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf (дата обращения: 15.06.2022).
7. Breast Cancer Wisconsin (Diagnostic) Data Set [Электронный ресурс]. — Режим доступа: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (дата обращения: 05.06.2022).
8. Abbass, H. An Evolutionary Artificial Neural Networks Approach for Breast Cancer Diagnosis. Artificial Intelligence in Medicine. Volume 25, Issue 3, July 2002, P. 265–281. https://doi.org/10.1016/S0933–3657(02)00028–3
9. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, C. Stone. —Wadsworth: Chapman and Hall/CRC, 1984. — 368 p.
10. Package «rpart», January 24, 2022, Version n 4.1.16 [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/rpart/rpart.pdf (дата обращения: 15.06.2022).
11. Breast Tissue Data Set [Электронный ресурс]. — Режим доступа: http://archive.ics.uci.edu/ml/datasets/breast+tissue (дата обращения: 15.06.2022).

12. Chang, L., Tiantian C., Changxing L. Breast Tissue Classification based on Electrical Impedance Spectroscopy. International Conference on Industrial Technology and Management Science (2015), P. 237–240. https://doi.org/10.2991/itms-15.2015.56
13. Package «nnet», January 13, 2022, Version 7.3–17 [Электронный ресурс]. — Режим доступа: https://cran.r-project.org/web/packages/nnet/nnet.pdf (дата обращения: 15.06.2022).
14. Breast Cancer Data Set [Электронный ресурс]. — Режим доступа: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer (дата обращения: 15.06.2022).
15. Olden, J., Joy, M., Death, R. An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data. Ecological Modelling, Vol.178, 3–4 (2004). https://doi.org/10.1016/S0304–3800(04)00156–5
16. The multi-purpose incremental learning system AQ15 and its testing application to three medial domains / R. Michalski, I. Mozetic, J. Hong, N. Lavrac // Proceedings of the Fifth National Conference on Artificial Intelligence. — Philadelphia, PA: Morgan Kaufmann, 1986. — P. 1041–1045.

© Касюк Сергей Тимурович (sergey.kasyk@gmail.com),
Диденко Галина Александровна (pga80@mail.ru), Степанова Оксана Александровна (okalst@mail.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»

«Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации