ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 1, С. 107–121

КОМПЬЮТЕРНЫЕ МЕТОДЫ

УДК: 004.852

ИНТЕРПРЕТИРУЕМОСТЬ ОБУЧЕНИЯ В СИСТЕМЕ ОБРАБОТКИ СИГНАЛОВ

© 2024 г. А.А. Докукин[a],*, А.В. Кузнецова[b],, Н.В. Окулов[c],***, О.В. Сенько[a],****, В.Я. Чучупал[a],*******

[a] *Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия*

[b] *Институт биохимической физики им. Н.М. Эмануэля Российской академии наук, Москва, Россия*

[c] *ООО "Правильное Пчеловодство", Пермь, Россия*

** e-mail: dalex@ccas.ru*

*** e-mail: azforus@yandex.ru*

**** e-mail: stereoperm@yandex.ru*

***** e-mail: senkoov@mail.ru*

****** e-mail: v.chuchupal@vail.ru*

Поступила в редакцию 01.08.2023 г.
После доработки 05.09.2023 г.
Принята к публикации 02.10.2023 г.

Рассмотрен программный комплекс, позволяющий генерировать алгоритмы автоматической классификации сигналов. Программный комплекс включает алгоритм, переводящий записи непрерывных сигналов в векторную форму, набор методов машинного обучения, а также средства интеллектуального анализа данных, направленные на достижение прозрачности и интерпретируемости обучения. В основе подхода лежит представление различий между сравниваемыми классами в виде совокупности относительно простых, статистически значимых и интерпретируемых эффектов, которые графически изображаются на двумерных диаграммах. Работа метода иллюстрируется на задаче оценки состояния улья по звуковым сигналам. Программный комплекс может быть использован при решении прикладных задач автоматической диагностики и анализа данных.

Ключевые слова: интерпретируемость, машинное обучение, обработка сигналов, программный пакет, звуки пчел, пчелиная матка

DOI: 10.31857/S0002338824010085, **EDN:** ITJGID

INTERPRETABILITY OF LEARNING IN A SIGNAL PROCESSING SYSTEM

© 2024 A. A. Dokukin[a], *, A. V. Kuznetsova[b], N.V. Okulov[c], O.V.Sen'ko[a], V. Ya. Chuchupal[a]

[a] *FRC CSC RAS, Moscow, Russia*

[b] *Emanuel Institute of Biochemical Physics of RAS, Moscow, Russia*

[c] *"Pravilnoe pchelovodstvo", Perm, Russia, 614002*

** e-mail: dalex@ccas.ru*

The paper presents a software package that allows one to generate algorithms for automatic classification of signals. The software package includes an algorithm that converts records of continuous signals into vector form, a set of machine learning methods, as well as data mining tools aimed at achieving transparency and interpretability of learning. The approach is based on the presentation of differences between compared classes as a set of relatively simple, statistically significant and interpretable effects, which are graphically represented on two-dimensional diagrams. The performance of the method is illustrated on the problem of assessing the state of the hive by sound signals. The software package can be used in solving applied problems of automatic diagnostics and data analysis.

Keywords: interpretability, machine learning, signal processing, software package, bees sounds, queen bee

Введение. Задачи компьютерной диагностики, когда требуется автоматически предсказать состояния изучаемых объектов по параметрам, наблюдаемым от их сигналов, все чаще возникают в различных областях человеческой деятельности. Наиболее эффективным способом решения таких задач является применение технологий машинного обучения.

В последние годы в этой области активно используются технологии машинного обучения, основанные на сверточных нейронных сетях [1, 2]. Отличительная особенность сверточных нейронных сетей — возможность применения их непосредственно к изображениям или сигналам.

Сверточные сети демонстрируют высокую эффективность при решении многих прикладных задач. Вместе с тем успех обычно достигается при решении задач, для которых существуют обучающие выборки значительного размера. Иными словами, должны существовать размеченные коллекции из многих тысяч сигналов. Разметка, в частности, предполагает, что каждому сигналу должен быть сопоставлен известный диагноз для соответствующего этому сигналу объекта. Сбор и разметка таких больших размеченных корпусов данных требует значительных затрат времени и средств, поэтому для многих, в частности, новых прикладных задач они отсутствуют.

Наряду с нейросетевыми методами в современном машинном обучении существует целый ряд альтернативных технологий, которые предназначены для решения задач автоматической классификации объектов, описания которых задаются векторами числовых переменных (признаков).

К числу таких технологий можно отнести получившие широкое распространение статистические методы, включая дискриминантный анализ Фишера, байесовские методы, логистическую регрессию и др.

Наряду со статистическими методами активно развивались эффективные альтернативные технологии, включая метод опорных векторов [3, 4], случайные решающие леса [3, 5]; леса, основанные на адаптивном или градиентном бустинге [3, 6—8]; различные методы, базирующиеся на принятии коллективных решений по наборам закономерностей [9]. Краткое описание некоторых из перечисленных технологий представлено в разд. 1. Перечисленные методы нередко превосходят многослойные нейросетевые архитектуры по обобщающей способности при относительно небольшом объеме обучающих данных.

Следует отметить продолжающееся совершенствование альтернативных технологий. В связи с этим представляются актуальными разработка и развитие методов и программных систем компьютерной диагностики, основанных на переводе сигналов во многомерные векторы признаков с дальнейшим использованием перечисленных выше методов автоматической классификации для генерации алгоритмов диагностики [10].

При этом для генерации векторных описаний целесообразно применять хорошо разработанные методы спектрального анализа, а также вейвлет-анализа.

Другим существенным недостатком большинства методов современного машинного обучения является непрозрачность полученных с их помощью алгоритмов автоматической классификации в смысле сложности понимания пользователем их работы, а также процесса обучения [11].

Одним из возможных способов повышения прозрачности служит дополнительное использование методов интеллектуального анализа данных, позволяющих представить работу сложного многофакторного алгоритма в виде совокупности относительно простых, достоверных и интерпретируемых эффектов.

Такими возможностями обладает представленный в разд. 2 метод оптимальных достоверных разбиений (ОДР), позволяющий находить разнообразные нелинейные взаимосвязи с оценкой их статистической значимости [12, 13]. Предполагается, что высокая статистическая значимость может обеспечить достоверность эффекта. Для оценки статистической значимости выявленных закономерностей применяется комбинация технологии перестановочных тестов и принципа бритвы Оккама. Перестановочные тесты во многих случаях обладают несомненными преимуществами по сравнению с традиционными параметрическими и непараметрическим критериями и прочно вошли в современный статистический анализ [14, 15, 16]. Принцип бритвы Оккама используется для оптимизации сложности моделей. В этом смысле цели применения совпадают с целями известных информационных критериев, включая информационный критерий Акаике [17], байесовский информационный критерий Шварца [18], принцип Риссанена [19]. Однако недостатком перечисленных критериев является невозможность по отдельности оценить статистическую значимость каждого элемента модели, что осуществляется с помощью рассматриваемого варианта принципа бритвы Оккама, ранее предложенного в работах [13, 20].

Использование принципа бритвы Оккама в машинном обучении подвергалось критике в том смысле, что более сложные модели позволяют нередко достигать более высокой обобщающей способности [21].

Основные цели рассматриваемого варианта принципа — резкое сокращение множества генерируемых моделей с сохранением только значимых эффектов, что существенно упрощает и облегчает анализ пользователем выявленных закономерностей.

Для решения задач компьютерной диагностики была разработана программная система, состоящая из пакета прикладных программ автоматической классификации и интеллектуального анализа данных Data Master Azforus и программы, генерирующей векторные признаковые описания. Пакет Data Master содержит набор методов автоматической классификации, включая методы, перечисленные в разд. 1; методы кластерного анализа; методы статистического анализа, включая метод ОДР.

В статье представлен пример применения для диагностики наличия матки в улье разработанной системы и подхода, основанного на анализе звуковых сигналов, производимых пчелами. Ранее приведенные подходы использовались для решения задачи автоматической диагностики деменции по данным электроэнцефалографии [22].

1. Методы автоматической классификации. Рассмотрим следующее определение задачи автоматической классификации.

Пусть даны некоторые объекты, которые можно описать вектором в признаковом пространстве (пространстве объясняющих переменных) $x \in \left\{R^n\right\}$.

Дополнительно у объекта есть выделенный, целевой, признак (целевая переменная) y, которая известна только для части объектов.

Множество признаковых описаний с известной целевой переменной $\widetilde{S_0} = \left(y_1, x_1\right), \ldots, \left(y_m, x_m\right)$ называется обучающей выборкой.

Необходимо по выборке $\widetilde{S_0}$ построить алгоритм A, который бы для произвольного объекта, в том числе ранее неизвестного, по его признаковому описанию x вычислял бы прогноз значения $y = A(x)$.

При этом если $y \in \left\{\overline{1, K}\right\}$, то задача называется задачей классификации. И этот случай нам будет наиболее интересен далее. Если же $y \in R$, то задача называется задачей регрессии. Перейдем к рассмотрению методов классификации.

1.1. Метод опорных векторов. Метод опорных векторов изначально основывался на принципе максимального зазора, который было предложено использовать для построения оптимальной линейной разделяющей поверхности в задачах распознавания с двумя линейно разделимыми классами. Было показано, что указанная задача сводится к известной математической задаче квадратичного программирования. Избежать требования априорной линейной разделимости и добиться таким образом универсальной применимости метода позволила модифицированная постановка задачи с помощью дополнительных параметров, “смягчающих” ограничения, которые необходимы для обеспечения линейного разделения классов. Было замечено, что при обучении и распознавании новых объектов метод опорных векторов применяет только взаимные скалярные произведения векторных описаний объектов обучающей выборки или векторных описаний новых распознаваемых объектов с векторными описаниями объектов обучающей выборки.

Данное свойство метода опорных векторов позволило разработать его дополнительную модификацию, основанную на использовании так называемого “ядерного” трюка, состоящего в применении вместо скалярных произведений других скалярных функций от пары векторов.

Наиболее успешной заменой оказался гауссиан, вычисляемый по формуле

$$\rho_G\left(x_1, x_2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}},$$

где x_1, x_2 — признаковые описания объектов.

Функция $\rho_G\left(x_1, x_2\right)$ имеет смысл ядерной функции, используемой в непараметрической статистике, а параметр σ имеет смысл размера ядра. При больших размерах ядра вычисление оценки за целевой класс для нового классифицируемого объекта с описанием x производится по всей обучающей выборке. При малых размерах ядра вычисление оценки за целевой

класс для нового классифицируемого объекта производится по объектам обучающей выборке с описаниями из окрестности x в многомерном признаковом пространстве.

Эффективность метода опорных векторов доказана практикой его применения при решении разнообразных прикладных задач. Недостатком метода является отсутствие в нем встроенных процедур отбора признаков. Поэтому при использовании метода опорных векторов целесообразно проводить предварительный отбор признаков.

1.2. Бинарные решающие деревья. Одним из популярных и информативных непараметрических методов является метод бинарных регрессионных деревьев. Бинарное решающее дерево представляет собой иерархически организованную систему вопросов, касающихся описывающих объекты признаков и подразумевающих ответ в форме Да / Нет. При этом вопрос, задаваемый на следующем уровне иерархии, зависит от ответа, полученного на предыдущем уровне. Преимуществом метода решающих деревьев является его прозрачность для пользователя. Процедура вычисления прогноза наглядно представлена на диаграмме и понятна для пользователя. Недостатком метода решающих деревьев является недостаточно высокая обобщающая способность.

1.3. Метод случайных решающих лесов. Обобщающая способность может быть повышена с помощью метода случайного решающего леса. Ансамбль решающих деревьев $T_1,\ldots,T_N$ строится по выборкам $\widetilde{S_1^r},\ldots,\widetilde{S_N^r}$, которые представляют собой выборки с возвращениями из $\widetilde{S_0}$. При построении дерева T_j используется случайное подмножество исходного набора признаков, а также простое коллективное решение: объект относится в тот класс, куда его отнесло большинство деревьев. Применение случайного решающего леса в большинстве случаев позволяет существенно снизить ошибку.

Вместе с тем теряется присущая методу решающих деревьев прозрачность.

1.4. Градиентный бустинг. Еще более высокие результаты показывает метод градиентного бустинга, в котором процедура построения ансамбля регрессионных деревьев в определенной степени аналогична использованию метода градиентного спуска в задачах многомерной оптимизации.

Предположим, что качество аппроксимации описывается с помощью функции потерь $L\left(A,\widetilde{S_0}\right)$, причем она имеет вид $L\left(A,\widetilde{S_0}\right)=F\left(y_1,y_1',\ldots,y_m,y_m'\right)$, *где* $y_i'=A\left(x_i\right)$ — прогноз вероятности принадлежности i-го объекта классу y_i, $i=\overline{1,m}$ (в случай регрессии — непосредственно прогноз величины y_i).

Вычислив градиент функции потерь по вектору прогнозов

$$\nabla F\left(y_1,y_1',\ldots,y_m,y_m'\right)=\left\{\frac{\partial F}{\partial y_1'},\ldots,\frac{\partial F}{\partial y_m'}\right\}$$

для вектора прогнозов на $(k-1)$-м шаге $\left\{A_{k-1}\left(x_1\right),\ldots,A_{k-1}\left(x_m\right)\right\}$, можно скорректировать алгоритм на k-м шаге с помощью метода градиентного спуска:

$$A_k\left(x_j\right)=A_{k-1}\left(x_j\right)-\eta\left.\frac{\partial F}{\partial y_j'}\right|_{A_{k-1}\left(x_j\right)}.$$

К сожалению, эта формула работает только на обучающей выборке. Для нового классифицируемого объекта x последнее слагаемое не определено. И даже если формально включить в функционал F дополнительные переменные с индексами $m+1$, для x неизвестно настоящее значение целевой переменной y. Однако оно известно в точках $x_1,\ldots,x_m$ и может быть спрогнозировано как решение задачи регрессии с помощью дерева T_k, обученного по выборке

$$\left(\left.\frac{\partial F}{\partial y_1'}\right|_{A_{k-1}(x_1)},x_1\right),\ldots,\left(\left.\frac{\partial F}{\partial y_m'}\right|_{A_{k-1}(x_m)},x_m\right).$$

Таким образом на шаге k в ансамбль добавляется дерево T_k.

Для вычисления коллективного прогноза в методе градиентный бустинг используется линейная комбинация прогнозов, полученных с помощью алгоритмов ансамбля. В качестве коэффициентов линейной комбинации при этом выступают оптимальные величины шага градиентного спуска.

1.5. Метод статистически взвешенных синдромов. В методе статистически взвешенных синдромов на этапе обучения ищутся оптимальные разбиения интервалов значений отдельных признаков и совместных областей значений пар признаков в рамках семейств разбиений I—IV (рис. 1). Классификация производится с использованием оценок вероятности целевого класса K внутри синдромов — областей найденных оптимальных разбиений.

Предположим, что описание нового классифицируемого объекта x принадлежит пересечению синдромов $q_1, \ldots, q_r$. Тогда оценка вероятности $P(K|x)$ вычисляется как отношение

$$P(K|x) = \frac{\sum_{i=1}^{r} w_i \nu_i}{\sum_{i=1}^{r} w_i},$$

где ν_i — доля объектов $\widetilde{S_0}$ из K внутри синдрома q_i, w_i — вес синдрома. Для вычисления весов с помощью метода максимального правдоподобия получена простая аналитическая формула.

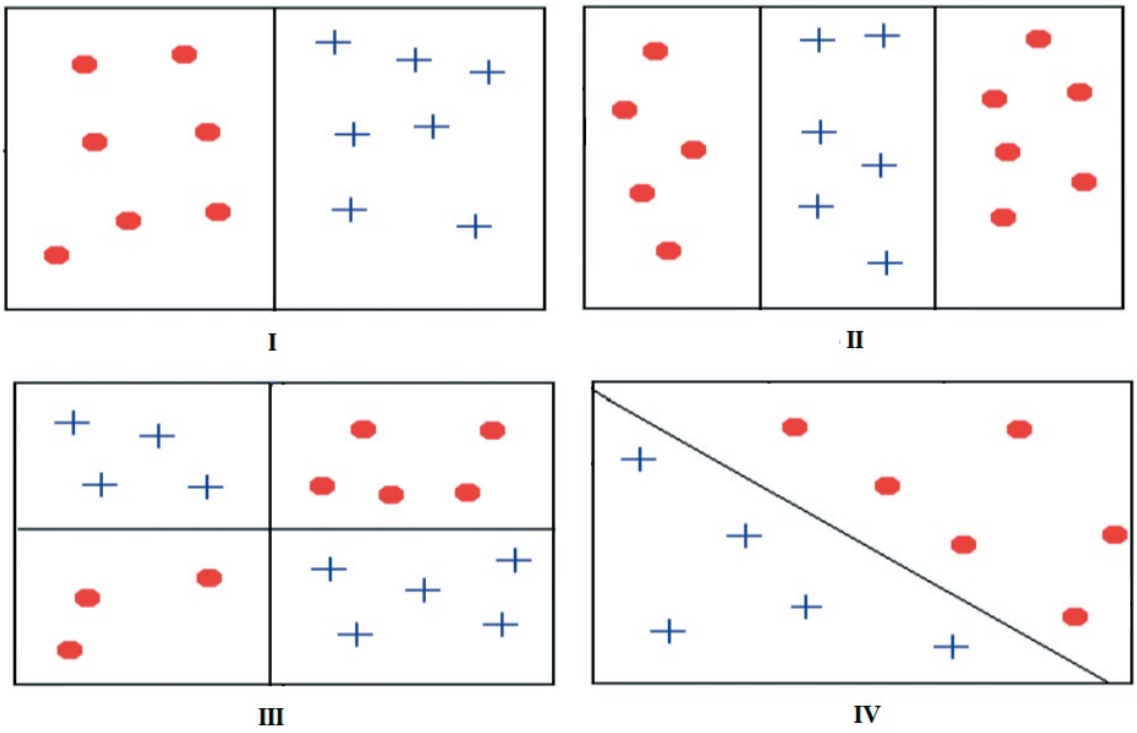

Рис. 1. Семейства разбиений I—IV

2. Метод ОДР. Целью метода ОДР является исследование зависимости бинарной переменной $y \in \{0,1\}$ от объясняющих переменных $x = \{X_1,\ldots,X_n\}$.

Метод ОДР включает поиск оптимальных разбиений интервалов значений отдельных переменных и двумерных областей значений пар переменных по обучающей выборке стандартного вида $\widetilde{S_0} = \{(y_1,x_1),\ldots,(y_m,x_m)\}$. Границы разбиений подбираются таким образом, чтобы достичь наилучшего разделения групп с различными значениями целевой переменной Y. При этом качество разделения для разбиения признакового пространства $r = q_1,\ldots,q_k$ описывается функционалом

$$Q\left(r,\widetilde{S_0}\right)=\frac{1}{\nu_0\left(1-\nu_0\right)}\sum_{i=1}^{k}m_i\left(\nu_0-\nu_i\right)^2, \tag{2.1}$$

где ν_0 — доля объектов с $y_j=1$ во всей обучающей выборке $\widetilde{S_0}$, ν_i — доля объектов с $y_j=1$ внутри области q_i, m_i — полное число объектов обучающей выборки внутри области q_i. Оптимальные разбиения ищутся внутри четырех семейств, представленных на рис. 1, где $Y=0$ — кружки, $Y=1$ — крестики.

Ось абсцисс на диаграммах семейств I и II соответствует объясняющей переменной X. Ось абсцисс на диаграммах семейств III и IV отвечает объясняющей переменной X', а ось ординат — объясняющей переменной X''. Наилучшее разбиение, найденное по выборке $\widetilde{S}$, далее будет обозначаться через $r_b(\widetilde{S})$.

Разбиения из семейств I и II соответствуют закономерностям, связывающим целевую переменную Y с одной объясняющей переменной X. Разбиения из семейств III и IV отвечают закономерностям, связывающим целевую переменную Y с парой переменных X' и X''. Простейшее одномерное семейство I включает все разбиения интервалов значений единичных переменных с одной граничной точкой. Более сложное семейство II включает все разбиения интервалов значений единичных переменных с двумя граничными точками. Двумерное семейство III содержит двумерные разбиения с двумя границами, параллельными координатным осям, двумерное семейство IV — все разбиения с одной прямой граничной линией, произвольно ориентированной относительно координатных осей.

Статистическая значимость закономерностей, описываемых оптимальными разбиениями, может быть оценена с помощью перестановочного теста. Пусть $\widetilde{f}=\{f_1,\ldots,f_N\}$ — множество всевозможных взаимно различимых перестановок чисел из множества $\{1,\ m\}$. Предположим, что $\widetilde{S}_p(f)$ является выборкой, полученной из исходной выборки $\widetilde{S}_0$ с помощью случайных перестановок величин $y_1,\ \ldots\ ,y_m$: $\widetilde{S}_p(f)=\{[y_{f(1)},\ x_1],\ \ldots\ ,[y_{f(m)},\ x_m]\}$.

Множество S_b включает всевозможные различимые между собой выборки, удовлетворяющие следующим двум условиям:

а) эмпирическое распределение Y в $\widetilde{S}$ совпадает с эмпирическим распределением Y в исходной выборке $\widetilde{S_0}$;

б) описания в $\widetilde{S}$ полностью совпадают с x-описаниями в $\widetilde{S_0}$.

Пусть $P_{pv}\left(\widetilde{S}\right)$ — предикат:

$$F\left[M_b\left(\widetilde{S}\right),\widetilde{S}\right]\geq F\left[M_b\left(\widetilde{S_0}\right),\widetilde{S_0}\right],$$

где F — функционал, характеризующий качество аппроксимации Y от $X_1,\ldots,X_n$, $M_b\left(\tilde{S}\right)$ — оптимальная модель, найденная по выборке $\widetilde{S}$ внутри множества моделей M.

В качестве меры значимости существования закономерности, описываемой оптимальным разбиением r_b, естественно использовать вероятность истинности предиката $P_{pv}\left(\widetilde{S}\right)$ при соблюдении нулевой гипотезы H_0: Y не зависит от $X_1,\ldots,X_n$ и наблюдения являются независимыми и одинаково распределенными, соблюдаются условия а) и б).

Лемма 1. Отношение

$$\frac{\left|\left\{\widetilde{S}\mid P_{pv}\left(\widetilde{S}\right)=1,\widetilde{S}\in\widetilde{S_b}\right\}\right|}{\left|\widetilde{S_b}\right|}$$

равно вероятности выполнения равенства $P_{pv}\left(\widetilde{S}\right)=1$ при справедливости нулевой гипотезы H_0.

Доказательство. Утверждение следует из равновероятности всех выборок из множества $\widetilde{S_b}$, что непосредственно следует из нулевой гипотезы. Действительно, из условия независимости Y от $X_1,\ldots,X_n$ и взаимной независимости наблюдений и соблюдения условий а) и б) вытекает, что вероятность произвольной выборки $\widetilde{S}=\left\{\left(y_1,x_1\right),\ldots,\left(y_m,x_m\right)\right\}$ равна $\prod_{j=1}^{m}P\left(Y=y_j\right)P\left(X=x_j\right)$. Лемма 1 доказана.

Лемма 2. Каждая выборка из $\widetilde{S_b}$ может быть получена $\widetilde{S_0}$ с помощью одинакового числа перестановок.

Доказательство. Действительно, произвольная выборка $\widetilde{S} \in \widetilde{S_b}$ может быть получена из $\widetilde{S_0}$ с помощью последовательного выполнения двух перестановок: произвольной перестановки $f_0 \in \widetilde{f}$, переводящей $\widetilde{S_0}$ в $\widetilde{S}$, и перестановки $f_0 \in \widetilde{f}$, переводящей $\widetilde{S}$ в неотличимую от нее выборку. Пусть Y принимает в $\widetilde{S_0}$ значения $\left\{y_1^d,\ldots,y_k^d\right\}$. Число перестановок f' одинаково всех $\widetilde{S} \in \widetilde{S_b}$ и равно $\prod_{i=1}^{k} m_i!$, где m_i — число объектов в $\widetilde{S_0}$, для которых Y принимает значение y_i^d. Лемма 2 доказана.

Из лемм 1 и 2 легко вытекает следующее утверждение.

Теорема 1. Отношение

$$\frac{\left|\{f \mid P_{pv}\left(\widetilde{S_p}p(f)\right) = 1, f \in \widetilde{f}\}\right|}{\left|\widetilde{f}\right|} \tag{2.2}$$

равно вероятности выполнения равенства $P_{pv}\left(\widetilde{S}\right) = 1$ при справедливости нулевой гипотезы H_0. Что и требовалось доказать.

Отношение (2.2) по сути является значением при использовании в качестве статистики критерия оптимального значения функционала F, т. е. критерия $F\left[M_b\left(\tilde{S}\right),\tilde{S}\right]$. Отношение не может быть точно рассчитано даже при относительно небольших объемах $\widetilde{S_0}$ из-за необходимости повторения при всех перестановках, общее количество которых равно $m!$. Однако для приближенного оценивания может быть применен подход, основанный на использовании датчиков случайных чисел. Пусть $\{\widetilde{f_g} = f_{(j)} | j = \overline{1,N_g}\}$ представляет собой набор перестановок чисел из множества $\left\{\overline{1,m}\right\}$, сгенерированных с помощью генератора случайных чисел. Вычисление p-значений тогда приближенно производится по формуле

$$p = \frac{|\{f \mid P_{pv}\left(\widetilde{S_p}(f)\right) = 1, f \in \widetilde{f_g}\}|}{\left|\widetilde{f_g}\right|} \tag{2.3}$$

Преимуществом перестановочного теста является отсутствие априорных предположений о вероятностных распределениях и возможность использования при произвольных размерах выборок. При применении метода ОДР (поиска значимых одномерных закономерностей, связывающих Y с признаком X_1) в качестве статистики критерия используется $Q\left[r_b\left(\widetilde{S_0}\right),\widetilde{S_0}\right]$, где оптимальное разбиение $r_b\left(\widetilde{S_0}\right)$ ищется внутри семейств I или II. При поиске значимых двумерных закономерностей, связывающих Y с признаками X_1 и X_2, в качестве статистики критерия выбирается $Q\left[r_b\left(\widetilde{S_0}\right),\widetilde{S_0}\right]$, где оптимальное разбиение $r_b\left(\widetilde{S_0}\right)$ ищется внутри семейств III или IV.

Для поиска оптимальных одномерных разбиений из семейства I значение функционала Q вычисляется при всевозможных границах, находящихся посередине между двумя ближайшими отличными друг от друга значениями признака X_1. Примеры одномерных закономерностей приведены на рис. 2 и рис. 3. Каждой одномерной закономерности из семейства I ставится в соответствие одно p-значение, вычисляемое по формуле (2.3).

Эксперименты с искусственно сгенерированными данными показали, что поиск закономерностей с оценкой их статистической значимости по формуле (2.1) может быть эффективный только в случае применения простейших одномерных моделей из семейства I. При использовании более сложных моделей из семейств I—IV такой поиск обычно приводит к появлению в выходном наборе частично ложных закономерностей, не предусмотренных сценарием эксперимента. Например, найденная двумерная закономерность из семейства III считается частично ложной, если на самом деле с целевой переменной связана только одна из двух объясняющих переменных. Вторая объясняющая

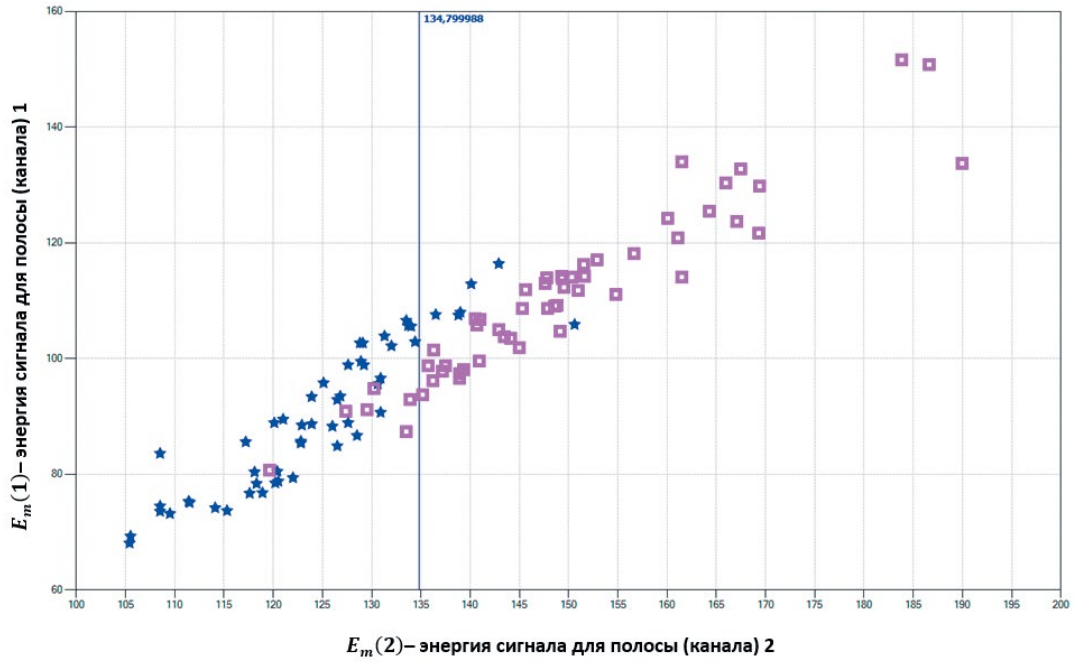

Рис. 2. Связь показателя $\bar{E}(2)$ с наличием матки в улье

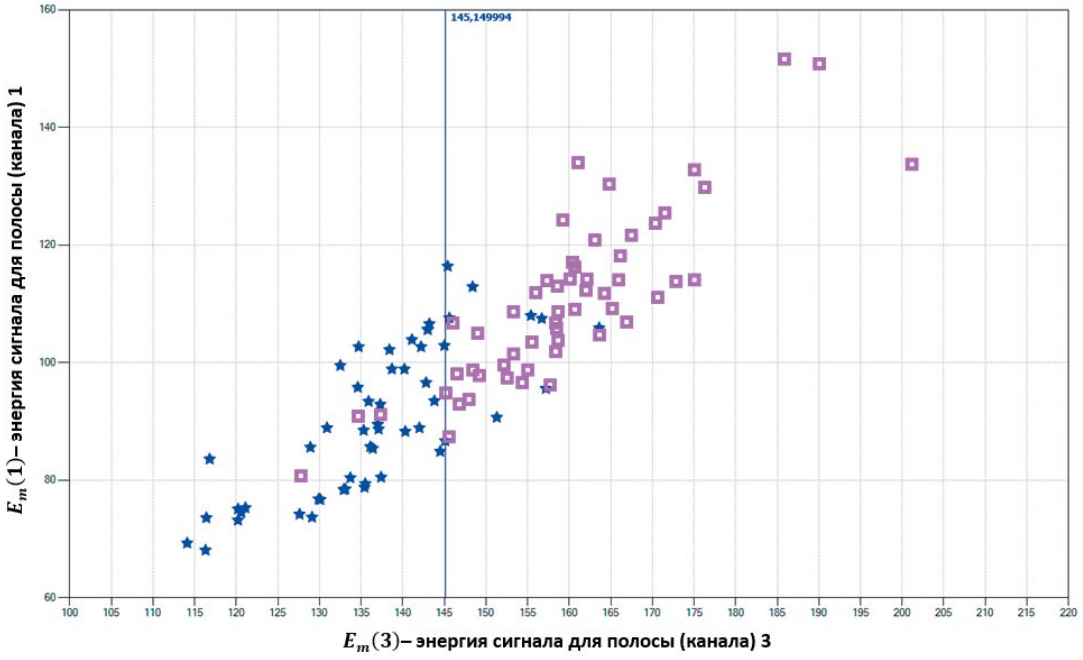

Рис. 3. Связь показателя $\bar{E}(3)$ с наличием матки в улье

переменная при этом избыточна и вхождение ее в найденную закономерность является чисто случайным.

Эксперименты на искусственно сгенерированных данных показали, что число частично ложных закономерностей может значительно превышать число закономерностей, предусмотренных сценарием экспериментов, что делает анализ с помощью более сложных моделей в значительной степени бессмысленным. Для преодоления указанной проблемы

используется принцип бритвы Оккама в формулировке, ориентированной на статистический анализ данных: более сложная модель должна применяться только в том случае, если она позволяет отвергнуть нулевую гипотезу об исчерпывающем описании зависимости простой моделью.

Пусть $\widetilde{M_s}$ — множество простых моделей, $\widetilde{M_c}$ — множество более сложных моделей. Используемый вариант бритвы Оккама сводится к более простому принципу: более сложная модель должна быть применена тогда и только тогда, когда в множестве $\widetilde{M_s}$ не существует модели, исчерпывающей зависимость Y от переменных $X_1,\ldots,X_n$.

Дополнительно предполагается, что нулевая гипотеза об исчерпывающем описании зависимости Y от переменных $X_1,\ldots,X_n$ моделью $M \in \widetilde{M_s}$ может быть отвергнута с помощью более сложной модели из $\widetilde{M_c}$.

Рассмотрим процедуру бритвы Оккама применительно к проверке предположения об исчерпывающем описании данных моделью из семейства I с помощью более сложной модели из одного из семейств III или IV.

Пусть простейшая одномерная модель из семейства I задается разбиением с граничной точкой δ_1 для интервала значений объясняющей переменной X_1, который обозначим (a_l,a_r).

Мы пытаемся проверить, позволит ли использование дополнительной переменной X_2 и моделей из одного из семейств III или IV более точно описать зависимость. Для этого проверяем нулевую гипотезу, что Y не зависит от переменных X_1, X_2 внутри интервалов (a_l,δ_1) и (δ_1,a_r).

В этом случае применяется модифицированная нулевая гипотеза H_0^1: Y не зависит от X_1, X_2 внутри интервалов (a_l,δ_1) и (δ_1,a_r) и наблюдения являются независимыми и одинаково распределенными, условия а) и б) соблюдаются слева и справа от границы.

В качестве меры значимости двумерной закономерности, описываемой оптимальным разбиением r_b, естественно использовать вероятность истинности предиката $P_{pv}(\widetilde{S})$:

$$Q\left[r_b(\tilde{S}),\tilde{S}\right] \geq Q\left[r_b(\widetilde{S_0}),\widetilde{S_0}\right]$$

при соблюдении нулевой гипотезы H_0.

Обозначим через J_l множество номеров объектов $\widetilde{S_0}$, для которых $X_1 < \delta$, а через J_r — множество номеров объектов $\widetilde{S_0}$, для которых $X_1 \geq \delta$. Пусть $\widetilde{f_l}$ — множество всевозможных перестановок J_l, $\widetilde{f_r}$ — множество всевозможных перестановок J_r, $\widetilde{f_\delta}$ — декартово произведение $\widetilde{f_l}$ и $\widetilde{f_r}$, $\widetilde{f_\delta} = \widetilde{f_l} \times \widetilde{f_r}$.

Теорема 2. Отношение

$$\frac{\left|\{f \mid P_{pv}\left(\widetilde{S_p}(f)\right) = 1, f \in \widetilde{f_\delta}\}\right|}{\left|\widetilde{f_\delta}\right|} \tag{2.4}$$

равно вероятности выполнения равенства $P_{pv}(\tilde{S}) = 1$ при справедливости нулевой гипотезы H_0.

Доказательство следует из возможности применения лемм 1 и 2 к выборкам слева и справа от границы δ_1.

Пусть $\{\widetilde{f_{g\delta}} = f(j) \mid j = \overline{1,N_g}\}$ является набором перестановок из множества $\widetilde{f_\delta}$, сгенерированных с помощью генератора случайных чисел. Вычисление p-значений тогда приближенно производится по формуле

$$p = \frac{\left|\{f \mid P_{pv}\left(\widetilde{S_p}(f)\right) = 1, f \in \widetilde{f_{g\delta}}\}\right|}{\left|\widetilde{f_{g\delta}}\right|} \tag{2.5}$$

Для того чтобы оценить значимость двумерной закономерности, необходимо одновременно с нулевой гипотезой H_0^1 опровергнуть нулевую гипотезы H_0^2 об исчерпывающем описании зависимости Y от X_1, X_2 одномерной моделью из семейства I с граничной точкой δ_2. Таким образом, значимость двумерной закономерности оценивается с помощью двух p-значений: при низких p_1 отвергается гипотеза H_0^1 и подтверждается вклад в двумерную закономерность X_2; при низких p_2 отвергается гипотеза H_0^2 и подтверждается вклад в двумерную закономерность X_2.

Для поиска оптимальных одномерных разбиений из семейства III и IV значение функционала Q рассчитывается при всевозможных границах, находящихся посередине между двумя ближайшими отличными друг от друга значениями признаков X_1 или X_2.

Примеры двумерных закономерностей приведены на рис. 4. Каждой двумерной закономерности из семейств III или IV ставится в соответствие два значения, вычисляемые по формуле (2.5).

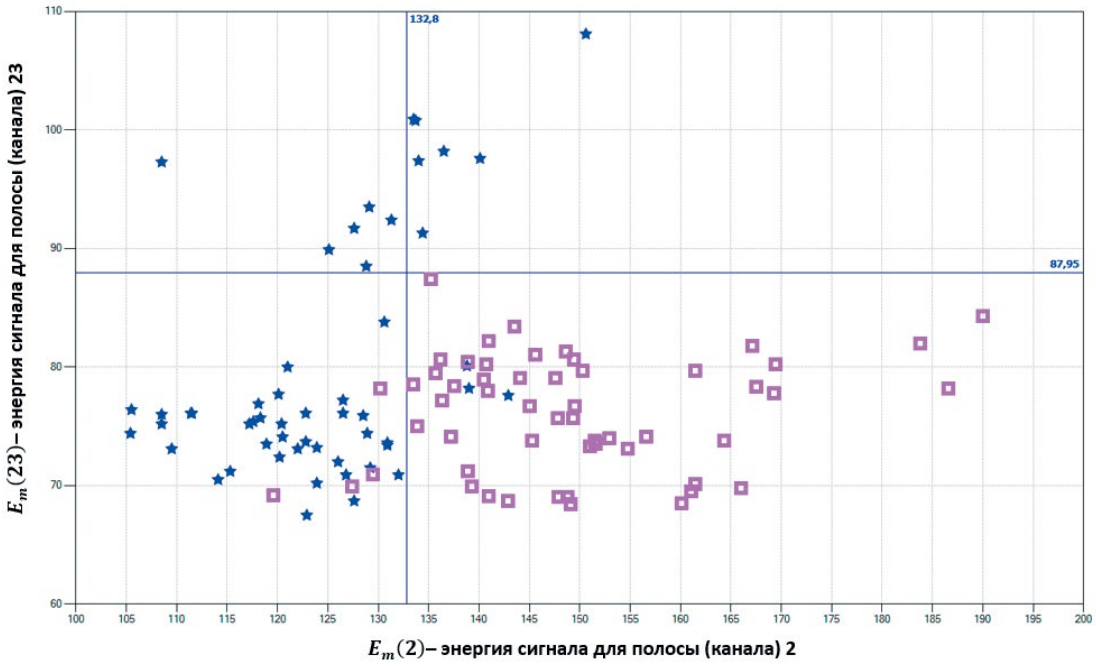

Рис. 4. Связь наличия матки в улье со спектральными показателями $\bar{E}(2)$ и $\bar{E}(23)$

3. Распознавание состояния объекта по временным сигналам. Рассматриваемый программный комплекс может применяться для распознавания состояния разнообразных биологических или технических систем по генерируемым ими сигналам различной природы. В частности, в задачах медицинской диагностики с помощью электрокардиограмм, электроэнцефалограмм, фотоплетизмограмм и т.д.; в задачах технической диагностики по акустическим и иным сигналам. В данной работе в качестве практического приложения представленных методов была использована задача автоматической классификации ульев по признаку наличия или отсутствия в них пчелиной матки.

В иерархии многих видов социальных насекомых, в частности пчел, центральное место занимает матка, которая производит на свет всех членов сообщества. При ее отсутствии сообщество быстро погибает. В случае пчел, если это произошло во второй половине лета, потеря матки ведет к гарантированному вымиранию пчелиной семьи. Поэтому важно вовремя уметь находить улья, в которых пропала матка. Для промышленных пасек с десятками и сотнями ульев — это трудоемкая процедура.

Существенно ее удешевить и облегчить в целом иные процедуры оценки состояния ульев может использование компьютерных методов, в частности интеллектуальный анализ звуковых сигналов, издаваемых пчелами.

Роль звуков в организации “общения” между пчелами подтверждается многими исследователями [23—25].

Данные для анализа собирались с помощью приложения для Android на смартфонах Dexp. Всего для анализа использовались 108 звуковых профилей продолжительностью 3 мин, включая 55 профилей, соответствующих ульям без матки, и 53 профиля, соответствующих ульям с маткой. В экспериментах применялось признаковое описание в амплитудо-частотной области. Для его получения в амплитудо-частотной области исходные аудиозаписи конвертировались в формат WAV с частотой дискретизации 8 кГц и разрядностью 16 бит.

Данные в каждой записи s_t, $t = \overline{0,T}$, разбивались на перекрывающиеся сегменты конечной длины N: $x_m(t) = s(ml + t)$, $0 \le t < N$, $m = \overline{0,M-1}$; $M = (T - N)/l + 1$, где N, m, l — соответственно длина, индекс и смещение сегментов в сэмплах.

В экспериментах использовались $N = 256$, $l = 128$, что соответствовало размеру сегментов анализа 32 мс.

Сэмплы каждого m-го сегмента взвешивались весовым окном Хемминга:

$$x_m(t) = x_m(t)\left[0.54 + 0.46\cos\left(\frac{2\pi t}{N-1}\right)\right], t = \overline{0,N-1}.$$

Для них вычислялись коэффициенты преобразования Фурье:

$$X_m(k) = FFT(x_m(t)), \mathrm{k} = \overline{0,N-1} \text{ и их амплитуды } |X(k)|, k = \overline{0,\mathrm{N}/2}.$$

Вычислялось значение энергии сигнала в 30 перекрывающихся частотных полосах, равномерно разнесенных по шкале мелов в диапазоне от 40 Гц до 3.2 кГц. Центральные частоты полос приведены в табл. 1. Для этого суммировались и логарифмировались взвешенные значения амплитуд $|X(k)|$, где k изменялся от индекса, соответствующего центру предыдущей полосы, до индекса центра следующей:

$$E_m(n) = \log\left(\sum_{k=d(n-1)}^{d(n+1)} w_n(k)|X_m(k)|\right),$$

Таблица 1. Различия между группами ульев с маткой и без матки с учетом оптимальных порогов

n	Граница для $\bar{E}(n)$	$p<$		n	Граница для $\bar{E}(n)$	$p<$
3	145.2	0.0004		14	88.55	0.0007
2	134.8	0.0004		16	93	0.001
6	150.4	0.0004		18	87.95	0.002
4	146.1	0.0004		20	89.45	0.002
5	149.5	0.0005		21	96.9	0.002
1	90.8	0.0005		23	87.15	0.002
7	154.75	0.0005		17	103.9	0.001
8	93.25	0.0005		22	109.55	0.002
10	125.3	0.0005		24	87.4	0.004
9	87.3	0.0005		29	88.55	0.007
11	100.25	0.0005		25	81.3	0.006
12	85.9	0.0005		26	83.2	0.006
13	133.7	0.0004		27	83	0.007
15	144.75	0.001		28	82.15	0.007
19	87.9	0.002		30	82.95	0.031

где n — номер полосы, $d(n)$ — индекс коэффициента преобразования Фурье, который соответствует центру n-й частотной полосы, а $w_n(k)$ — вес амплитуды k-го коэффициента в n-й полосе, $n = \overline{1,30}$, $d(0) = 0$, $d(31) = 104$. Веса $w_n(k)$ неотрицательные, с суммой, равной единице, их значения обратно пропорциональны $|k - d(n)|$. Максимальное значение веса соответствовало центру текущей частотной полосы, а минимальное, равное 0, границам.

Дополнительно к вектору из логарифмов энергии сигнала в частотных полосах вычисляется вектор разности между соседними сегментами логарифмов энергии сигнала в частотных полосах в качестве аппроксимации первой производной значений параметров по времени:

$$\Delta E_m(n) = E_{m+1}(n) - E_m(n), n = \overline{1,30}; m = \overline{0, M-2}.$$

Таким образом, каждый m-й сегмент аудиозаписи описан 60-мерным вектором, в котором содержатся 30 логарифмов энергий сигнала в частотных полосах и столько же их первых разностей, а вся запись представляется как последовательность из таких векторов, длина которой зависит от длительности аудиозаписи.

Поскольку сигналы были фактически квазистационарны, при обучении моделей и классификации звуков ульев вместо последовательностей векторных значений параметров $E(m)$ и $\Delta E(m)$, $m = \overline{0, M-1}$, используется вектор средних этих параметров для всей записи:

$$\bar{E}(n) = \frac{1}{M}\sum_{m=0}^{M-1} E_m(n).$$

Таким образом, для каждой записи звука улья получается вектор признаков фиксированной размерности, который состоит из средних значений логарифмов энергии сигнала в 30 частотных полосах, равно-разнесенных по шкале мел и оценок их первых производных по времени.

4. Использование метода ОДР. Полученные спектральные показатели были проанализированы с помощью метода ОДР.

4.1. Однофакторный анализ. На первом этапе с использованием метода ОДР анализировалась информативность отдельных спектральных показателей. Работа метода проиллюстрирована на рис. 2 и рис. 3. Звездочками обозначены ульи с маткой, квадратиками — без матки.

На рис. 2 показана связь спектрального показателя $\bar{E}(2)$, соответствующего частотной полосе 2 с центральной точкой 83 Гц, с наличием матки в улье. Вдоль оси Y, не имеющей содержательного смысла и необходимой для развертки изображения, отложены значения показателя $\bar{E}(1)$.

Из рис. 2 видно, что значения показателя $\bar{E}(2)$ выше границы 134.8 в подавляющем большинстве случаев соответствуют ульям без матки: из 55 значений $\bar{E}(2)$ выше границы 134.8 49 значений отвечают ульям без матки, из 53 значений $\bar{E}(2)$ ниже границы 134.8 47 значений — ульям с маткой. Значимость закономерности на рис. 2 была оценена с помощью перестановочного теста на уровне $p < 0.0005$.

На рис. 3 показана связь спектрального показателя $\bar{E}(3)$, соответствующего частотной полосе 3 с центральной точкой 128.1 Гц, с наличием матки в улье. Из рис. 3 видно, что значения показателя $\bar{E}(3)$, отвечающего частотной полосе 3 с центральной точкой 128.1 Гц, выше границы 145.2 в подавляющем большинстве случаев соответствуют ульям без матки: из 60 значений $\bar{E}(3)$ выше границы 145.2 52 значения относятся к ульям без матки, из 48 значений $\bar{E}(3)$ ниже границы 145.2 45 значений — к ульям с маткой. Значимость закономерности на рис. 3 была оценена с помощью перестановочного теста на уровне $p < 0.0005$.

В табл. 1 даны подробные результаты применения ОДР. Частотные полосы упорядочены по величинам функционала Q (2.1). Согласно табл. 1, ОДР выявляет значимые различия между сравниваемыми группами ульев по величинам $\bar{E}$ для всех полос. При этом лучшее разделение достигается с помощью порога 145.2 для $\bar{E}(3)$ и порога 134.8 для $\bar{E}(3)$. Видно, что значимые различия между сравниваемыми группами существуют по всем спектральным показателям. Однако значимость на уровне $p < 0.001$ существует только для первых 15 полос.

На рис. 5 представлена зависимость информативности спектрального показатели в зависимости от соответствующих частот. Вдоль оси X отложен номер частотной полосы. Из рис. 5 видно, что максимальная способность разделять ульи с маткой и без матки наблюдается для показателей $\bar{E}(2)$ и $\bar{E}(3)$, соответствующих центральным частотам 83 и 128.1 Гц. При дальнейшем увеличении частот информативность постепенно снижается, достигая минимального уровня для полос с номером более 15.

4.2. Двухфакторный анализ. На рис. 4 представлена двухмерная закономерность из семейства III, связывающая наличие матки со спектральными показателями $\bar{E}(2)$ и $\bar{E}(23)$. Показатель $\bar{E}(23)$ соответствует полосе 3 с центральной точкой 1838.3 Гц. Из рис. 4 видно, что значения спектральных показателей для 51 из 55 сигналов, отвечающих отсутствию матки, попали в правый нижний квадрант, соответствующий одновременному выполнению неравенств: $\bar{E}(2) > 132$, $\bar{E}(23) < 87.95$. В этот квадрант попали значения спектральных показателей только для трех сигналов, отвечающих наличию матки в улье. Значимость вхождения в модель каждого из показателей $\bar{E}(2)$ и $\bar{E}(23)$ оценивается с помощью метода, описанного в разд. 2 на уровне $p < 0.0004$.

Из рис. 4 видно, что более высокие частоты вносят дополнительный вклад и позволяют улучшить разделение групп ульев.

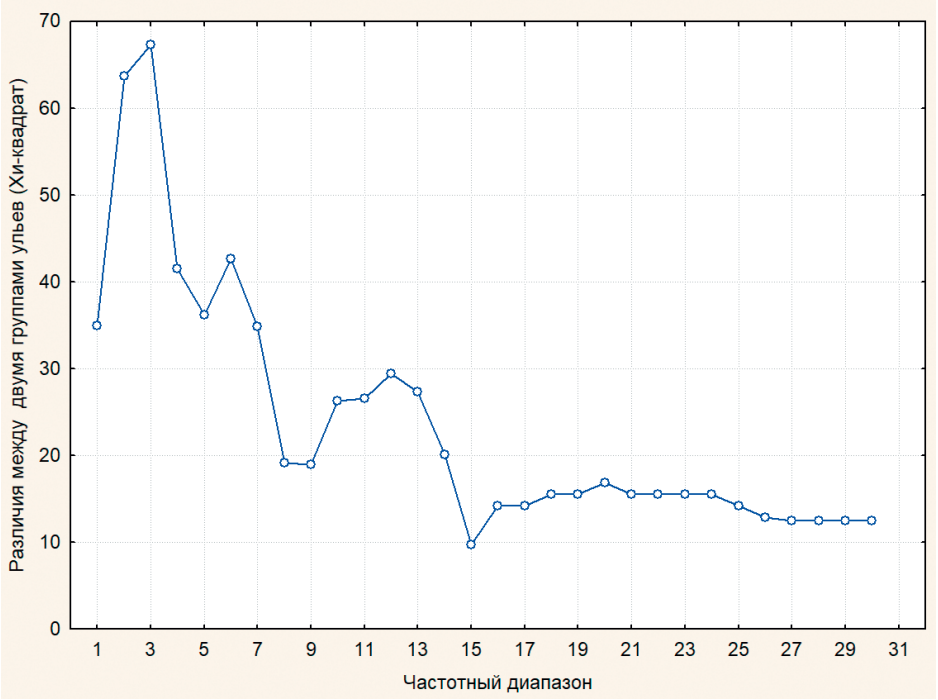

Рис. 5. Зависимость информативности спектрального показателя от номера полосы частот

4.3. Мультифакторный анализ с использованием методов машинного обучения. Для предсказания наличия или отсутствия матки в улье по полной совокупности спектральных показателей применялись методы машинного обучения. Выбор диагноза или прогностического решения осуществляется компьютерным алгоритмом по показателям, значения которых на соответствующий момент известны. При этом настройка алгоритмов производится в автоматическом режиме по данным, для которых известны как значения прогнозируемой величины, так и значения показателей, по которым предполагается вычислять прогноз или диагностику. Процесс настройки алгоритмов принято называть обучением. В настоящем исследовании использовались следующие методы: градиентный бустинг (ГБ), основанный на генерации ансамблей решающих деревьев, вариант метода опорных векторов (МОВ) с гауссианами в качестве базисных функций, метод статистически взвешенных синдромов (СВС), который основан на вычислении коллективных решений по синдромам — областям значений показателей с преобладанием одного из значений прогнозируемой величины.

Оценивание точности алгоритмов машинного обучения происходит с помощью метода Leave One Out, состоящего в проверке точности обученного алгоритма при распознавании состояния отдельных ульев, звуковые сигналы, которые не использовались при обучении этого алгоритма. При этом обучение выполняется по выборке, состоящей из звуковых сигналов для ульев, за исключением того, на котором производится проверка. Применяются стандартные показатели, включая правильность (accuracy), чувствительность, специфичность, а также площадь под ROC-кривой (ROC AUC). При этом под чувствительностью понимается доля правильно распознанных ульев без матки; под специфичностью — доля правильно распознанных ульев с маткой; под точностью — доля ульев без матки в группе, для которых наличие в них матки было распознано.

Результаты распознавания ульев с маткой с помощью перечисленных выше методов машинного обучения представлены в табл. 2. Из таблицы видно, что более высокая точность достигается для логистической регрессии и метода опорных векторов. Более подробно результаты для логистической регрессии приведены на рис. 6.

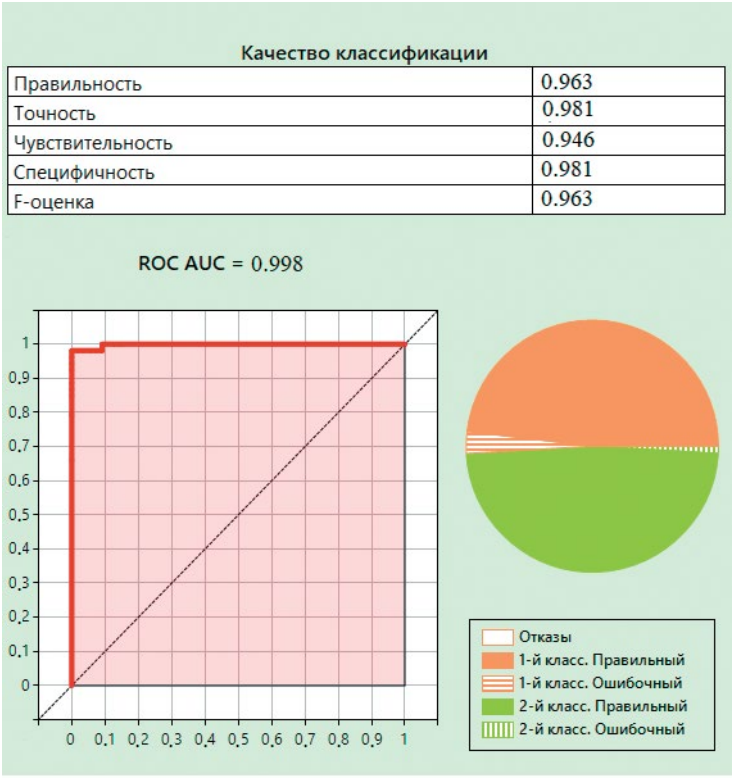

Рис. 6. Результаты распознавания состояния ульев с использованием логистической регрессии

Таблица 2. Результаты работы методов машинного обучения для распознавания ульев

Метод машинного обучения	Чувствительность	Специфичность	ROC AUC
Логистическая регрессия	0.946	0.981	0.998
МОВ	0.946	1.0	0.997
ГБ	0.909	0.868	0.967
СВС	0.891	0.868	0.963

Заключение. Рассмотрен программный комплекс, позволяющий генерировать алгоритмы автоматической классификации сигналов. Генерация алгоритмов производится с помощью нескольких методов машинного обучения, эффективность которых подтверждена многочисленными исследованиями. Преимуществом подхода является возможность выделения и графического представления совокупности статистически значимых эффектов (законо-

мерностей), связанных с решаемой задачей классификации. Такая возможность позволяет в определенной степени преодолеть проблему непрозрачности и плохой интерпретируемости в машинном обучении. Оценивание статистической значимости производится с помощью перестановочных тестов, не требующих априорных предположений о нормальности распределений. Перестановочные тесты могут быть также использованы при выборках произвольного размера. Применение при оценке значимости принципа бритвы Оккама позволяет отклонить закономерности избыточной сложности. Метод является в значительной степени универсальным и может быть использован при решении разнообразных прикладных задач.

СПИСОК ЛИТЕРАТУРЫ

1. *Glorot X.*, *Bordes A.*, *Bengio Y.* Deep Sparse Rectifier Neural Networks // Proc. Fourteenth Intern. Conf. on Artificial Intelligence and Statistics ({AISTATS} 2011). Fort Lauderdale, USA, 2011. V. 15. P. 315—323.
2. *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* Imagenet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems. 2012. No. 1. P. 1097—1105.
3. *Hastie T.*, *Tibshirani R.*, *Friedman J.* Chapter 15. Random Forests. The Elements of Sta-tistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009.
4. *Cortes C.*, *Vapnik V.* Support-vector Networks // Machine Learning. 1995. V. 20. No. 3. P. 273—297.
5. *Breiman L.* Random Forests // Machine Learning. 2001. V. 45. No. 1. P. 5—32.
6. *Freund Y.*, *Schapire R.* A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // J. Computer and System Sciences. 1997. No. 55. P. 119—139.
7. *Schapire R. E.*, *Freund Y.* Boosting: Foundations and Algorithms. Cambridge, Massachusetts, London: MIT Press, 2012.
8. *Friedman J.* Greedy Function Approximation: A Gradient Boosting Machine // The Annals of Statistics. 2001. V. 29. No. 5. P. 1189—1232.
9. *Журавлев Ю. И.*, *Рязанов В. В.*, *Сенько О. В.* “Распознавание”. Математические методы. Программная система. Практические применения. М.: Фазис, 2006.
10. *Davis S.*, *Mermelstein P.* Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences // IEEE Transactions on Acoustics, Speech and Signal Processing. 1980. V. 28. No. 4. P. 357—366.
11. *Haibe-Kains B.*, *Adam G. A.*, *Hosny A.* et al. Transparency and Reproducibility in Artificial Intelligence // Nature. 2020. No. 586. E14—E16.
12. *Kuznetsova A. V.*, *Kostomarova I. V.*, *Senko O. V.* Modification of the Method of Optimal Valid Partitioning for Comparison of Patterns Related to the Occurrence of Ischemic Stroke in Two Groups of Patients // Pattern Recognition and Image Analysis. 2014. V. 24. \textnumero 1. P. 114—123.
13. *Senko O. V.*, *Kuznetsova A. V.*, *Matveev I. A.*, *Litvinchev I. S.* Chapter 8 — Search of Regularities in Data: Optimality, Validity, and Interpretability. Eds P. Vasant, J. Thomas, E. Munapo, G.-W. Weber Advances of Artificial Intelligence in a Green Energy Environment. Amsterdam: Acad. Press, 2022. P. 151—171.
14. *Good P.* Permutation, Parametric and Bootstrap Tests of Hypotheses, 3rd Ed. N. Y.: Springer-Verlag, 2005.
15. *Pesarin F.*, *Salmaso L.* Permutation Tests for Complex Data: Theory, Applications and Software. Hoboken: John Wiley and Sons, Ltd 2010.
16. *Kim H. J.*, *Fay M. P.*, *Feuer E. J.*, *Midthune D. N.* Permutation Tests for Jointpoint Regression with Applications to Cancer Rates // Stat Med. 2000. No. 19. P. 335—351.
17. *Akaike H.* A New Look at the Statistical Model Identification // IEEE Transactions on Automatic Control. 1974. V. 19. No. 6. P. 716—723.
18. *Schwarz G.* Estimating the Dimension of a Model // Annals of Statistics. 1978. V. 6. P. 461—464.
19. *Rissanen J.* Modeling by Shortest Data Description // Automatica. 1978. V. 14. No. 5. P. 465—658.
20. *Senko O. V.*, *Dzyba D. S.*, *Pigarova E.*, *Rozhinskaya L. Ya.*, *Kuznetsova A. V.* A Method for Evaluating Validity of Piecewise-linear Models // Intern. Conf. on Knowledge Discovery and Information Retrieval (KDIR-2014). Rome, 2014. P. 437—443.
21. *Domingos P.* The Role of Occam’s Razor in Knowledge Discovery // Data Mining and Knowledge Discovery. 1999. V. 3. No. 4. P. 409—425.
22. *Доровских И. В.*, *Сенько О. В.*, *Чучупал В. Я.*, *Докукин А. А.*, *Кузнецова А.* В. Исследование возможности диагностики деменции по сигналам ЭЭГ с помощью методов машинного обучения // Мат. биология и биоинформатика. 2019. Т. 14. № 2. С. 543—553.
23. *Еськов Е. К.*, *Тобоев В. А.* Изменение структуры звуков, генерируемых пчелиными семьями в процессе социотомии // Зоол. журн. 2011. Т. 90. № 2. С. 192—198.
24. *Еськов Е. К.* Генерация, восприятие и использование акустических и электрических полей в коммуникациях медоносной пчелы // Биофизика. 2013. Т. 58. № 6. С. 1051—1064.
25. *Дрейзин В. Э.*, *Рыбочкин А. Ф.*, *Захаров И. С.* Акустический контроль пчелиных семей с помощью вычислительной техники // Пчеловодство. 2000. № 4. С. 25—27.