Экология человека
2021, № 3, с. 54–64

УДК 519.2 + 61 DOI: 10.33396/1728-0869-2021-3-54-64

ИНТЕЛЛЕКТУАЛЬНЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ В БИОМЕДИЦИНСКИХ ИССЛЕДОВАНИЯХ: ДЕРЕВЬЯ КЛАССИФИКАЦИИ

©2021 г. **[1]А. Н. Наркевич, [1]К. А. Виноградов, [2,3,4,5]А. М. Гржибовский**

[1]ФГБОУ ВО «Красноярский государственный медицинский университет им. проф. В. Ф. Войно-Ясенецкого», г. Красноярск; [2]ФГБОУ ВО «Северный государственный медицинский университет», г. Архангельск; [3]Западно-Казахстанский медицинский университет им. Марата Оспанова, г. Актобе, Казахстан; [4]Казахский Национальный Университет им. аль-Фараби, г. Алматы, Казахстан; [5]ФГАОУ ВО «Северо-Восточный федеральный университет им. М. К. Аммосова», г. Якутск

Задачи современных биомедицинских исследований требуют все более сложных методов анализа данных. В последнее время под анализом данных все реже понимают проверку статистических гипотез с помощью классических статистических критериев и оценку связи между признаками с помощью корреляционного анализа и все чаще в понятие анализ данных вкладывается более всестороннее изучение полученных в результате эмпирических исследований данных с применением многомерных статистических методов. Одним из таких методов анализа с большим потенциалом использования в технологиях искусственного интеллекта, анализе больших данных и машинном обучении данных является дерево классификации, или дерево решений. Целью данной статьи является рассмотрение вопросов применения деревьев классификации в медико-биологических исследованиях, а также представление примеров их построения в наиболее часто применяемых статистических программах. В статье приведены описание задачи, которая может быть решена с помощью деревьев классификации, пример набора данных для их построения, а также построение модели дерева классификации в IBM SPSS Statistics и StatSoft Statistica. Применение при анализе данных медико-биологических экспериментов дерева классификации, как одного из относительно легко используемых и интерпретируемых методов многомерного анализа данных, позволит более глубоко изучать закономерности явлений и состояний в области медицины и биологии.

Ключевые слова: деревья классификации, деревья решений, SPSS, Statistica, математическое моделирование, искусственный интеллект, машинное обучение

INTELLIGENT DATA ANALYSIS IN BIOMEDICAL RESEARCH: CLASSIFICATION TREES

[1]A. N. Narkevich, [1]K. A. Vinogradov, [2,3,4,5]A. M. Grjibovski

[1]Voino-Yasenetsky Krasnoyarsk State Medical University, Krasnoyarsk, Russia; [2]Northern State Medical University, Arkhangelsk, Russia; [3]West Kazakhstan Marat Ospanov Medical University, Aktobe, Kazakhstan; [4]Al-Farabi Kazakh National University, Almaty, Kazakhstan; [5]M. K. Ammosov North-Eastern Federal University, Yakutsk, Russia

Modern analytical tasks in biomedical research require increasingly sophisticated methods of data analysis. In recent years, the term data analysis is not only related to classical statistical tests for hypothesis testing and correlation analysis for studying associations between variables. Classification tree or decision tree analysis is getting more and more frequently used in biomedical research. In this paper we present the use of classification trees in biomedical research and provide examples of their construction in the most commonly used statistical programs. The article is constructed as a problem solving exercise using classification trees with an example of a data set for creation of classification trees and description of how to build a classification tree model in IBM SPSS Statistics and StatSoft Statistica software. Moreover, we provide recommendations on how the results of this analysis should be presented in a scientific article. The use of the classification trees has a potential to contribute to better understanding of the factors behind the observed phenomena in medicine and biology.

Key word: classification trees, decision trees, SPSS, Statistica, mathematical modeling, artificial intelligence, machine learning

Библиографическая ссылка:

Наркевич А. Н., Виноградов К. А., Гржибовский А. М. Интеллектуальные методы анализа данных в биомедицинских исследованиях: деревья классификации. 2021. № 3. С. 54–64.

For citing:

Narkevich A. N., Vinogradov K. A., Grjibovski A. M. Intelligent Data Analysis in Biomedical Research: Classification Trees. *Ekologiya cheloveka (Human Ecology)*. 2021, 3, pp. 54-64.

Уровень анализа данных в современных медико-биологических исследованиях не стоит на месте — на сегодняшний день под понятием «анализ данных» уже намного реже подразумевается применение классических статистических критериев и коэффициентов для нахождения различий между группами или оценки связи между признаками [1, 20]. В настоящий момент все чаще в понятие «анализ данных» исследователями вкладывается куда более всестороннее изучение полученных в результате медико-биологического эксперимента данных с применением многомерных статистических и математических методов. Под многомерными методами нами подразумеваются методы, которые позволяют одновременно учитывать не один, а совокупность изучаемых у объектов признаков [11, 14, 15]. Одним из многомерных методов анализа

данных является дерево классификации, или дерево решений [5, 9, 13, 16, 19]. Примеры использования данной модели можно найти при решении различных аналитических задач в социологии [8], пульмонологии [10], инфектологии [2, 3], анестезиологии [12], оториноларингологии [17], кардиологии [22], стоматологии [21], онкологии [18], общественном здоровье и здравоохранении [4] и других областях.

Целью данной статьи является рассмотрение вопросов применения деревьев классификации в медико-биологических исследованиях, а также представление примеров их построения в наиболее часто применяемых статистических программах.

Описание задачи, которая может быть решена с помощью деревьев классификации

В целом деревья классификации позволяют решать так называемую задачу прогнозирования качественного признака (иначе такая задача называется задачей классификации). Как правило, задача классификации в медико-биологических исследованиях используется, когда выяснение реального значения качественного признака является либо очень дорогим, опасным для здоровья человека, либо в принципе затруднительным по каким-либо причинам (долгое по времени и т. д.). Например, диагностика туберкулеза является довольно затруднительной, весьма затратной и самое главное – более-менее достоверные данные о наличии или отсутствии туберкулеза могут быть получены лишь через 20–90 дней, а пациента необходимо начинать лечить от туберкулеза или другого заболевания уже сейчас. При этом имеется информация, что наличие туберкулеза связано с наличием у человека различных факторов риска. Этот пример можно применить практически к любому заболеванию – если у человека имеются факторы риска заболевания, то вероятность, что человек имеет данное заболевание, выше (в этом как раз и есть суть факторов риска). В данном случае «туберкулез» является качественным признаком (есть туберкулез или нет туберкулеза). То есть можно попытаться на основании данных о наличии у человека различных факторов риска спрогнозировать – страдает он туберкулезом или нет. В целом даже неважно, знает ли исследователь, каким именно образом факторы риска связаны с развитием туберкулеза, достаточно знания или даже предположения о том, что эта связь есть [6].

Для осуществления такого прогноза (решения задачи классификации – построения математической модели, которая сможет всех пациентов разделить на два класса: больные туберкулезом и не больные туберкулезом) надо собрать данные согласно следующему дизайну. Набираются две группы пациентов: одна – с заранее достоверно установленным диагнозом туберкулез, другая – с заранее достоверно установленным отсутствием диагноза туберкулез. Затем выясняется наличие или отсутствие конкретных факторов риска и в одной, и в другой группе.

Таким образом, у каждого пациента должны быть данные о наличии или отсутствии факторов риска, на основе которых будет осуществляться прогнозирование, и данные о достоверном подтверждении диагноза или о достоверном отсутствии диагноза туберкулез. После применения деревьев классификации на таких данных можно будет выяснять у пациентов наличие факторов риска и получать прогнозируемое значение качественного признака – есть туберкулез или нет туберкулеза.

Дальнейшая иллюстрация построения деревьев классификации будет осуществлена на приведенном примере. Однако примеров подобных задач можно привести достаточно много. Так, база данных включает информацию о женщинах, имеющих по результатам ультразвукового исследования образование в яичниках. Информация включает в себя признаки, отражающие анамнез жизни, акушерско-гинекологический анамнез, данные лабораторных исследований, в том числе результаты анализов на онкомаркеры. Помимо этого у каждой женщины имеется информация о характере образования (злокачественное или доброкачественное), полученная в результате его удаления и гистологического исследования. Задача заключается в том, чтобы построить математическую модель, которая позволит с достаточной уверенностью определить, к какой группе относится женщина – со злокачественным или доброкачественным образованием. На основании такой классификации врачом может быть определена наиболее эффективная дальнейшая тактика ведения.

Еще одним примером может служить задача классификации пациентов с инфарктом миокарда после проведения операции аортокоронарного шунтирования на два класса – класс пациентов, у которых будут сохраняться когнитивные нарушения через 12 месяцев после операции, и класс пациентов, у которых через 12 месяцев когнитивных нарушений не будет. В данном случае в качестве входных признаков для построения дерева классификации также могут использоваться анамнестические, клинические, лабораторные данные. Более того, в данном примере может быть использована информация, характеризующая непосредственно произошедший инфаркт миокарда (сторона и глубина поражения, стадия, на которой поступил пациент, и т. д.), и, к примеру, информация, характеризующая когнитивную сферу непосредственно после операции и при выписке.

Приведем пример из наркологической практики. Задача заключается в том, чтобы по поведенческим и психиатрическим характеристикам пациента при поступлении в наркологическое отделение определить наиболее вероятное наркотическое вещество, прием или введение которого привело к возникновению девиантного поведения. Примером применения деревьев классификации в хирургии может служить решение следующей потенциальной задачи. Необходимо классифицировать новорожденных, имеющих дефекты двенадцатиперстной кишки, на два класса – детей, у которых возникнут послеоперационные осложнения,

Экология человека
2021, № 3, с. 54–64

и детей, у которых такие осложнения не возникнут. В качестве признаков, на основе которых будет осуществляться построение дерева классификации, могут быть использованы данные о родившемся ребенке, степени дефекта и планируемой операции.

Описание набора данных для построения деревьев классификации

Примером данных для построения деревьев классификации может служить следующий набор информации о пациентах: наличие заболевания (болен туберкулезом или не болен), рост, вес, индекс массы тела (ИМТ), пол, частое переохлаждение на работе, выполнение тяжелого физического труда, постоянная нервно-психическая нагрузка на работе, образование (основное общее, среднее, средне-профессиональное, средне-специальное, неоконченное высшее, высшее), наличие контакта с больным туберкулезом, пребывание в течение жизни в пенитенциарных учреждениях, наличие сопутствующих заболеваний (ВИЧ, сахарный диабет (СД), язвенная болезнь желудка и двенадцатиперстной кишки (ЯБЖиДПК), другие заболевания желудочно-кишечного тракта (ЖКТ), злоупотребление алкоголем, наркомания, психические заболевания, хронические неспецифические заболевания легких (ХНЗЛ), пылевые заболевания, вирусные заболевания печени) и табакокурение. При формировании базы данных очень важно уделить внимание ее правильной структуре. Необходимо максимально разбить информацию на разные столбцы базы данных и не следует указывать все, к примеру, сопутствующие заболевания в одном столбце. В таблице приведены столбцы тестовой базы данных и их кодировка.

Столбцы базы данных и их кодировка

№ столбца	Название столбца	Кодировка	Интерпретация кодов
1	Наличие заболевания	1 / 0	Болен туберкулезом / Не болен туберкулезом
2	Рост, см	Без кодировки	Абсолютные значения роста
3	Вес, кг	Без кодировки	Абсолютные значения веса
4	ИМТ	Без кодировки	Абсолютные значения ИМТ
5	Пол	1 / 0	Мужской / Женский
6	Возраст, лет	Без кодировки	Абсолютные значения возраста
7	Частое переохлаждение	1 / 0	Да / Нет
8	Тяжелый физический труд	1 / 0	Да / Нет
9	Нервно-психическая нагрузка	1 / 0	Да / Нет
10	Основное общее образование	1 / 0	Да / Нет
11	Среднее образование	1 / 0	Да / Нет
12	Средне-профессиональное образование	1 / 0	Да / Нет
13	Средне-специальное образование	1 / 0	Да / Нет
14	Неоконченное высшее образование	1 / 0	Да / Нет
15	Высшее образование	1 / 0	Да / Нет
16	Контакт с больным туберкулезом	1 / 0	Да / Нет
17	Пребывание в пенитенциарных учреждениях	1 / 0	Да / Нет
18	ВИЧ	1 / 0	Да / Нет
19	СД	1 / 0	Да / Нет
20	ЯБЖиДПК	1 / 0	Да / Нет
21	Другие болезни ЖКТ	1 / 0	Да / Нет
22	Злоупотребление алкоголем	1 / 0	Да / Нет
23	Наркомания	1 / 0	Да / Нет
24	Психические заболевания	1 / 0	Да / Нет
25	ХНЗЛ	1 / 0	Да / Нет
26	Пылевые заболевания	1 / 0	Да / Нет
27	Вирусные заболевания печени	1 / 0	Да / Нет
28	Табакокурение	1 / 0	Да / Нет

Как видно из представленной таблицы, каждый минимальный объем информации о пациентах максимально разбит на столбцы таким образом, чтобы каждый столбец содержал данные о наличии или отсутствии какого-либо одного признака у пациента. Исключение составляют лишь те столбцы, в которых информация может быть представлена в количественном виде, например, столбцы «Рост, см», «Вес, кг» и т. д.

Несомненно, увеличение числа признаков, на основании которых будет осуществляться построение математической модели дерева классификации, повысит шансы получить модель с наиболее приемлемым качеством классификации. Помимо тех признаков, которые были включены в наш пример, исследователями могут включаться признаки, отражающие результаты лабораторных, клинических и инструментальных методов исследования пациентов, если они есть в наличии или могут быть получены из медицинской документации. При этом необходимо отметить, что признаки в базе данных должны быть максимально детализированы и формализованы. К примеру, признак «наличие у пациента гипертонической болезни II стадии и 3 степени» вряд ли можно назвать максимально детализированным и формализованным. Скорее всего его даже нельзя назвать признаком, если рассматривать его с позиции построения математических моделей. Наиболее приемлемо разбить такой признак на три отдельных: наличие гипертонической болезни (будет содержать 0, если гипертоническая болезни отсутствует, и 1, если присутствует у пациента), стадия (будет содержать в виде арабских цифр обозначение стадии – «1», «2» и т. д.) и степень (также будет содержать в виде арабских цифр обозначение степени).

Таким образом, база данных для иллюстрации включает в себя 28 признаков у 728 пациентов: 342 пациента, у которых достоверно установлено наличие туберкулеза, 386 пациентов, у которых достоверно установлено отсутствие туберкулеза. Задача

заключается в том, чтобы построить математическую модель дерева классификации, которое позволит по наличию 27 (кроме «Наличие заболевания») признаков пациента классифицировать его в группу больных или не больных туберкулезом. Так как при включении большего числа признаков в примерную базу данных теряется возможность иллюстрации результатов построения деревьев классификации в рамках статьи, нами объем базы данных ограничен вышеописанными 28 признаками.

Построение модели дерева классификации в IBM SPSS Statistics

Для построения модели дерева классификации в IBM SPSS Statistics необходимо в основном меню выбрать «Анализ» – «Классификация» – «Деревья классификации» (рис. 1). В появившемся окне в поле «Зависимая переменная» необходимо включить признак, который планируется прогнозировать – «Наличие заболевания», а в поле «Независимые переменные» признаки, на основе которых планируется прогнозировать «Рост, см», «Вес, кг», «ИМТ», «Пол», «Возраст, лет» и т. д. То есть будет построена модель дерева классификации, позволяющая на основе независимых переменных прогнозировать значение класса, к которому относится пациент – к классу 1 (болен туберкулезом) или 0 (не болен туберкулезом).

Необходимо отметить, что в поле «Метод построения» можно изменить метод построения дерева классификации. К основным, наиболее часто используемым, методам построения относятся «CHAID», «Исчерпывающий CHAID», «CRT» и «QUEST». Сущностью всех методов построения деревьев классификации заключается в последовательном разделении всех единиц наблюдения на подгруппы таким образом, чтобы по возможности в отдельные подгруппы входили наблюдения разных классов.

Метод построения «CHAID» формирует разделение

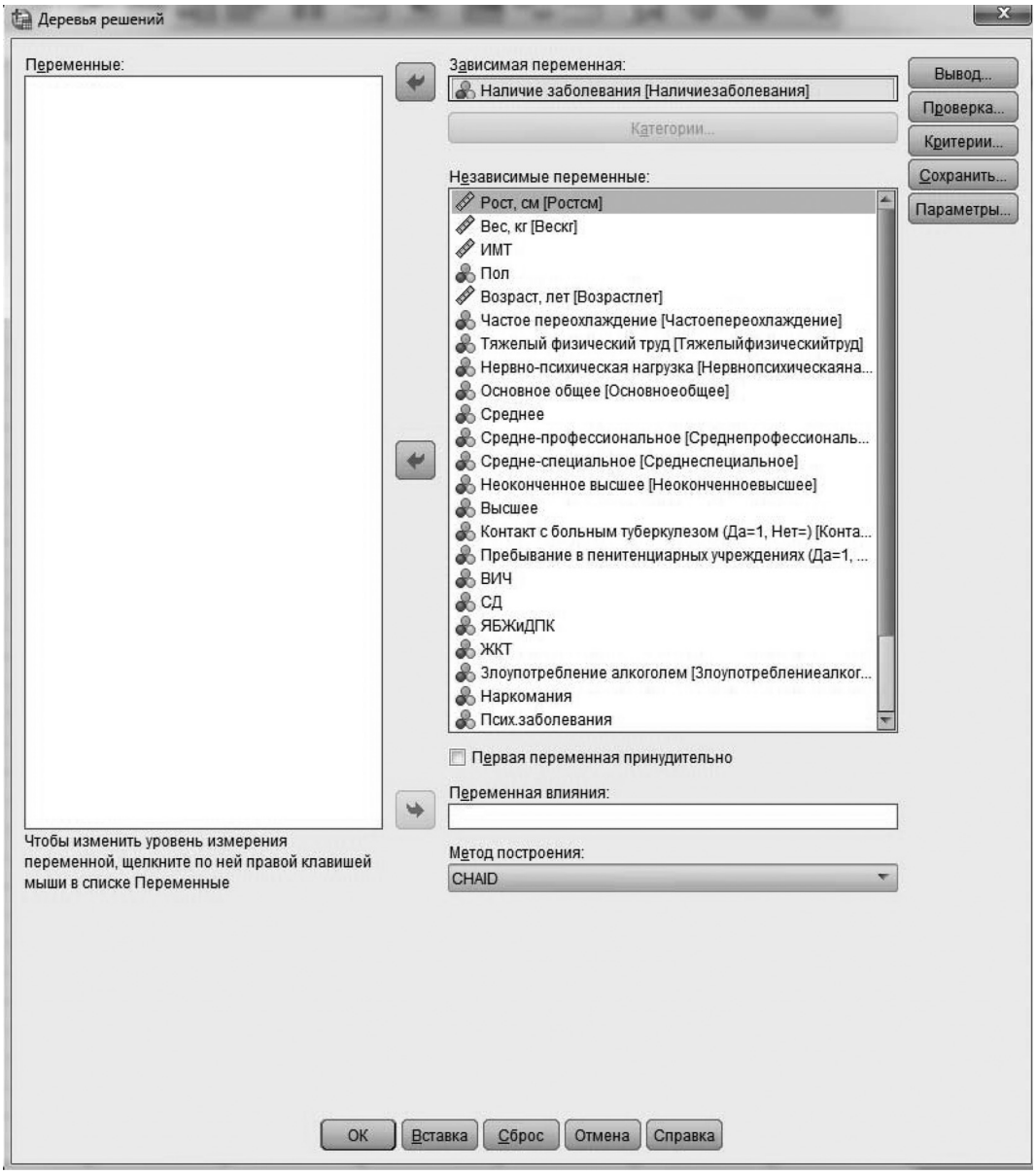

Рис. 1. Окно построения деревьев классификации

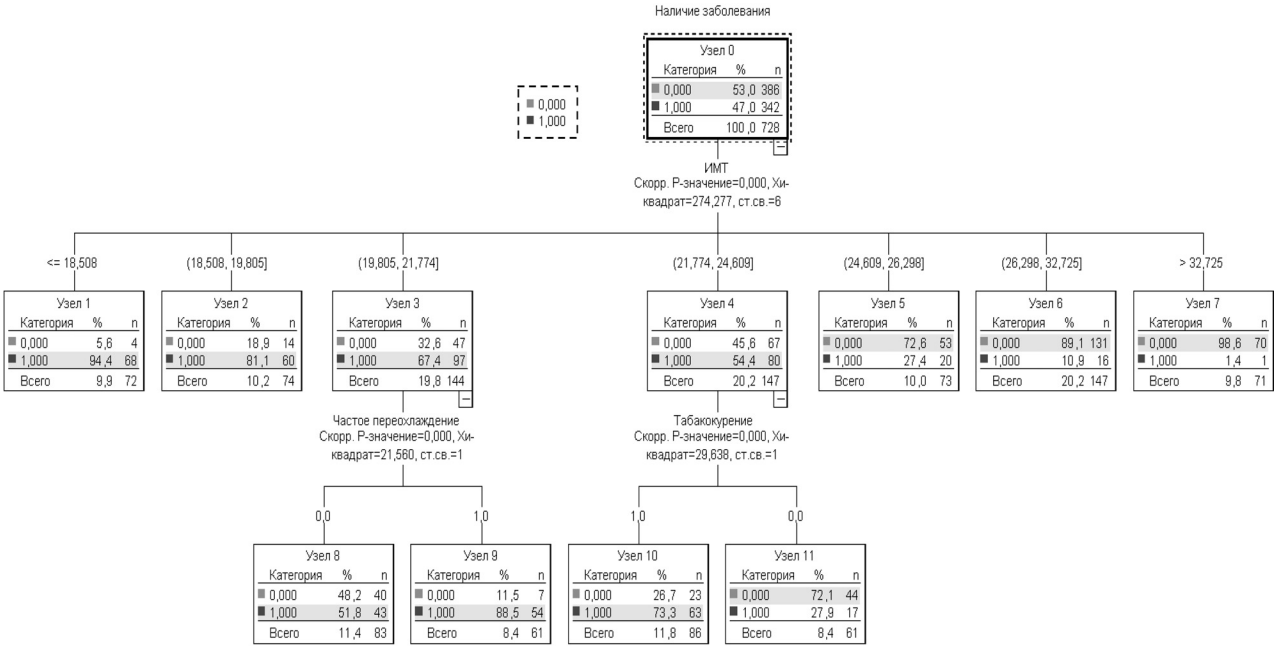

Рис. 2. Дерево классификации, построенное методом «CHAID»

узлов на основе статистики хи-квадрат и в отличии от методов «CRT» и «QUEST» может генерировать не только бинарные разделения, но и множественные. Метод построения «Исчерпывающий CHAID» является более углубленной модификацией метода «CHAID». Метод построения «CRT» позволяет осуществлять только бинарные разделения узлов, а метод «QUEST» является ускоренным методом «CRT». После выбора зависимой и независимых переменных необходимо нажать кнопку «ОК» для получения дерева классификации (рис. 2).

Подробнее остановимся на полученном дереве классификации. В самом верху дерева классификации находится «Узел 0» (так называемый корень дерева классификации). С данного узла начинается разделение единиц наблюдения на подгруппы. Под нулевым узлом представлен признак, по которому пациенты в первую очередь разделяются на подгруппы. В данном примере этим признаком является «ИМТ». Далее приведены интервалы значений ИМТ, при которых пациенты относятся в тот или иной узел. Так, если у пациента ИМТ меньше или равен значению 18,508, то он относится к «Узел 1», если больше 18,508 и меньше либо равно 19,805, то к «Узел 2», если больше 19,805, но меньше либо равно 21,774, то к «Узел 3», и т. д.

«Узел 1», «Узел 2», «Узел 5», «Узел 6» и «Узел 7» являются терминальными узлами, так как после них разделения на дочерние узлы не происходит. «Узел 3» далее в соответствии с наличием у пациента такого фактора риска, как частое переохлаждение, также разделяется на два дочерних узла: если данный фактор риска отсутствует у пациента, то пациент относится к «Узел 8», если имеется, то к «Узел 9». «Узел 4» также имеет разделение на два дочерних узла в соответствии с табакокурением пациента: если пациент курит, то он относится к «Узел 10», если не курит, то к «Узел 11»

Таким образом, для классификации пациента с использованием полученного дерева классификации достаточно трех параметров – «ИМТ», «Частое переохлаждение» и «Табакокурение». В соответствии с этими значениями необходимо дойти до терминального узла, параметры которого и будут свидетельствовать о классе пациента.

Необходимо учитывать, что пример, на котором иллюстрируется построение деревьев классификации, и полученные математические модели не могут претендовать на истинный научный результат, так как данные пациентов являются лишь частью базы данных, которая использовалась при проведении оригинального исследования [7].

Прежде чем приводить примеры классификации пациентов с применением полученного дерева классификации, рассмотрим еще несколько таблиц, полученных при его построении. В таблице «Риск» (рис. 3) в столбце «Оценка» приведена доля неверно классифицированных единиц наблюдения, то есть при применении полученного дерева классификации неверно классифицировано 19,5 % пациентов.

Риск

Оценка	Стандартная ошибка
,195	,015

Метод построения: CHAID
Зависимая переменная:
Наличие заболевания

Рис. 3. Таблица «Риск»

Доля неверно классифицированных пациентов также подтверждается данными таблицы «Классификация» (рис. 4). Так, в данной таблице приведены

наблюдаемые классы пациентов и их предсказанные классы. На пересечении строки «1» и столбца «1» приведено количество пациентов, у которых имеется туберкулез, и дерево классификации предсказало, что у них есть туберкулез. На пересечении строки «0» и столбца «0» – количество пациентов, у которых отсутствует туберкулез, и дерево классификации предсказало, что у них его нет. Эти значения составляют верно предсказанные наблюдения. Общий процент верно предсказанных наблюдений составил 80,5.

Классификация

Наблюденные	Предсказанные		
	0	1	Процент правильных
0	298	88	77,2%
1	54	288	84,2%
Общая процентная доля	48,4%	51,6%	80,5%

Метод построения: CHAID
Зависимая переменная: Наличие заболевания

Рис. 4. Таблица «Классификация»

Для иллюстрации примера использования полученного дерева классификации методом «CHAID» воспользуемся следующими параметрами одного из пациентов, имеющимися в базе данных: ИМТ – 19,98, частое переохлаждение – да, табакокурение – да. Так как ИМТ пациента попадает в интервал от 19,805 до 21,774, то от «Узел 0» осуществляется переход к «Узел 3». Так как этот узел не является терминальным, то далее необходимо оценить следующий параметр пациента. Пациент имеет фактор риска частого переохлаждения. В связи с этим осуществляется переход к «Узел 9». Так как «Узел 9» является терминальным узлом, то для окончательного прогноза наличия или отсутствия туберкулеза у пациента необходимо проанализировать данные этого узла. В «Узел 9» входит больше пациентов с наличием туберкулеза (88,5 %), в связи с чем можно сделать прогноз, о том, что у пациента, на основе данных которого делается прогноз, имеется туберкулез.

Для построения дерева классификации другими методами необходимо произвести те же самые действия, что и для построения уравнения методом «CHAID», только необходимо произвести изменения в поле «Метод построения». Приведем лишь дерево классификации, которое построено методом «CRT» и имеет наибольшую точность среди деревьев классификации, построенных всеми четырьмя методами (рис. 5). Общий процент верно предсказанных наблюдений с применением данного дерева классификации составил 84,1, тогда как метод «Исчерпывающий CHAID» позволил получить 83,1, а «QUEST» – 83,4. Как видно из представленного на рис. 5 дерева классификации, оно включает лишь два параметра пациента – табакокурение и ИМТ.

Необходимо отметить, что построение деревьев классификации имеет довольно широкий спектр на-

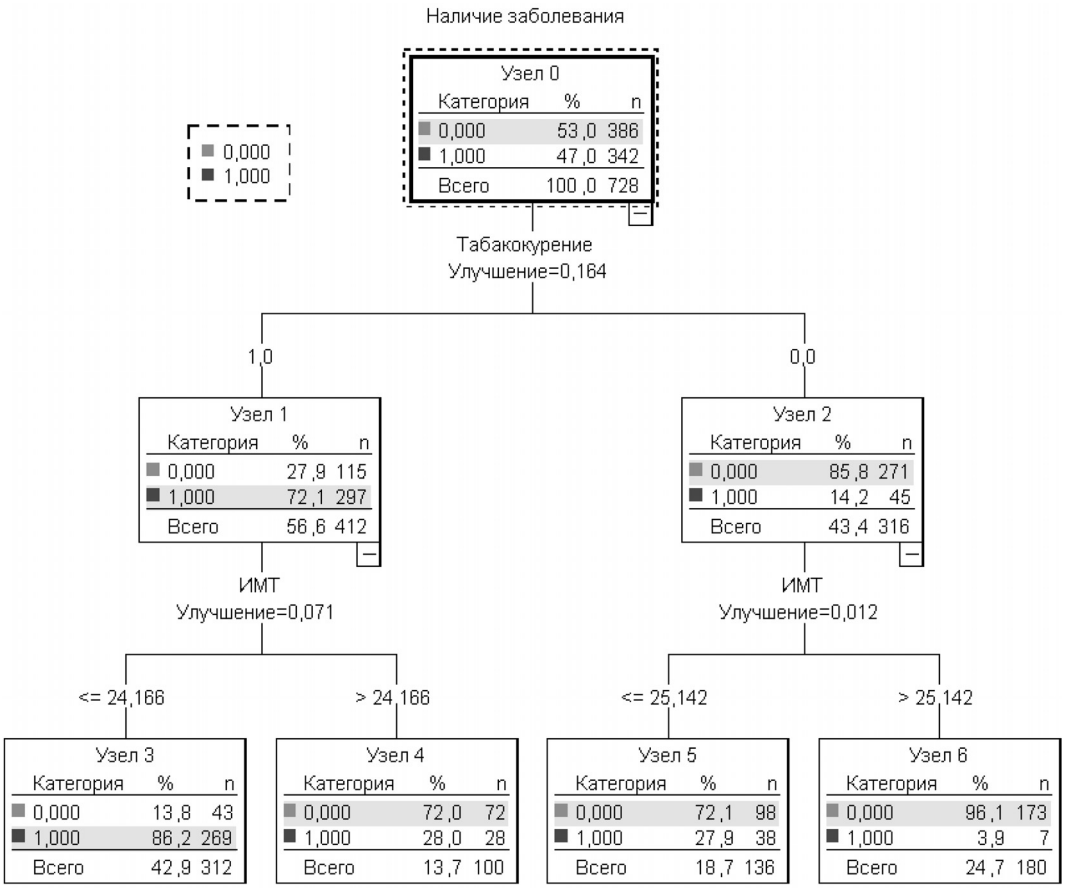

Рис. 5. Дерево классификации, построенное методом «CRT»

строек и позволяет рассчитывать довольно большое число показателей, характеризующих данное дерево. Ввиду довольно большого списка таких настроек и показателей они не приведены в данной статье, а примеры построения деревьев классификации приведены с целью показать их принципиальные возможности, процесс их построения и возможного использования. Рекомендуется при решении задачи классификации с использованием деревьев классификации воспользоваться всеми методами их построения с различными настройками, сравнить их результаты и выбрать оптимальную модель.

Построение модели дерева классификации в StatSoft Statistica

Построение деревьев классификации в системе StatSoft Statistica осуществляется путем использования различных пунктов меню. Для выбора данных пунктов меню необходимо в основном меню выбрать «Добыча данных», а затем один из следующих пунктов:

1. Общие деревья классификации и регрессии.
2. Общие CHAID модели.
3. Интерактивные деревья (C&RT, CHAID).
4. Растущие деревья классификации и регрессии.
5. Случайные леса регрессии и классификации.

Так как общие принципы построения деревьев классификации не меняются в зависимости от выбранного пункта меню, в статье будет рассмотрен пример построения дерева классификации, полученного путем выбора «Общие деревья классификации и регрессии». После выбора данного пункта меню в появившемся окне в поле «Type of analysis» необходимо выбрать «Standard C&RT», а в поле «Specification method» – «Quick specs dialog». После нажатия «OK» в открывшемся окне необходимо нажать кнопку «Variables», после чего будет открыто окно выбора переменных (рис. 6).

В окне выбора переменных в поле «Dependent» необходимо выбрать параметр, который планируется прогнозировать – «Наличие заболевания», в поле «Categorical pred.» необходимо выбрать качественные параметры, на основе которых будет осуществляться прогнозирование. В текущем примере такими параметрами является большинство из имеющихся в наборе данных. В поле «Continuous pred.» необходимо выбрать количественные и ранговые параметры, на ос-

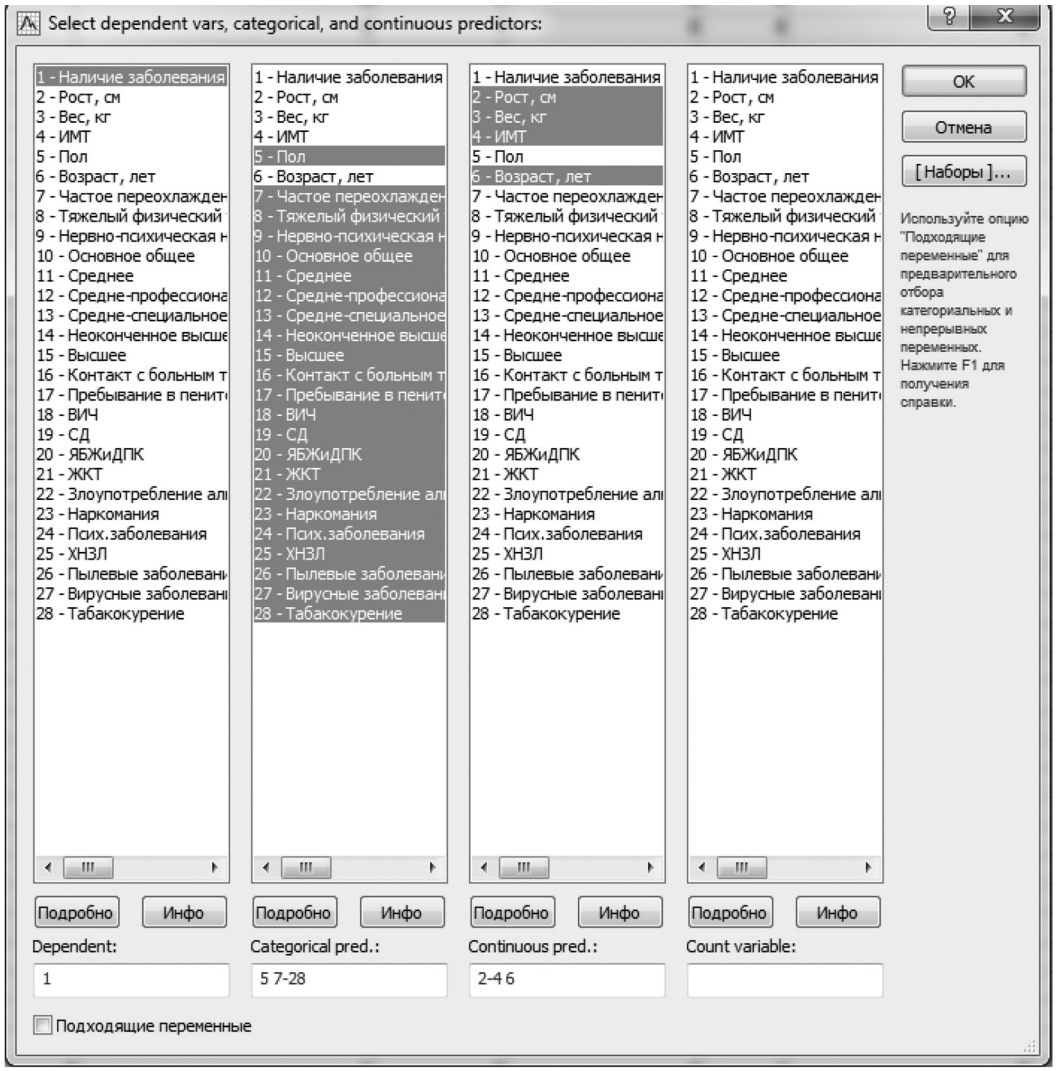

Рис. 6. Окно выбора переменных

нове которых будет осуществляться прогнозирование. В текущем примере такими параметрами являются «Рост, см», «Вес, кг», «ИМТ» и «Возраст, лет». То есть будет построена модель дерева классификации, позволяющая на основе этих данных прогнозировать значение класса, к которому относится пациент, – больных или не больных туберкулезом.

После выбора нужных переменных необходимо нажать кнопку «OK». После возврата в окно настроек дерева классификации необходимо установить «галочку» в поле «Categorical response (categorical dependent variable)». После того как выбраны все переменные и отмечены все необходимые настройки построения дерева классификации, необходимо в окне настроек, имеющих представленный на рис. 7 вид, нажать кнопку «OK».

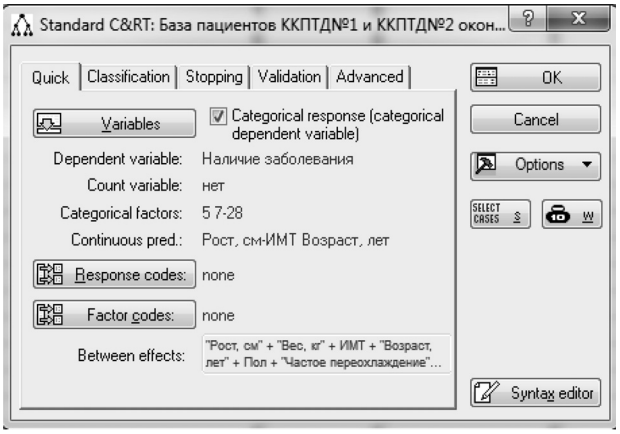

Рис. 7. Окно настроек построения дерева классификации

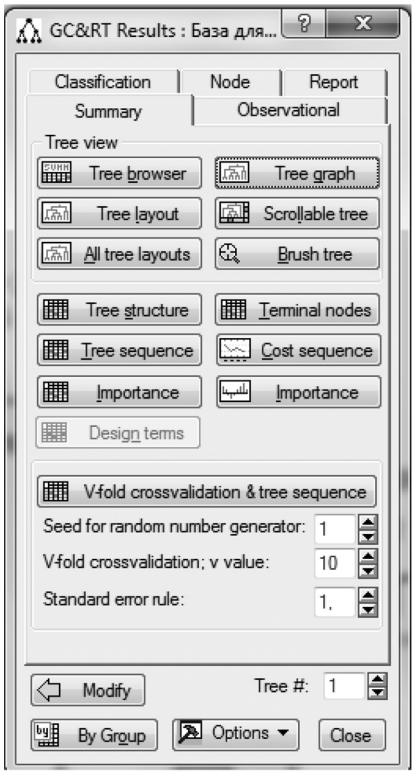

Рис. 8. Окно результатов построения дерева классификации

После нажатия кнопки «OK» будет открыто окно результатов построения дерева классификации (рис. 8). На вкладке «Summary» необходимо нажать кнопку «Tree graph» для получения графического представления полученного дерева классификации (рис. 9).

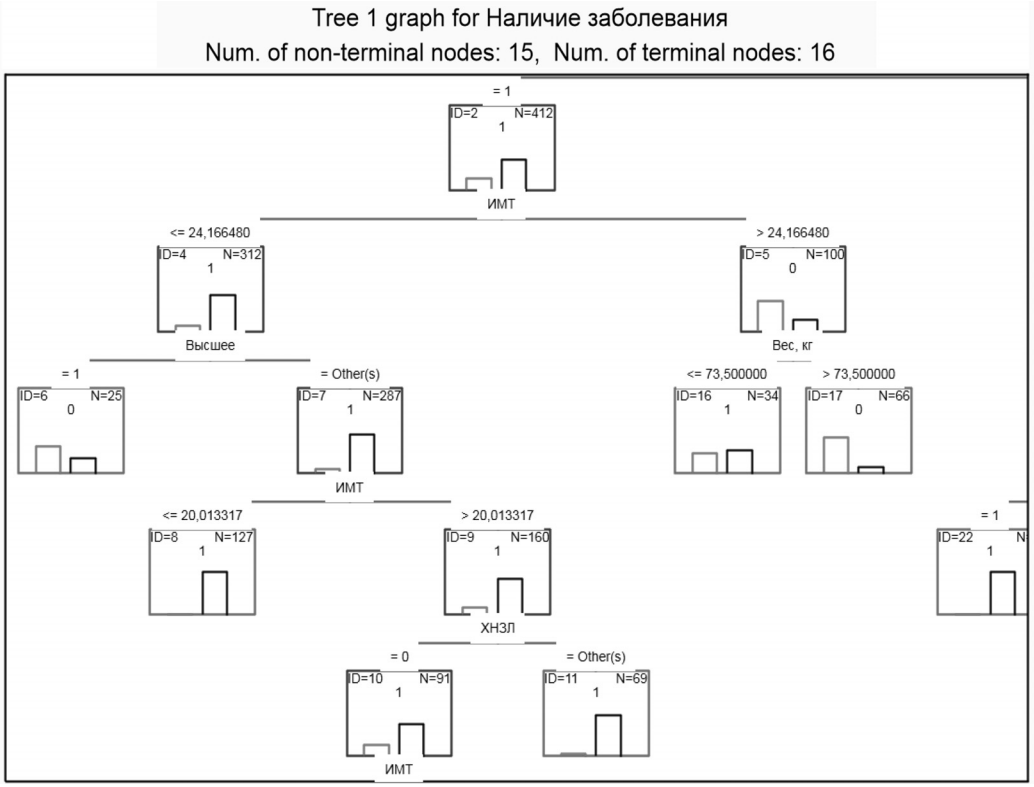

Рис. 9. Графическое представление полученного дерева классификации

В связи с тем, что с применением статистической системы StatSoft Statistica было получено довольно большое дерево классификации, включающее в себя 31 узел, на рис. 9 приведена только часть графического представления модели. Затем необходимо перейти на вкладку «Classification» для получения результатов оценки качества полученного дерева классификации.

После нажатия кнопки «Tree graph» можно рассмотреть подробнее полученное дерево классификации. Его использование осуществляется так же, как и использование деревьев классификации, полученных с помощью IBM SPSS Statistics.

Для оценки качества полученного дерева классификации после перехода на вкладку «Classification» необходимо нажать кнопку «Predicted vs. Observed by classes», после чего будет получена таблица классификации, в которой приведены все необходимые параметры (рис. 10).

Classification matrix 1 (База пациентов ККПТД№1 и
Dependent variable: Наличие заболевания
Options: Categorical response, Analysis sample

	Observed	Predicted 0	Predicted 1	Row Total
Number	0	336	50	386
Column Percentage		88.65%	14.33%	
Row Percentage		87.05%	12.95%	
Total Percentage		46.15%	6.87%	53.02%
Number	1	43	299	342
Column Percentage		11.35%	85.67%	
Row Percentage		12.57%	87.43%	
Total Percentage		5.91%	41.07%	46.98%
Count	All Groups	379	349	728
Total Percent		52.06%	47.94%	

Рис. 10. Таблица классификации

Таблица классификации разбита на три блока. В первом блоке представлены результаты классификации пациентов, не имеющих туберкулез – «Observed=0», во втором блоке – пациентов, имеющих туберкулез – «Observed=1», в третьем блоке – всех пациентов – «Observed=All Groups».

Рассмотрим результаты классификации подробнее. Данные первого блока свидетельствуют о том, что у пациентов, не имеющих туберкулез, полученным деревом классификации предсказывается отсутствие туберкулеза («Predicted 0») в 87,05 % случаев, а наличие туберкулеза («Predicted 1») – в 12,95 %. Данные второго блока свидетельствуют о том, что у пациентов, имеющих туберкулез, полученным деревом классификации предсказывается отсутствие туберкулеза («Predicted 0») в 12,57 % случаев, а наличие туберкулеза («Predicted 1») – в 87,43 %.

Представление результатов построения деревьев классификации

При представлении результатов построения деревьев классификации в научной статье или диссертации следует учитывать то, что искушенного читателя будут интересовать, во-первых, стартовые условия, при которых осуществилось построение дерева классификации: число входных признаков, их наименование, на какие классы исследователь классифицировал пациентов, метод построения дерева классификации. Все это позволит читателю воспроизвести полученный исследователем результат или провести похожее исследование.

Во-вторых, наибольший интерес будет вызывать непосредственно дерево классификации. Если размер дерева классификации позволяет использовать его в тексте статьи или диссертации (как в случае с деревом классификации, полученным с применением IBM SPSS Statistics), то лучше непосредственно привести его изображение как на рис. 2. В таком случае самостоятельный анализ читателем такого изображения даст ответы на многие возникающие вопросы. Если формат статьи или диссертации не позволяет уместить дерево классификации в виде изображения (как в случае с деревом классификации, полученным с применением StatSoft Statistica), то его стоит как следует описать. Степень детализации описания, естественно, определяет автор, но максимально, насколько это возможно, стоит указать следующее: число признаков, которые включены в дерево классификации, наименование этих признаков, число уровней дерева, число узлов, наиболее интересные с позиции автора правила, разделяющие на несколько узлов.

В-третьих, при представлении результатов построения деревьев классификации важно отразить качество классификации с применением модели. В качестве показателей, отражающих качество классификации, как правило, используются чувствительность, специфичность и точность. В качестве дополнительных показателей могут быть использованы прогностическая ценность положительного и отрицательного результата, отношение правдоподобия положительного и отрицательного результата. На расчете данных показателей и их интерпретации останавливаться не будем, но каждый из них характеризует качество классификации с разных позиций, а расчет их можно осуществить с применением онлайн калькуляторов, используя значения, которые приводятся в таблицах результатов классификации (см. рис. 4 и 10). Помимо самих показателей желательно снабдить их 95 % доверительными интервалами.

Какие значения чувствительности, специфичности или точности считать приемлемыми, сказать довольно сложно, так как это зависит от конкретной области медицины и даже задачи, которую решает исследователь в рамках этой области. К примеру, если в арсенале медицинских специалистов уже существует онкомаркер, использование которого позволяет установить наличие онкологического заболевания с чувствительностью 87 %, специфичностью 92 % и точностью 90 %, то использование дерева классификации, которое имеет более низкие значения данных показателей, не будет иметь смысла. В связи с этим для каких-то задач будет достаточно и 70–80 %, а для каких-то задач недостаточно будет и 90–95 % значений. Это можно оценить, только сравнивая результаты, полученные методами, имеющимися в арсенале медицинских специалистов.

Для представления в научной публикации результатов, полученных в ходе реализации примера, приведенного в данной статье, раздел материалов и методов может содержать примерно следующий текст.

«Для классификации пациентов на больных туберкулезом и пациентов с отсутствием данного заболевания осуществлялось построение математической модели дерева классификации. Для построения дерева классификации в качестве входных признаков использовались 28 параметров пациентов. При этом 23 параметра представлены в бинарном виде, где 0 – отсутствие признака, 1 – наличие признака: наличие частого переохлаждения, тяжелого физического труда, нервно-психической нагрузки, основного общего, среднего, средне-профессионального, среднеспециального, неоконченного высшего и высшего образования, наличие контакта с больным туберкулезом, ВИЧ-инфекции, сахарного диабета, язвенной болезни желудка и двенадцатиперстной кишки, других болезней ЖКТ, психических заболеваний, хронических неспецифических заболеваний легких, пылевых заболеваний, вирусных заболеваний печени, наркомании, пребывание в пенитенциарных учреждениях в течение жизни, злоупотребление алкоголем, табакокурение. Один параметр (пол) представлен в бинарном виде, где 0 – женский пол, 1 – мужской пол, а четыре параметра представлены в количественном виде: рост, вес, индекс массы тела и возраст.

Классификация осуществлялась на два класса: 0 – отсутствие туберкулеза, 1 – наличие туберкулеза. Построение дерева классификации осуществлялось методом CHAID. Для оценки качества классификации использовались показатели точности, чувствительности и специфичности с 95 % доверительными интервалами.

Несомненно, часть приводимого в материалах и методах текста может быть представлена в виде таблицы (по примеру табл. 1 данной статьи). В разделе результатов при оформлении научной публикации стоит привести в виде рисунка полученное дерево классификации (если это позволяет его размер, оно может быть приведено по примеру рис. 2, 5 или 9 данной статьи). Также раздел результатов может содержать примерно следующий текст.

«В результате построения дерева классификации для определения наличия или отсутствия у пациентов туберкулеза в качестве входных признаков в математическую модель включено три признака: индекс массы тела, наличие частого переохлаждения и табакокурения. Данное дерево классификации состоит из трех уровней и содержит 11 узлов. Точность классификации пациентов с применением данного дерева классификации составила 80,5 (78,9; 81,8) %, чувствительность – 84,2 (80,0; 87,7) %, специфичность – 77,2 (72,8; 81,1) %».

Таким образом, рассмотрены вопросы применения деревьев классификации в медико-биологических исследованиях, а также представлены примеры их построения в статистических пакетах прикладных программ IBM SPSS Statistics и StatSoft Statistica. Применение при анализе данных медико-биологических экспериментов дерева классификации, как одного из легко используемых и интерпретируемых методов многомерного анализа данных, позволит более глубоко изучать закономерности явлений и состояний в области медицины и биологии.

Авторство

Наркевич А. Н. внес существенный вклад в концепцию и дизайн исследования, получение, анализ и интерпретацию данных, подготовил первый вариант статьи, окончательно утвердил присланную в редакцию рукопись; Виноградов К. А. внес существенный вклад в концепцию и дизайн исследования, получение, анализ и интерпретацию данных, существенно переработал первый вариант статьи на предмет важного интеллектуального содержания, окончательно утвердил присланную в редакцию рукопись; Гржибовский А. М. внес существенный вклад в концепцию и дизайн исследования, анализ и интерпретацию данных, существенно переработал первый вариант статьи на предмет важного интеллектуального содержания, окончательно утвердил присланную в редакцию рукопись.

Наркевич Артем Николаевич – ORCID 0000-0002-1489-5058; SPIN 9030-1493
Виноградов Константин Анатольевич – ORCID 0000-0001-6224-5618; SPIN 6924-0110
Гржибовский Андрей Мечиславович – ORCID 0000-0002-5464-0498; SPIN 5118-0081

Список литературы / References

1. *Гржибовский А. М., Иванов С. В., Горбатова М. А.* Однофакторный линейный регрессионный анализ с использованием программного обеспечения Statistica и SPSS // Наука и здравоохранение. 2017. № 2. С. 5–33.

Grjibovski A. M., Ivanov S. V., Gorbatova M. A. One-factor linear regression analysis using software Statistica and SPSS. *Nauka i zdravookhranenie* [Science and health]. 2017, 2, pp. 5-33. [In Russian]

2. *Калагина Л. С., Сморкалова В. М., Зобкова Т. И.* Деревья классификации в прогнозировании исходов гепатита А у детей // Медицинский альманах. 2011. № 4. С. 207–210.

Kalagina L. S., Smorkalova V. M., Zobkova T. I. Classification trees in predicting outcomes of hepatitis A in children. *Meditsinskiy al'manakh* [Medical almanac]. 2011, 4, pp. 207-210. [In Russian]

3. *Калагина Л. С., Сморкалова В. М., Зобкова Т. И.* Математическая модель прогнозирования исходов легкой формы гепатита В у детей // Педиатрия. Журнал им. Г. Н. Сперанского. 2012. Т. 91, № 4. С. 156–159.

Kalagina L. S., Smorkalova V. M., Zobkova T. I. Mathematical model for predicting outcomes of mild hepatitis B in children. *Pediatriya (Pediatriya - Zhurnal im. G. N. Speranskogo)*. 2012, 91 (4), pp. 156-159. [In Russian]

4. *Кондрова Н. С., Зулькарнаев Т. Р., Франц М. В.* Потенциал здоровья работников как компонент человеческого потенциала организации // Гигиена и санитария. 2018. Т. 97, № 2. С. 164–171.

Kondrova N. S., Zul'karnaev T. R., Frants M. V. Potential of workers ' health as a component of the human potential of the organization. *Gigiena i Sanitariya*. 2018, 97 (2), pp. 164-171. [In Russian]

5. *Константинова Е. Д., Вараксин А. Н., Жовнер И. В.* Определение основных факторов риска развития неинфекционных заболеваний: метод деревьев классификации // Гигиена и санитария. 2013. Т. 92, № 5. С. 69–72.

Konstantinova E. D., Varaksin A. N., Zhovner I. V. Determination of the main risk factors for the development of non-communicable diseases: method of classification trees. *Gigiena i Sanitariya*. 2013, 92 (5), pp. 69-72. [In Russian]

6. *Наркевич А. Н., Виноградов К. А.* Настольная книга автора медицинской диссертации: пособие. М.: Инфра-М, 2019. 454 с.

Narkevich A. N., Vinogradov K. A. *A reference book of the author of a medical dissertation: manual*. Moscow, Infra-M Publ., 2019, 454 p. [In Russian]

7. *Наркевич А. Н., Виноградов К. А., Корецкая Н. М., Наркевич А. А.* Использование прогностических математических моделей для выявления больных туберкулезом легких // Туберкулез и болезни легких. 2014. № 9. С. 44–45.

Narkevich A. N., Vinogradov K. A., Koretskaya N. M., Narkevich A. A. Use of predictive mathematical models for detecting patients with pulmonary tuberculosis. *Tuberkulez i bolezni legkikh* [Tuberculosis and lung diseases]. 2014, 9, pp. 44-45. [In Russian]

8. *Фомина Е. Е.* Возможности метода деревьев классификации при обработке социологической информации // Гуманитарный вестник. 2018. № 11. С. 5.

Fomina E. E. Possibilities of the method of classification trees in the processing of sociological information. *Gumanitarnyi vestnik* [Humanities Bulletin]. 2018, 11, p. 5. [In Russian]

9. *Халафян А. А., Виноградов Р. А., Акиньшина В. А., Кошкаров А. А.* Система поддержки принятия решений при выборе тактики коррекции стеноза внутренних сонных артерий // Врач и информационные технологии. 2018. № 2. С. 29–38.

Khalafyan A. A., Vinogradov R. A., Akin'shina V. A., Koshkarov A. A. Decision support System for choosing tactics for correction of internal carotid artery stenosis. *Vrach i informatsionnye tekhnologii* [Doctor and information technology]. 2018, 2, pp. 29-38. [In Russian]

10. *Харевич О. Н., Лаптева И. М., Лаптева Е. А., Королева Е. Г.* Клинические фенотипы тяжелой астмы (по результатам кластерного анализа) // Вестник Санкт-Петербургского университета. Медицина. 2015. № 2. С. 28–39.

Kharevich O. N., Lapteva I. M., Lapteva E. A., Koroleva E. G. Clinical phenotypes of severe asthma (based on the results of cluster analysis). *Vestnik Sankt-Peterburgskogo universiteta. Meditsina* [Bulletin of the Saint Petersburg University. Medicine]. 2015, 2, pp. 28-39. [In Russian]

11. *Шарашова Е. Е., Холматова К. К., Горбатова М. А., Гржибовский А. М.* Применение множественного логистического регрессионного анализа в здравоохранении с использованием пакета статистических программ SPSS // Наука и здравоохранение. 2017. № 4. С. 5–26.

Sharashova E. E., Kholmatova K. K., Gorbatova M. A., Grjibovski A. M. Application of multiple logistic regression analysis in healthcare using a package of statistical programs SPSS. *Nauka i zdravookhranenie* [Science and health]. 2017, 4, pp. 5-26. [In Russian]

12. Bamber J. H., Evans S. A. The value of decision tree analysis in planning anaesthetic care in obstetrics. International *Journal of Obstetric Anesthesia*. 2016. 27, pp. 55-61. DOI: 10.1016/j.ijoa.2016.02.007.

13. Ben-Gal I., Dana A., Shkolnik N., Singer G. Efficient Construction of Decision Trees by the Dual Information Distance Method. *Quality Technology & Quantitative Management*. 2014, 11 (1), pp. 133-147.

14. Bewick V., Cheek L., Ball J. Statistics review 14: Logistic regression. *Critical Care*. 2005, 9 (1), pp. 112-118.

15. Bewick V., Cheek L., Ball J. Statistics review 7: Correlation and regression. *Critical Care*. 2003, 7 (6), pp. 451-459.

16. Deng H., Runger G., Tuv E. Bias of importance measures for multi-valued attributes and solutions. *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*. 2011, pp. 293-300.

17. Distinguin L., Blanchard M., Rouillon I., Parodi M., Loundon N. Pediatric cochlear reimplantation: Decision-tree efficacy. *European Annals of Otorhinolaryngology, Head and Neck Diseases*. 2018, 135 (4), pp. 243-247. DOI: 10.1016/j.anorl.2018.05.002.

18. Garonzik-Wang J. M., Majella Doyle M. B. Decision Tree for Liver Resection for Hepatocellular Carcinoma. *JAMA Surgery*. 2016, 151 (9), pp. 853-854. DOI: 10.1001/jamasurg.2016.1149.

19. Laurent H., Rivest R. L. Constructing Optimal Binary Decision Trees is NP-complete. *Information Processing Letters*. 1976, 5 (1), pp. 15-17. DOI: 10.1016/0020-0190(76)90095-8.

20. Schneider A., Hommel G., Blettner M. Linear Regression Analysis. Part 14 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*. 2010, 107 (44), pp. 776-782.

21. Suzuki S., Ukiya T., Kawauchi Y., Ishii H., Sugihara N. Decision tree analysis for factors associated with dental caries in school-aged children in Japan. *Community Dental Health Journal*. 2018, 35 (4), pp. 247-251. DOI: 10.1922/CDH_4409Suzuki05.

22. Tayefi M., Esmaeili H., Saberi Karimian M., Amirabadi Zadeh A., Ebrahimi M., Safarian M., Nematy M., Parizadeh S. M. R., Ferns G. A., Ghayour-Mobarhan M. The application of a decision tree to establish the parameters associated with hypertension. *Computer Methods and Programs in Biomedicine*. 2017, 139, pp. 83-91. DOI: 10.1016/j.cmpb.2016.10.020.

Контактная информация:

Наркевич Артем Николаевич – доктор медицинских наук, зав. лабораторией медицинской кибернетики и управления в здравоохранении, зав. кафедрой медицинской кибернетики и информатики ФГБОУ ВО «Красноярский государственный медицинский университет им. проф. В. Ф. Войно-Ясенецкого» Минздрава России

Адрес: 660022, Красноярский край, г. Красноярск, ул. Партизана Железняка, д. 1

E-mail: narkevichart@gmail.com