ГЕНЕТИКА, 2024, том 60, № 7, с. 3–16

ОБЗОРНЫЕ И ТЕОРЕТИЧЕСКИЕ СТАТЬИ

УДК 575.174.015.3, 004.855.5

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И КЛАССИЧЕСКИЕ МЕТОДЫ В ГЕНЕТИКЕ И СЕЛЕКЦИИ ЖИВОТНЫХ

© 2024 г. А. Д. Солошенков[1,2,*], Э. А. Солошенкова[1], М. Т. Семина[1], Н. Н. Спасская[3], В. Н. Воронкова[1], Ю. А. Столповский[1]

[1] *Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, 119991 Россия*
[2] *Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, 127434 Россия*
[3] *Московский государственный университет имени М.В. Ломоносова, Научно-исследовательский Зоологический музей, Москва, 125009 Россия*
**e-mail: alesol@rgau-msha.ru*

Поступила в редакцию 29.11.2023 г.
После доработки 07.03.2024 г.
Принята к публикации 15.03.2024 г.

В настоящей статье проведены обзор и анализ основных методов популяционной генетики и селекции животных, а также математических методов машинного обучения, используемых в животноводстве. На примере двух доместицированных видов – домашняя лошадь (*Equus caballus*) и северный олень (*Rangife rtarandus*) проведено обучение моделей библиотеки CatBoost. Для обучения модели на основе данных по одомашненным и диким северным оленям, европейским и российским породам лошадей использованы результаты, полученные с помощью микросателлитных панелей, соответственно локусов 16 и 17. Для определения успешности модели были рассчитаны стандартные показатели: Accuracy, Precision, Recall и *F1*, построены матрицы ошибок. Показаны новые возможности идентификации породной принадлежности животных.

Ключевые слова: машинное обучение, нейронные сети, животноводство, лошадь, *Equus caballus*, северный олень, *Rangifer tarandus,* микросателлитный анализ, генетическое разнообразие.

DOI: 10.31857/S0016675824070017 **EDN:** BIMRAP

Современное развитие индустриального, органического, традиционного животноводства связано с анализом большого спектра хозяйственно-полезных признаков (экстерьера и интерьера животных), качественных и количественных показателей продуктивности, генетических маркеров, генов-кандидатов, данных о секвенировании геномов, в совокупности различных критериев отбора в зоотехнии и ветеринарии. Необходимость работы с большими массивами данных, а также новейшие возможности получения “цифровых фенотипов” определили интенсивное развитие цифровых технологий, математических методов анализа и интегрирование машинного обучения и нейросетей в практику животноводства, в том числе в глобальную проблему сохранения генетических ресурсов, уникального адаптивного потенциала редких локальных пород животных [1].

Классические методы биоинформационного анализа в генетике и селекции животных включают в себя оценку частот, число эффективных и уникальных аллелей, расчеты генетических дистанций, определение филогенеза, определение различных коэффициентов оценки уровня генетического разнообразия и т. д. Филогенетический анализ, изучение генетических и селекционных взаимоотношений между породными группами и популяциями проводятся с помощью построения дендрограмм. Наибольшее распространение получили методы UPGMA и Neighbor-joining, а также программа STRUCTURE, которая активно используется для исследования популяционной структуры методом байесовского анализа марковских цепей. Для оценки влияния генотипа на продуктивность животных применяются методы дисперсионного анализа, в частности BLUP (Best Linear Unbiased Prediction) и его подвиды, основанные на статистической модели, предложенной C.R. Henderson [2].

Широкое распространение нейросетей и искусственного интеллекта позволяет проводить исследования в биологии, генетике и животноводстве. В настоящее время машинное обучение используется для мониторинга состояния и благополучия животных при их содержании, в идентификации отдельных особей и других направлениях, например для предсказания корреляций нуклеотидных замен и продуктивности животных. Очевидно, что использование современных методов мониторинга,

оценки родословных, идентификации животных может значительно улучшить качество управления стадом, разведения и селекции животных.

Одной из важных (насущных) проблем животноводства является идентификация породной принадлежности животных. В связи с этим в нашей работе на большом массиве данных по микросателлитным маркерам европейских и российских пород лошадей, а также домашних и диких северных оленей был использован один из методов машинного обучения (CatBoost) и проведено обучение модели с целью идентификации особей, популяций, пород.

В настоящей статье описаны классические методы, которые используются в генетике и селекции животных, проведен анализ современных методов машинного обучения и их перспектив в животноводстве.

ПОПУЛЯЦИОННО-ГЕНЕТИЧЕСКИЕ ПАРАМЕТРЫ

Параметры, применяемые для оценки “генетического благополучия” породы или популяции, рассчитываются на основе генетических профилей животных и частот аллелей. Под “генетическим благополучием” мы подразумеваем определенную степень инбридинга и генетического разнообразия в породе или популяции. Оценка генетического разнообразия и филогенетический анализ, основанные на молекулярных маркерах, позволяют идентифицировать породные группы с низким уровнем аллельного разнообразия, которое может снижаться ввиду длительной изолированности популяции, влияния внешних факторов, а также жесткой системы подбора пар животных в племенном животноводстве [3]. Как правило, в племенных хозяйствах используются несколько выдающихся по продуктивности производителей на большом количестве особей маточного поголовья. Интенсивные методы селекционной работы, а также сокращение поголовья сельскохозяйственных животных в России, особенно в коневодстве, приводят к необходимости постоянного контроля за уровнем генетического разнообразия с целью нивелирования негативных эффектов инбридинга [4].

В настоящее время во всем мире для контроля происхождения и определения статуса пород основных доместицированных видов животных применяются микросателлитные (short tandem repeats) маркеры, которые рекомендованы ISAG (международное общество генетики животных https://www.isag.us/). В то же время все интенсивнее используются подходы, связанные с детекциями однонуклеотидных замен (SNP, single nucleotide polymorphism), особенно ассоциированных с конкретными фенотипическими признаками. Микросателлитные маркеры отличаются высокой вариабельностью и широкой представленностью в геноме. Панели, применяемые для идентификации животных, используют наиболее полиморфные локусы, которые при этом считаются условно нейтральными (не локализованы рядом с кодирующими участками ДНК, участвующими в отборе) [5].

Благодаря использованию SNP-маркеров в животноводстве возможно ускорение темпов селекции при привлечении таких смежных областей, как эмбриология, биоинформатика и математическая генетика [6].

Наиболее широко распространенными показателями при оценке генетического разнообразия являются: ожидаемая (H_e) и наблюдаемая (H_o) гетерозиготность. Данные параметры основаны на уравнении Харди – Вайнберга и позволяют выявить недостаток гетерозигот в популяции. Ожидаемая гетерозиготность (разнообразие по M. Nei) показывает вероятность гетерозиготности особи в популяции, рассчитывается по формуле:

$H_e = 1 - \sum_1^i p_i^2$, где p_i – частота i-го аллеля, n_i – общее число аллелей во всех локусах.

Значения для H_e и H_o варьируют от 0 (нет гетерозиготности) до практически 1.

H_o – наблюдаемая гетерозиготность, т. е. фактическая доля гетерозиготных образцов.

Так, при изучении генетического разнообразия пород лошадей России наибольший уровень гетерозиготности был идентифицирован для русской верховой породы лошадей ($H_o = 0.71$), в формировании которой участвуют несколько пород: чистокровная верховая, немецкие спортивные, ахалтекинская [7].

Также проводится расчет параметров инбридинга, в частности коэффициента F_{is}, с помощью которого измеряют внутрипопуляционный инбридинг, т. е. снижение гетерозиготности индивида по причине близкородственных скрещиваний, F_{it} – межпопуляционного коэффициента инбридинга, где учтены поправки на дифференциацию между популяциями [8].

Для оценки различий между популяциями используют методы расчетов генетических расстояний F_{st} от 0 – различий нет до 1 – максимальное различие. F_{st} также является коэффициентом инбридинга в популяциях в сравнении с общей выборкой. Подобные методы помогают установить дифференциацию популяций [9]. Так, низкий параметр F_{st} может наблюдаться для двух близкородственных пород, например как донская и буденновская породы лошадей (0.02). И напротив, высокий показатель наблюдается у пород, имеющих различное историческое происхождение (тяжеловозные и верховые породы лошадей – 0.15).

Программная среда R, активно используемая для анализа генетических данных, позволяет

визуализировать различные показатели. При использовании пакета PopGenReport частоты аллелей представляются в виде тепловых карт (рис. 1), по которым выявляются общие тренды и индивидуальные для каждой отдельной популяции [10].

Так, при анализе генетического разнообразия ценного промыслового вида – соболя было выявлено среднее число аллелей на популяцию от 7.73 до 10.73 (табл. 1). Так как соболь обладает высокой миграционной активностью, это позволило сделать выводы о его миграциях. Наименьшее количество аллелей наблюдалось на Камчатке (7.73), а наибольшие показатели – для популяций, находящихся на пересечении миграционных потоков и обогащающихся за счет этого новыми аллельными вариантами [11].

Как указывает ряд авторов [12], важно учитывать тот факт, что для анализа разнообразия изначально исследователями выбираются наиболее полиморфные локусы. В исследовании сравнили оценку нуклеотидного разнообразия и анализ по микросателлитным маркерам. Была выявлена положительная корреляция между данными показателями, однако гетерозиготность была выше в 1.4 раза для нуклеотидных замен в сравнении с панелью полиморфных микросателлитных локусов.

Кроме того, изначально выбираются нейтральные маркеры, которые подчиняются закону Харди – Вайнберга. Это может являться причиной искажения результатов и смещения оценки ввиду утери части данных.

КЛАСТЕРИЗАЦИЯ ПОПУЛЯЦИЙ

Исследования филогенеза пород одомашненных и популяций диких животных позволяют оценить микроэволюционные процессы и их историческое формирование [10].

Для кластерного анализа (рис. 2) используются программы STRUCTURE и Geneland, с применением алгоритма Монте-Карло по схеме марковских цепей (MCMC) для байесовской статистики. В программах задают предполагаемое количество популяций и число итераций, далее рассчитываются графики для каждого числа популяций по каждой итерации [13].

Для построения непосредственно филогенетических деревьев широкое распространение получили методы UPGMA и Neighbor-joining. Эти

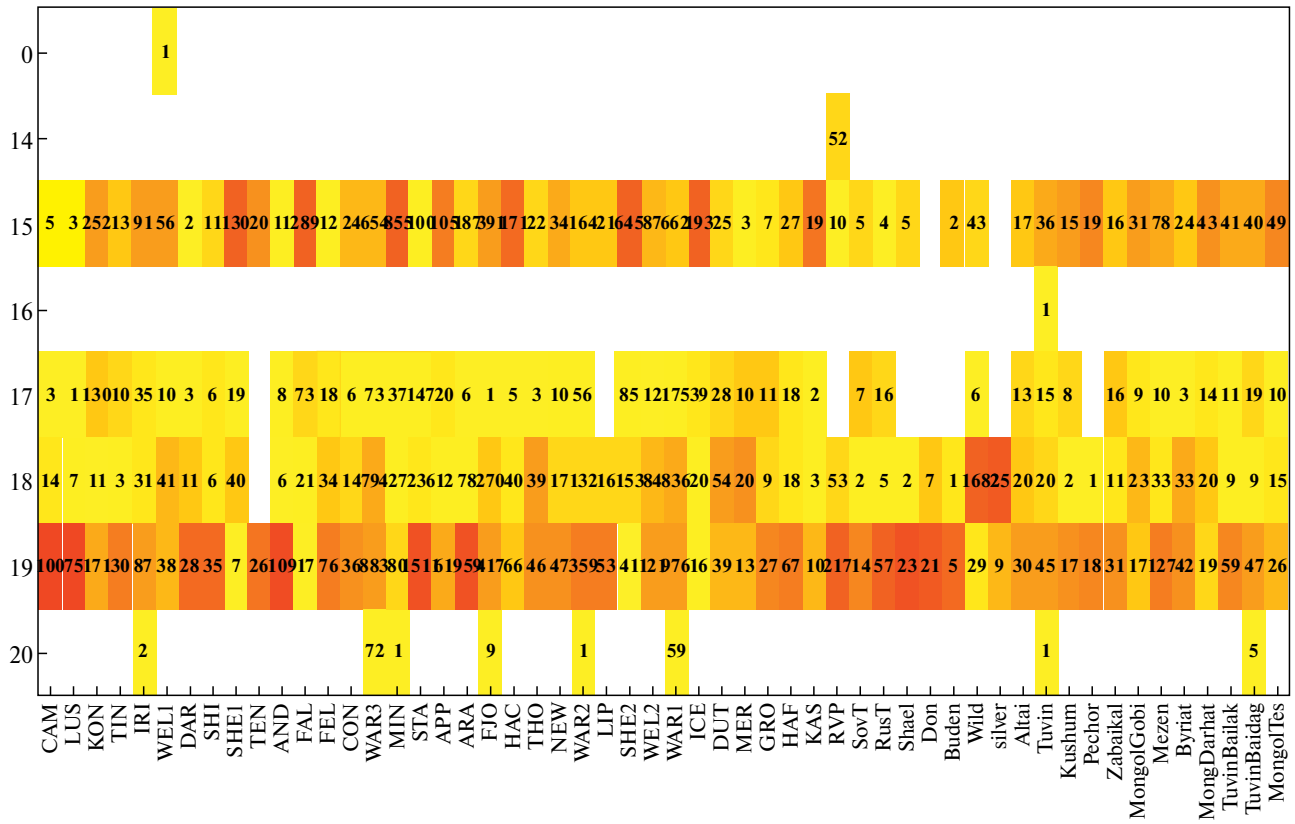

Рис. 1. Тепловая карта частот аллелей для 54 пород лошадей. Цветом от желтого к красному обозначена частота встречаемости аллеля в популяции. Идентифицирован приват-аллель 14-го локуса HTG7 для русской верховой породы лошадей (RVP).

Таблица 1. Показатели генетического разнообразия для различных популяций соболя

Популяция/регион	N	A	AR	H_o	H_e	P
Ванавара (Ц. Сибирь)	31	8.55	6.44	0.704	0.761	0.461
Ербогачен (Ц. Сибирь)	28	8.09	5.99	0.660	0.729	0.415
Братск (Ц. Сибирь)	23	8.18	6.43	0.675	0.770	0.277
Саяны	31	9.18	6.69	0.707	0.786	0.286
Сихотэ-Алинь (Приморье)	40	10.73	7.37	0.741	0.809	0.270
Северный Урал	71	9.36	6.17	0.695	0.733	0.471
Камчатка	37	7.73	5.29	0.713	0.706	0.926

Примечание. N – размер выборки; A – среднее число аллелей на локус; AR – обогащенность популяций аллелями (allelic richness), вычисленная как среднее число аллелей, нормированное на объем выборки N; H_o – наблюдаемая гетерозиготность; H_e – ожидаемая гетерозиготность; P – значение вероятности для теста Харди – Вайнберга с учетом всех локусов.

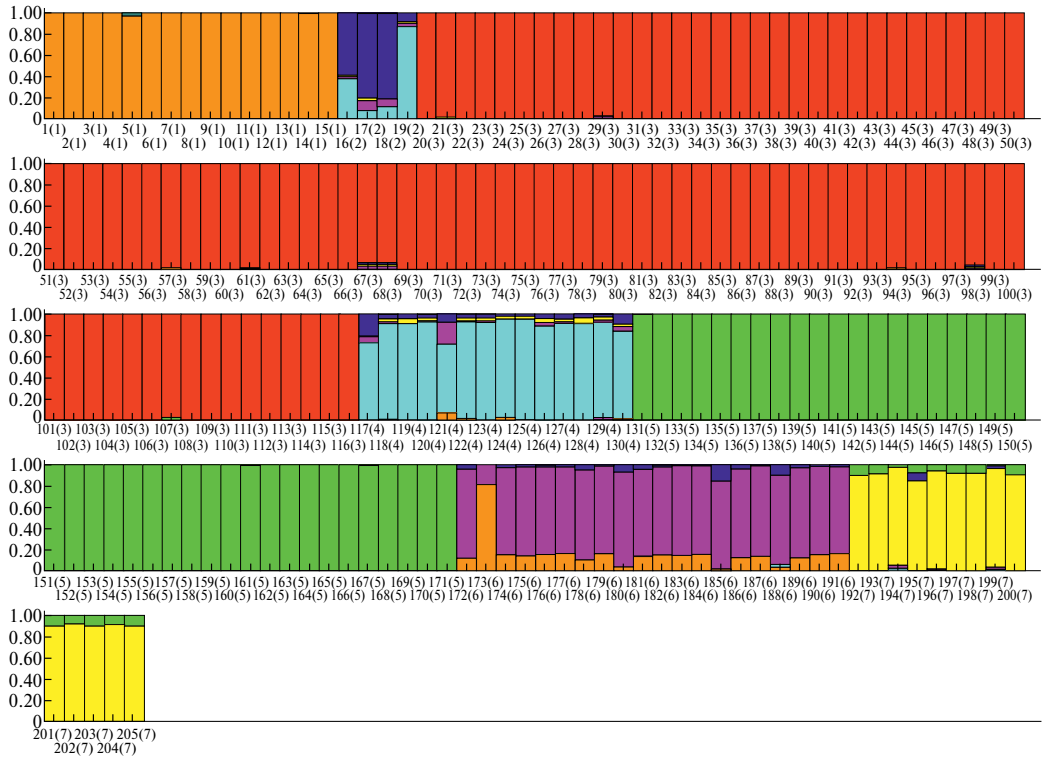

Рис. 2. Популяционная структура заводских пород лошадей. Оранжевый цвет - ахалтекинская, синий – буденновская, красный – одичавшие лошади о. Водный, голубой – донская, зеленый – русская тяжеловозная, розовый – русская верховая, желтый – советская тяжеловозная.

методы являются дистанционными, т. е. первоначальные данные рассчитываются в виде матрицы генетических расстояний, которая впоследствии преобразуется в дихотомичное дерево [14].

Основная идея бутстреп-процедуры по B. Efron состоит в том, чтобы методом статистических испытаний Монте-Карло многократно извлекать повторные выборки из эмпирического распределения, возможно, при использовании метода главных компонент, позволяющего отойти от стандартной модели построения дендрограмм [15].

Метод главных компонент уменьшает размерность данных, преобразуя их в ковариационную матрицу – матрицу главных компонент. PCA (principal component analysis) отличается высокой степенью воспроизводимости результатов ввиду

отсутствия введения в расчеты различных поправок либо коэффициентов. Единственной мерой расчета является доля дисперсии данных [16].

Таким образом, при использовании метода возможно визуализировать данные в пространстве двух главных компонент (двухмерное пространство) и трех главных компонент (трехмерное пространство) [17].

Сотрудниками лаборатории сравнительной генетики животных был проведен анализ главных компонент для заводских пород лошадей (тяжеловозные и верховые породы) в сравнении с одичавшими лошадьми о. Водный, чей статус и происхождение остаются неуточненными [18]. Построенные методом UPGMA дендрограммы показали низкий уровень бутстреп-поддержки для верховых пород лошадей, что не позволяло достоверно разделить их на отдельные породы и сделать выводы о возникновении одичавшей популяции. Однако в пространстве двух главных компонент было обнаружено частичное перекрытие с буденновской и донской породами лошадей, что позволило сделать предположение о возникновении популяции одичавших лошадей от данных пород (рис. 3).

В большинстве случаев анализ данных происходит существующими пакетами и библиотеками для R или Python. Так, библиотека *poppr* для R позволяет строить UPGMA и NJ деревья. Расчет происходит на основе генетических дистанций Nei (1972) [19]. Тем не менее авторами пакета указывается ряд моментов, которые необходимо учитывать при обсчетах: различные модели мутационных процессов (пошаговая мутация) либо отсутствие учета мутаций, а также проблемы с обсчетами для организмов с различной плоидностью [20]. Кроме того, как уже указывалось выше, деревья ограничены дихотомией, что приводит к неоднозначным результатам. Л.А. Животовским в книге “Генетика природных популяций” [8] наглядно проиллюстрированы неоднозначность сжатия матрицы генетических дистанций и дальнейшее построение деревьев (рис. 4).

При построении дерева методом UPGMA наблюдается неоднозначность отнесения популяций 1 и 2 к различным кластерам. При включении обеих популяций или только одной из них наблюдаются три разных дерева с отнесением популяций 1 и 2 либо к отдельному кластеру, либо популяции 1 к кластеру A, а популяции 2 к кластеру B, что в дальнейшем объясняется распределением выборок в пространстве главных компонент. Тем не менее данные методы являются основными при наглядном представлении структуры выборок. Поэтому необходимо проводить расчеты различными методами с дальнейшим их глубоким анализом для понимания популяционных процессов в популяциях.

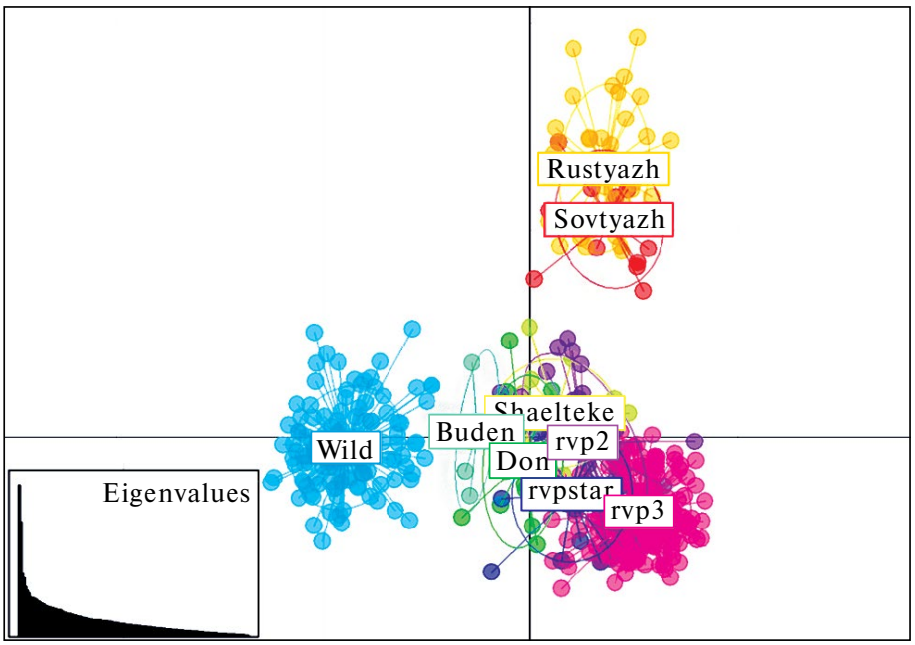

Рис. 3. Распределение верховых и тяжеловозных пород лошадей в пространстве двух главных компонент в сравнении с одичавшими лошадьми о. Водный для уточнения происхождения данной популяции. Wild – одичавшие лошади, Buden – буденновская порода; Shaelteke – ахалтекинские лошади завода “Шаэль”; Don – донская; rvp2, rvpstar, rvp3 – выборки русской верховой породы; Rustyazh – русская тяжеловозная; Sovtyazh – советская тяжеловозная порода.

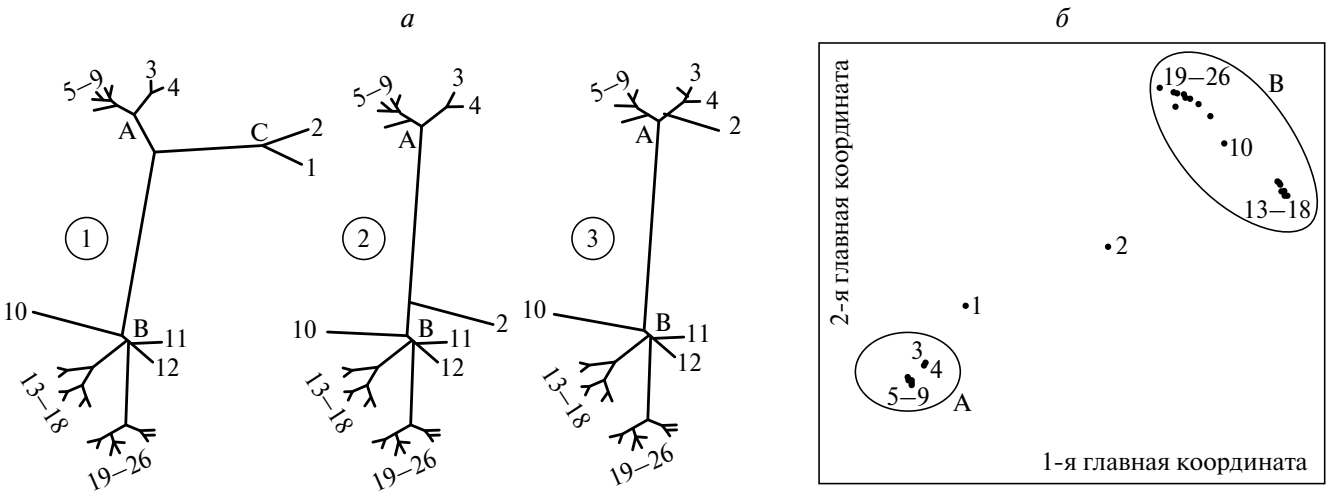

Рис. 4. Построение дерева методом VPGVA. *а* – дихотомическая кластеризация выборок кеты; *б* – метод главных компонент.

BLUP (BEST LINEAR UNBIASED PREDICTION)

Для оценки проводимых селекционных программ традиционно использовались методы “матери – дочери”, “дочери – сверстницы”, основанные на расчете разницы средних величин продуктивности животных. Различия в эколого-географических условиях, рационе, условиях содержания животных являлись факторами смещения данных показателей. Для повышения эффективности оценки селекционных процессов был введен метод оценки BLUP (наилучший линейный несмещенный прогноз), позволяющий учитывать влияние как генотипических, так и паратипических факторов, таких как возраст, возраст отела, год, дата постановки на откорм, среднесуточный удой и прочие. В зависимости от целей селекции, отбора и подбора животных, а также наличия тех или иных данных используют различные виды данного метода. Данный метод был предложен C.R. Henderson в 1984 г. [21].

По данным П.И. Отраднова и соавт. приведено уравнение смешанной модели в матричном виде [22]:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Матрицы X и Z являются матрицами дизайна для фиксированных (X) и рандомизированных (Z) эффектов. Соответственно X' и Z' – транспонированные матрицы, h^2 – коэффициент наследуемости (его нет в данной формуле) оценивает отношение изменчивости, обусловленной генетическими факторами, к общей изменчивости (генетическая и паратипическая), из него рассчитывают λ обратная матрица родства).

При использовании методов GWAS (Genome-Wide Association Study) и наличии данных нуклеотидных замен возможно использование метода GBLUP (Genomic Best Linear Unbiased Prediction), при котором используется матрица геномного сходства G вместо матрицы родства.

Так, при исследовании голштинизированного черно-пестрого скота Московской области с использованием моделей BLUP Sire Model (BLUP SM), GBLUP был выявлен среднегодовой генетический тренд +37 кг молока, +1.7 и +0.8 кг продукции молочного жира и белка, а в Ленинградской области – +84, +3.3 и +2.3 кг. Причем было доказано, что оценка по нуклеотидным заменам (GBLUP) достовернее, чем по быкам-производителям (Sire Model) [23].

BLUP удобно использовать в современных условиях автоматизации процессов животноводства, например при использовании откормочных станций, которые позволяют учитывать нахождение на кормовой станции, среднесуточное потребление корма, скорость потребления корма, его конверсию и ряд других показателей, которые затем возможно использовать в будущей модели. Так, при исследовании свиней породы дюрок в модель включались год и месяц рождения животных, дата постановки на откормочную станцию, данные кормовой станции, количество недель выращивания на ней. Помимо этого в анализ была включена матрица родства [24].

Одним из плюсов BLUP следует отметить несмещенность прогноза и отсутствие необходимости в нормальности распределения данных, так как учитываются многие факторы, как генетические, так и средовые.

МАШИННОЕ ОБУЧЕНИЕ

Развитие цифровых технологий и непосредственно технической составляющей современных компьютерных систем позволили ввести анализ больших данных в современное животноводство, генетику и селекцию [25]. Обучение искусственному интеллекту позволяет отойти от классических стандартов применения математических моделей. По данным базы PubMed (https://www.ncbi.nlm.nih.gov/) за 2023-й год опубликовано 110 статей по использованию машинного обучения в животноводстве.

Машинное обучение (Machine Learning, ML) – класс методов искусственного интеллекта, основанный на статистических моделях и логических операциях, позволяющий автоматически улучшать вычислительные алгоритмы при отсутствии четких инструкций с использованием примеров данных или прошлого опыта [26].

Обучая программу (алгоритмы) на основе экспериментальных данных по генотипированию животных, мы получаем модель, которая может делать прогнозы (например, определить породу животного) на основе наблюдений (например, по генотипу и фенотипу). Если рассматривать информатику как предмет алгоритмов, то машинное обучение является предметом обучения алгоритмов.

Модель может быть прогностической – моделирует данные в будущем, описательной – получает знания из существующих данных или комбинированной. Машинное обучение использует теорию статистики при построении математических моделей, поскольку основная задача – делать выводы на основе выборки.

В классическом ML выделяют два основных способа:

- “Обучение с учителем” (supervised learning) – такой вид обучения подразумевает наличие размеченных данных (обучающей выборки), потенциально связанных некоторой закономерностью. Модель обучается по принципу “стимул – реакция” и позволяет решать задачи классификации или регрессии, а качество модели определяется по тестовой (иногда валидационной) выборке.

- “Обучение без учителя” или неконтролируемое обучение (unsupervised learning) – обучение на неразмеченных данных. В классических задачах unsupervised learning есть данные, но нет обучающей выборки (т. е. правильные ответы неизвестны). При таком обучении модель обучается выявлять скрытые взаимосвязи без контроля со стороны исследователя и позволяет решать задачи кластеризации, ассоциации и уменьшения размерности (обобщения).

Для определения точности, полученной в ходе обучения модели, используются метрики качества. Для каждой задачи используются свои метрики. В контексте задач классификации выделим Accuracy, Precision, Recall и F_1.

Для понимания концепции метрик качества необходимо сказать о матрице ошибок (confusion matrix). Матрица представляет таблицу, позволяющую проиллюстрировать качество обучения модели, как правило, контролируемого обучения (supervised learning). В случаях unsupervised learning ее называют матрицей соответствия (matching matrix) [27].

Разберем, как устроена матрица ошибок на примере задачи бинарной классификации. В таких задачах алгоритм учится предсказывать принадлежность объекта с определенным набором данных (features) к одному из двух классов. Матрица ошибок представлена на рис. 5.

Здесь True class – истинная метка класса на этом объекте, то есть истинные значения классов, изначально содержащиеся в данных. Predicted class – это ответ алгоритма на объекте, т. е. значения классов, которые предсказывает обученная модель для элементов выборки. Если истинная метка класса для объекта 1 и модель отнесла ее к соответствующему классу ($1 \rightarrow 1$), то предсказание учитывается в поле True Positive (TP). Если для объекта с истинным классом 0 модель отнесла его к нулю ($0 \rightarrow 0$), то предсказание учитывается в True Negative (TN). В случае если объект с меткой 1 был отнесен к классу 0 ($1 \rightarrow 0$), предсказание учитывается в поле False Negative (FN), а для меток 0, отнесенных к 1 ($0 \rightarrow 1$), – в False Positive (FP).

Таким образом, суммы правильных предсказаний классов записываются в True Positive и True Negative, а ошибки – в False Positive и False Negative.

Например, мы обучили модель на данных, содержащих 100 объектов, разделенных на два класса поровну (50 объектов класса 0 и 50 объектов класса 1), но по различным причинам алгоритм не обучился распознавать классы со 100% точностью. Представим, что проверка модели дала следующие результаты: объектов класса 1 с правильно предсказанными метками было 40 единиц, объектов класса 0 – 30 единиц соответственно, объектов класса 1, отнесенных к классу 0, – 10 единиц, а объектов класса 0, отнесенных к классу 1, – 20 единиц. Матрица ошибок представлена на рис. 6.

Метрика Accuracy (общая точность) – это метрика, которая характеризует долю правильных ответов алгоритма, т. е. то, насколько близок данный набор измерений (классов, наблюдений, показаний) к их истинному значению [28]:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

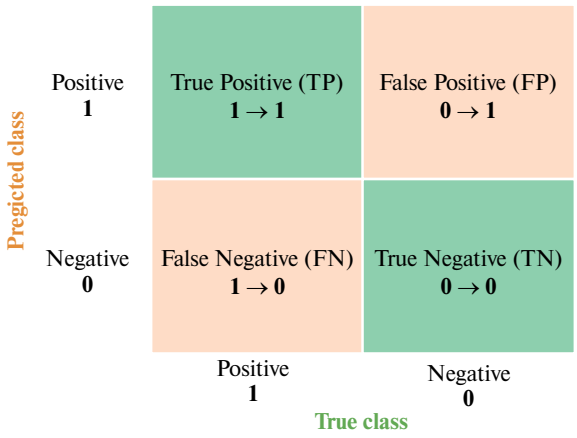

Рис. 5. Матрица ошибок.

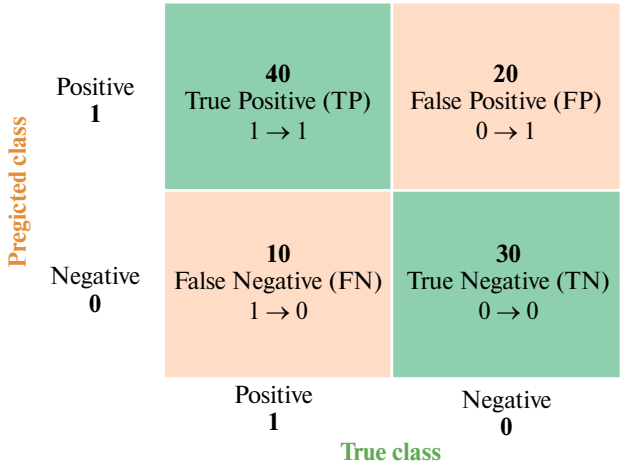

Рис. 6. Матрица ошибок примера.

Используется, когда классы сбалансированы. В случае дисбаланса классов лучше воспользоваться другими метриками. Для оценки качества модели на каждом из классов по отдельности обычно используют метрики Precision и Recall, они не зависят, в отличие от Accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок.

Метрика Precision (положительная точность, positive predictive value) – это доля правильно предсказанных положительных объектов относительно всех объектов, отнесенных моделью к положительному классу [29]:

$$precision = \frac{TP}{TP + FP} \cdot$$

Чем меньше ложноположительных срабатываний будет допускать модель, тем больше будет ее Precision.

Метрика Recall (полнота, sensitivity in diagnostic binary classification) характеризует долю правильно предсказанных положительных объектов среди всех объектов положительного класса [28]:

$$recall = \frac{TP}{TP + FN} \cdot$$

Recall не зависит от True Negative и False Positive в отличие от Precision, то есть, если модель правильно прогнозирует все Positive объекты, Recall будет равен единице (100%) даже для случаев, когда все объекты класса Negative были ошибочно определены как Positive. Для объединения Precision и Recall в обобщенный критерий качества применяют F-меру.

Метрика F1-мера (*F1*-measure) – среднее гармоническое Precision и Recall, является мерой точности в задачах бинарной классификации [30]:

$$F_1 = 2\frac{recall \times precision}{recall + precision} = \frac{TP}{TP + \frac{FP + FN}{2}} \cdot$$

F1-мера достигает максимума при Precision и Recall, равных единице, и близка к нулю, если один из аргументов близок к нулю. *F1*-мера применяется, если метрики precision и recall одинаково важны.

Для случаев, когда одна из используемых метрик приоритетнее, используют F_β-меру:

$$F_\beta = \left(1 + \beta^2\right) \times \frac{precision \times recall}{\beta^2 \times precision + recall},$$

где β – весовой коэффициент Precision в F_β-мере.

F_β-мера применяет дополнительные веса, придавая одной из метрик Precision или Recall большее значение, чем другой.

Таким образом, качество обучения моделей определяется по соответствующим метрикам.

Машинное обучение применимо для решения задач в области генетики, например, когда в качестве исходных для обучения модели данных (features) используются нуклеотидные последовательности или генотипы. Предсказание уровня и направленности экспрессии генов на основе большого объема данных по SNP позволяет определять эффекты различных нуклеотидных замен, а также последствия соматических мутаций и изменения в хроматине. Существующая модель DeepSEA – система алгоритмов глубокого обучения, созданная для этих целей, предсказывает изменения в хроматине на основании данных секвенирования с чувствительностью до одного нуклеотида. Широкое развитие GWAS и NGS методов позволяет накапливать большое количество датасетов по rs (Reference SNP cluster ID), ассоциированных с

различными заболеваниями, что в совокупности с данной моделью может быть актуально для исследования в области медицинской генетики [31].

В области животноводства активно исследуются и разрабатываются различные технологии: компьютерное зрение для решения задач классификации и мониторинга состояния и поведения животных, электронные ошейники для наблюдений за состоянием здоровья и активностью носителя, модели машинного и глубокого обучения для ускорения и повышения качества селекционных процессов, экспертные системы в области оптимизации процессов учета животных, анализа родословных, точечных мутаций, их влияния на резистентность и продуктивность животных и другие. Использование подобных технологий позволяет снизить затраты труда при содержании животных и ветеринарном контроле [32].

Так как паспортизация и создание баз данных генетических профилей животных не всегда доступны для фермеров и селекционеров, в особенности в отдаленных регионах, все чаще начинают применяться методы обучения моделей на основе фотографического материала.

Например, модель VGG-16 (рис. 7), предложенная K. Simonyan и A. Zisserman в 2014 г., достигает высокой точности при определении объектов на изображении (93%) [33]. В качестве DataSet используется база данных ImadeNET (https://image-net.org/), в которой в настоящее время находится 1.2 млн изображений, относящихся к 1000 категорий.

На основе данной модели ряд исследователей обучили свои варианты нейросети для задач идентификации как отдельных животных, так и пород на основе их фотографий. Применение предобученных моделей для решения узконаправленных задач – распространенная практика на сегодняшний день.

S.A. Jwade с соавт. обучили модель определять одну из четырех пород овец: меринос, суффолк, белый суффолк, полл дорсет, используя данные 1642 животных [34]. Для стандартизации условий для фото был создан специальный загон. Таким образом удалось снизить вероятность ошибочного обучения ввиду различий в средовых условиях съемки. Показатель Accuracy при этом достиг 95.8%.

Подобные исследования проведены и на других видах животных. В 2020 г. были проведены исследования для идентификации собак по коллекции фото [35]. Актуальность этой работы заключается в возможности внедрения предложенного метода для

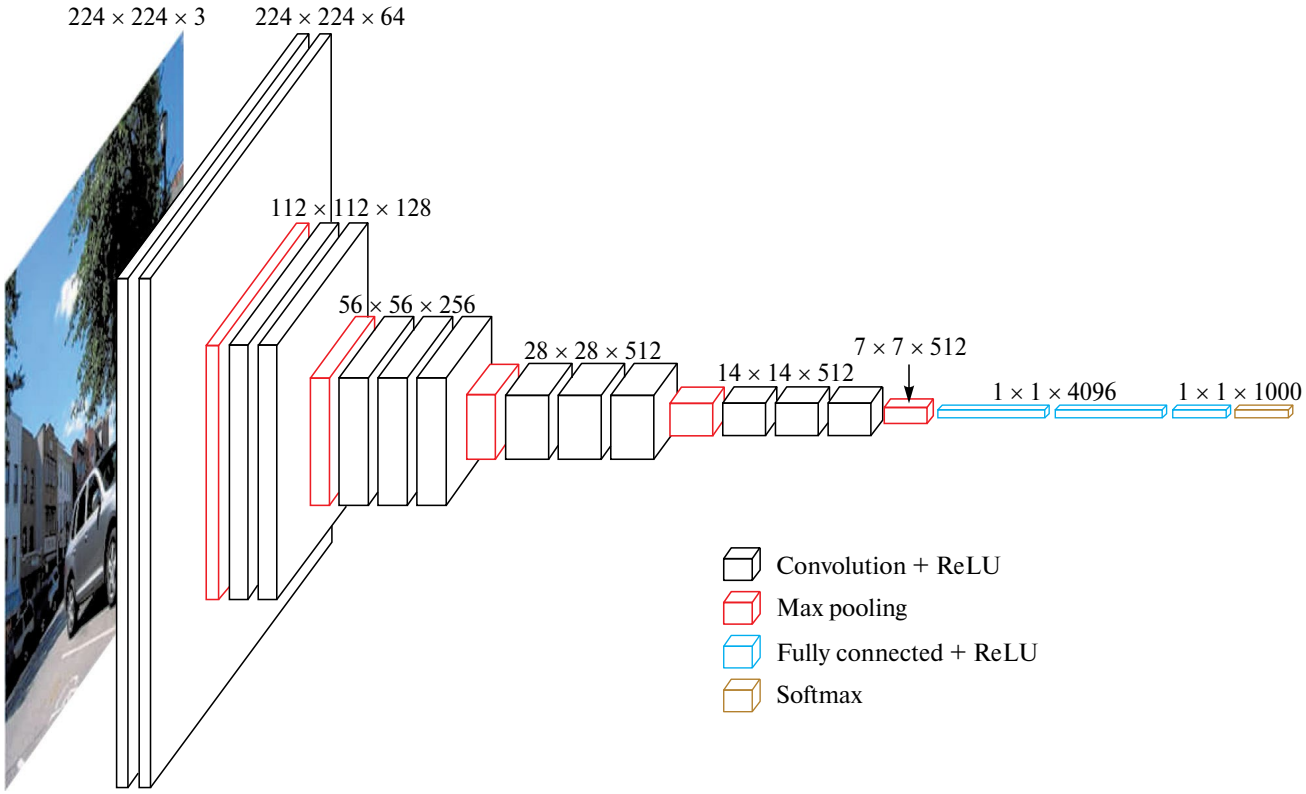

Рис. 7. Архитектура VGG-16.

поиска пропавших животных. Авторы ссылаются на ассоциацию American Pet Products Association (APPA), по данным которой из 78 млн домашних собак 15% теряются и 7% (819 000 особей) не удается вернуть владельцам. Несмотря на небольшой объем входных данных, включающих 21 собаку пород хаски и мопсов, а также данные из интернет-ресурсов по 10 породам, с пятью фотографиями по каждой породе, удалось обучить модель до 77.19% точности на кросс-валидации.

Сотрудники лаборатории сравнительной генетики животных ИОГен им. Н.И. Вавилова РАН провели исследование по оценке возможности использования моделей машинного обучения для определения классификации пород и популяций животных по данным анализа микросателлитных локусов. Исследования были проведены для двух видов: лошадь (*Equus caballus*) – классификация пород и северный олень (*Rangifer tarandus*) – идентификация диких и домашних популяций.

Для исследования лошадей использовалась стандартная панель из 17 микросателлитных локусов (AHT4, AHT5, ASB2, ASB17, ASB23, CA425, HMS1, HMS2, HMS3, HMS6, HMS7, HTG4, HTG6, HTG7, HTG10, LEX3, VHL20), одобренная для генотипирования и паспортизации ISAG (International Society for Animal Genetics). Датасет содержал информацию по генотипам микросателлитных локусов более чем 3400 особей 14 пород. Исследование северного оленя проводилось по 16 микросателлитным локусам (BMS1788, RT30, RT1, RT9, C143, RT7, OHEQ, FCB193, RT6, C217, RT24, C32, BMS745, NVHRT16, T40, C276) для популяций дикого и домашнего оленя с соответствующим разбиением на две популяции [36]. Выборки домашних оленей представлены четырьмя зарегистрированными в РФ породами: ненецкой, чукотской, эвенской и эвенкийской, а также двумя экотипами из Тофаларии и Тоджи. Популяции диких оленей были представлены образцами из Магаданской, Амурской, Таймырской, Якутской, Турской, Мурманской и Чукотской областей.

В качестве обучаемой модели была выбрана модель CatBoostClassifier из библиотеки CatBoost от Yandex. Функция потерь – MultiClass. [37] Соотношение обучающей и валидационной выборок составляло 80 к 20, с перетасовкой строк и выравниванием классов (стратификацией).

Наблюдаемые значения precision (табл. 2) находятся в диапазоне от 0.73 (чистокровная верховая порода) до 1 (советская тяжеловозная, одичавшие лошади о. Водный, русская тяжеловозная, нью форест, донская, андалузская), минимальные значения свидетельствуют о высокой доле неправильно отнесенных к данному классу лошадей других пород. Метрика Recall, оценивающая способность улавливать необходимый класс, зависит от доли неверно отрицательных (False Negative), т. е. упущенных объектов и варьирует от 0.5 (советская тяжеловозная) до 1 для большей части пород. Наглядное представление данных в виде матрицы ошибок представлено на рис. 7.

Цветом от синего к желтому показана точность идентификации. Одичавшие лошади о. Водный, ахалтекинская, русская тяжеловозная, фризская, фьорд, донская, арабская породы идентифицируются со 100% вероятностью.

Советская тяжеловозная порода в 50% случаев неверно определяется как ахалтекинская порода лошадей, что может быть вызвано низким уровнем генетического разнообразия и количества приват-аллелей в данной породе. Практический вопрос дифференциации советской тяжеловозной от ахалтекинской породы не был приоритетным, так как это абсолютно разные по фенотипическим показателям лошади; кроме того, ахалтекинская порода является чистокровной породой лошадей с “закрытым генофондом”. При идентификации советской тяжеловозной породы модель не совершает ошибок при сравнении с другими упряжными и тяжеловозными лошадьми и может использоваться при оценке уровня межпородной гибридизации. Порода аппалуза верно определяется в 63% случаев, потери (ошибки) связаны с чистокровной верховой породой лошадей, которая часто используется для улучшения скаковых и верховых качеств в селекции многих пород лошадей. Стоит отметить идентификацию в 100% случаев одичавших лошадей о. Водный и их дифференциацию от породы-основательницы – донской, так как эти породы схожи по экстерьерным показателям и масти, и в данном случае идентификация с использованием молекулярных маркеров является актуальной. В целом обучение модели можно считать успешным, средневзвешенная метрика Accuracy составляет 0.96.

При исследовании популяций северного оленя была обучена модель с целью идентификации породной принадлежности домашнего оленя и их дифференциации от диких популяций. Матрицы ошибок, полученные в ходе обучения модели, представлены на рис. 9 и 10.

Несмотря на то что в оленеводстве используется пастбищный тип содержания и нередки случаи для так называемого “освежения крови” или использования диких самцов в одомашненных стадах, модель успешно разделяет домашних и диких оленей, что предоставляет возможность исследовать процесс доместикации и решать практическую задачу, в частности контролировать браконьерство. При разделении оленей на две группы (дикие и домашние) ошибочно идентифицированы были только две особи из 770. Ошибка модели стремится к 0.

При разделении группы домашних оленей на различные породы и экотипы наблюдается

Таблица 2. Значения метрик качества модели для пород лошадей (*Equus caballus)*

Порода	Сокращенное название	Precision	Recall	*F1*-score
Андалузская	AND	1.00	1.00	1.00
Аппалуза	APP	0.86	0.63	0.73
Арабская	ARA	0.94	0.95	0.95
Донская	DON	1.00	1.00	1.00
Фьорд	FJO	0.99	1.00	1.00
Фризская	FRI	0.99	1.00	1.00
Нью форест	NEW	1.00	0.70	0.82
Русская тяжеловозная	RusT	1.00	1.00	1.00
Русская верховая	Rwp	0.88	0.88	0.88
Стандартбредная	STA	0.98	0.99	0.98
Ахалтекинская (к/з Шаэль)	Shael	0.67	1.00	0.80
Чистокровная верховая	THO	0.73	0.80	0.76
Одичавшие лошади о.Водный	Wild	1.00	1.00	1.00
Советская тяжеловозная	sovt	1.00	0.50	0.67
	Accuracy			0.96
	Macro avg	0.93	0.89	0.90
	Weighted avg	0.96	0.96	0.96

снижение вероятности верно идентифицировать отдельные породы.

Породная идентификация вызывает затруднения с эвенской породой, возможно это связано с отсутствием консолидированной структуры, геногеографическими особенностями, случайными скрещиваниями с другими породами из-за разведения данной породы на огромных территориях: Саха-Якутии, Магаданской области и на севере Камчатского края. Именно эти ареалы разведения северного оленя являются центром миграционных путей северных оленей [38]. Чукотская порода определяется с вероятностью 90% (в 10% случаев идентифицируется как эвенская), эвенкийская – 75%, эвенская – 50%, ненецкая – 98%, экотипы из Тоджи – 89% и Тофаларии – 100%. При этом модель ошибочно относит в 10% случаев эвенскую породу к диким оленям. В остальных случаях ошибки в идентификации домашних оленей как диких отсутствуют, что позволяет использовать модель, а именно метод CatBoost, для решения вопросов идентификации индивидов и различий между домашними и дикими северными оленями.

ЗАКЛЮЧЕНИЕ

Одной из проблем, возникающих в процессе применения искусственного интеллекта в науке, является интерпретация результатов обучения моделей. Сложности возникают из-за того, что модели искусственного интеллекта часто работают на основе сложных алгоритмов и большого количества данных, которые трудно обработать человеку. Глубокое обучение и машинное обучение предоставляют разные подходы к созданию моделей искусственного интеллекта.

Машинное обучение зачастую использует простые модели, такие как деревья решений или линейные модели, которые легче интерпретировать, чем модели глубокого обучения. Обычно они имеют меньше параметров и используют более простой математический аппарат, что делает их понятнее для человека.

Модели глубокого обучения, такие как нейронные сети, могут быть более сложными и трудными для интерпретации, поскольку они имеют множество скрытых слоев и множество нейронов в каждом слое. Они способны аппроксимировать

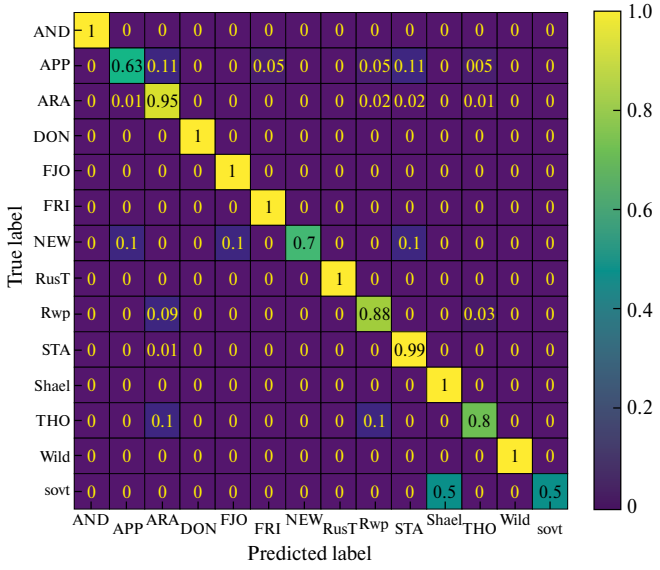

Рис. 8. Матрица ошибок для каждой пары исследуемых пород.

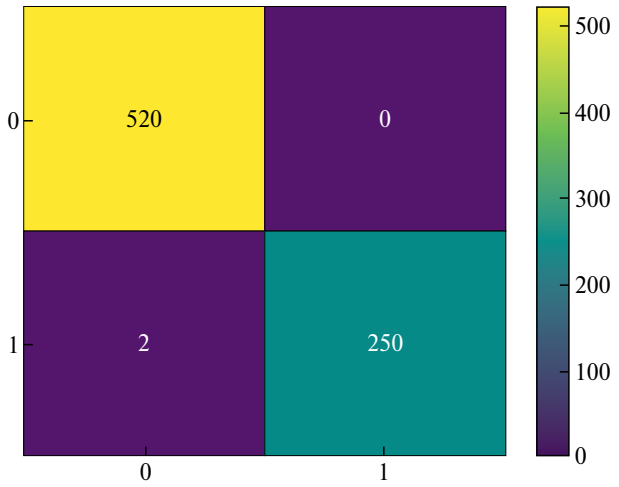

Рис. 9. Матрица ошибок модели бинарной классификации для домашних и диких оленей. 0 – домашние, 1 – дикие.

сложные функции, но это может затруднить понимание того, как они принимают решения.

Таким образом, интерпретируемость является важным фактором при выборе между глубоким и машинным обучением. Если требуется более интерпретируемая модель, то модель машинного обучения может стать лучшим выбором. Если же требуется более мощная модель, способная аппроксимировать сложные зависимости, возможно, стоит обратиться к глубокому обучению.

С точки зрения генетики и селекции животных, равно как и для многих других отраслей науки, существует проблема накопления и создания больших массивов данных. В области животноводства стоит отметить отсутствие достаточных объемов

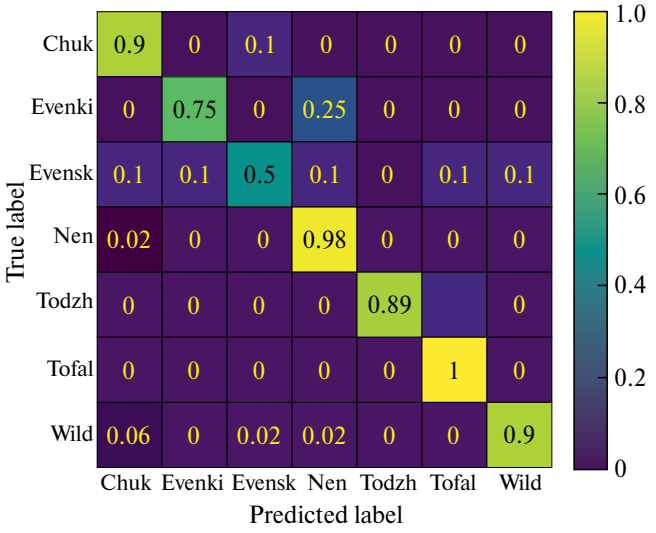

Рис. 10. Матрица ошибок модели для пород домашних оленей и их диких популяций.

фенотипических (зоотехнических и ветеринарных данных), генетических (различные типы маркеров, гены, секвенированные геномы), фотографических и видеоматериалов для создания баз данных, которые могут быть использованы или исследованы при помощи различных методов искусственного интеллекта. При этом достоверность полученных данных напрямую влияет на качество обучаемых моделей.

Методы машинного обучения в генетике и селекции в ближайшем будущем станут основой для решения широкого спектра научных и практических задач, таких как: оценка племенной и генофондной ценности животных, адаптивности, жизнеспособности, психотипа, генетического потенциала, прогноз по их использованию, созданию оптимальных условий содержания и кормления, подбор и отбор, меж- и внутрипородные варианты скрещивания, создания новых пород или селекционных достижений.

В настоящей работе использование микросателлитных баз данных по двум видам одомашненных животных и мультиклассовых моделей машинного обучения позволило с достаточной точностью идентифицировать породную принадлежность животных, а именно породы лошадей, и различить одомашненную и дикую формы северного оленя.

Перспективы использования методов машинного обучения в традиционной, геномной, маркер-зависимой, эпигенетической селекции огромны. Оценка генетической ценности, здоровья животных, их адаптивности к различным агроклиматическим условиям, поиск новых генов, анализ их взаимодействия и т. д. выходят на новый уровень, где сложно переоценить возможности искусственного интеллекта.

Работа выполнена при поддержке гранта Российского научного фонда № 23-16-00059.

Все применимые международные, национальные и/или институциональные принципы ухода и использования животных были соблюдены.

Настоящая статья не содержит каких-либо исследований с участием в качестве объекта людей.

Авторы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Моисеева И.Г., Уханов С.В., Столповский Ю.А.* и др. Генофонды сельскохозяйственных животных. Генетические ресурсы животноводства России. М.: Наука, 2006. 462 с.
2. *Weigel K.A., VanRaden P.M., Norman H.D., Grosu H.* A 100-year review: Methods and impact of genetic selection in dairy cattle-from daughter-dam comparisons to deep learning algorithms // J. Dairy Sci. 2017. V. 100. № 12. P. 10234–10250.
3. *Храброва Л.А., Зайцев А.М., Суходольская И.В. и др.* Проблемы учета и сохранения аборигенных пород лошадей // Аборигенное коневодство России: история, современность, перспективы: Сб. науч. трудов по матер. II Всеросс. научно-практ. конф. с междунар. участием. Мезень, 2018. С. 170–176.
4. *Николаева Э.А., Спасская Н.Н., Столповский Ю.А., Воронкова В.Н.* Структура популяций заводских и вторично одичавших лошадей // Генетические процессы в популяциях: Материалы науч. Конф. с междунар. участием, посвященной 50-летнему юбилею лаборатории популяционной генетики им. Ю.П. Алтухова ИОГен РАН и 85-летию со дня рождения академика Юрия Петровича Алтухова. 2022. С. 45.
5. *Ashley M.V., Dow B.D.* The use of microsatellite analysis in population biology: background, methods and potential applications // Mol. Ecol. Evol.: Approaches and Applications. 1994. P. 185–201.
6. *Столповский Ю.А., Пискунов А.К., Свищева Г.Р.* Геномная селекция. I: Последние тенденции и возможные пути развития // Генетика. 2020. Т. 56. № 9. С. 1006–1017. https://doi.org/10.31857/S0016675820090143
7. *Николаева Э.А., Воронкова В.Н., Политова М.А. и др.* Генетическая структура русской верховой породы лошадей // Генетика. 2023. Т. 59. № 9. С. 1048–1058. https://doi.org/10.31857/S0016675823090096. EDN WUWYIE.
8. *Животовский Л.А.* Генетика природных популяций. Йошкар-Ола: Вертикаль, 2021. 600 с.
9. *Meirmans P.G., Hedrick P.W.* Assessing population structure: FST and related measures // Mol. Ecol. Res. 2011. V. 11. № 1. P. 5–18. https://doi.org/10.1111/j.1755-0998.2010.02927.x
10. *Adamack A.T., Gruber B.* Popgenreport: Simplifying basic population genetic analyses in R // Methods Ecol, Evol, 2014. V. 5. N 4. P. 384-387. https://doi.org/10.1111/2041-210X.12158
11. *Каштанов С.Н., Свищёва Г.Р., Пищулина С.Л. и др.* Географическая структура генофонда соболя (*Martes zibellina* L.): данные анализа микросателлитных локусов // Генетика. 2015. Т. 51. №. 1. С. 78–78. https://doi.org/10.1134/S1022795415010044
12. *Väli Ü., Einarsson A., Waits L., Ellegren H.* To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? // Mol. Ecol. 2008. V. 17. № 17. P. 3808–3817.
13. *Porras-Hurtado L., Ruiz Y., Santos C. et al.* An overview of STRUCTURE: Applications, parameter settings, and supporting software // Front. in Genet. 2013. V. 4. P. 98. https://doi.org/10.3389/fgene.2013.00098
14. *Gronau I., Moran S.* Optimal implementations of UPGMA and other common clustering algorithms // Inform. Proc. Letters. 2007. V. 104. № 6. P. 205–210. https://doi.org/10.1016/j.ipl.2007.07.002
15. *Efron B.* Bootstrap methods: Another look at the jackknife // Ann. Statist. 1979. V. 7. P. 1–26. https://doi.org/10.1214/aos/1176344552
16. *Reich D., Price A., Patterson N.* Principal component analysis of genetic data // Nat. Genet. 2008. V. 40. P. 491–492. https://doi.org/10.1038/ng0508-491
17. *Sievert C.* Interactive Web-based Data Visualization With R, plotly, and shiny. CRC Press, 2020.
18. *Spasskaya N.N., Voronkova V.N., Letarov A.V. et al.* Features of reproduction in an isolated island population of the feral horses of the Lake Manych-Gudilo (Rostov Region, Russia) // App. An. Beh. Sci. 2022. V. 254. https://doi.org/10.1016/j.applanim.2022.105712
19. *Maloy S., Hughes K.* Brenner's Encyclopedia of Genetics. MS, Cambridge: Academic Press,. 2013.
20. *Ruzica Bruvo, Nicolaas K. Michiels, Thomas G. D'Souza, Hinrich Shulenberg.* A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level // Mol. Ecol. 2004. V. 13(7). P. 2101–2106.
21. *Henderson C.R.* Applications of linear models in animal breeding. Guelph, Canada: Univ. Guelph Press. 1984. 462 p.
22. *Отраднов П.И., Рудиянов Д.М., Белоус А.А.* Валидация оценок племенной ценности свиней породы дюрок по признакам кормового поведения // Свиноводство. 2023. № 5. С. 22–26. https://doi.org/10.37925/0039-713X-2023-5-22-26
23. *Сермягин А.А., Белоус А.А., Контэ А.Ф. и др.* Валидация геномного прогноза племенной ценности быков-производителей по признакам молочной

продуктивности дочерей на примере популяции черно-пестрого и голштинского скота // С.-х. биология. 2017. Т. 52. № 6. С. 1148–1156.

24. *Контэ А.Ф., Белоус А.А., Отраднов П.И.* Племенная ценность кормового поведения свиней // Аграрный вестник Урала. 2022. №. 10 (225). С. 44–53.
25. *Nayeri S., Sargolzaei M., Tulpan D.* A review of traditional and machine learning methods applied to animal breeding // Animal Health Res. Rev. 2019. V. 20, P. 31–46. https://doi.org/10.1017/ S1466252319000148
26. *Zhou Z.H. Machine Learning.* London: Springer Nature, 2021. 460 p. https://doi.org/10.1016/ S0034-4257(97)00083-7
27. *Stehman S.V.* Selecting and interpreting measures of thematic classification accuracy // Remote Sensing of Environment. 1997. V. 62. № 1. P. 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7
28. *Erickson B.J., Kitamura F.* Magician's corner: 9. Performance metrics for machine learning models // Radiology: Artificial Intelligence. 2021. V. 3. № 3. https://doi.org/10.1148/ryai.2021200126
29. *Powers D.M.W.* Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation // arXiv preprint arXiv:2010.16061. 2020. https://doi.org/10.48550/arXiv.2010.16061
30. *Sasaki Y.* The truth of the F-measure // Teach Tutor Mater. 2007. V. 1. № 5. P. 1–5.
31. *Penzar D.D., Zinkevich A.O., Vorontsov I.E.* What do neighbors tell about you: The local context of cis-regulatory modules complicates prediction of regulatory variants // Front. Genet. 2019. V. 10. https://doi.org/10.3389/fgene.2019.01078
32. *Михальский А.И., Новосельцева Ж.А.* Применение методов машинного обучения в задачах продуктивного животноводства // Пробл. биол. продуктивных животных. 2018. № 4. С. 98-109. https://doi.org/10.25687/1996-6733. prodanimbiol.2018.3.98-109
33. *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. 2014. doi 10.48550/ arXiv.1409.1556
34. *Jwade S.A., Guzzomi A., Mian A.* On farm automatic sheep breed classification using deep learning // Computers and Electronics in Agriculture. 2019. V. 167. https://doi.org/10.1016/j.compag.2019.105055
35. *Batic D., Culibrk D.* Identifying individual dogs in social media images // arXiv:2003.06705. 2019.
36. *Столповский Ю.А., Бабаян О.В., Каштанов С.Н. и др.* Генетическая оценка пород северного оленя (*Rangifer tarandus*) и их дикого предка с помощью новой панели STR-маркеров // Генетика. 2020. Т. 56. № 12. С. 1409–1425. https://catboost.ai/en/docs/concepts/loss-functions-multiclassification#usage-information
37. *Южаков А.А., Мухачев А.Д., Лайшев К.А.* Породы и проблемы селекции северных оленей России. М.: Наука, 2023. 165 с.

Artificial Intelligence and Classical Methods in Animal Genetics and Breeding

A. D. Soloshenkov[1,2,*], E. A. Soloshenkova[1], M. T. Semina[1], N. N. Spasskaya[3], V. N. Voronkova[1], Y. A. Stolpovky[1]

[1]*Vavilov Institute of General Genetic, Russian Academy of Sciences Moscow, 119991 Russia*
[2]*Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, Moscow, 127434 Russia*
[3]*Zoo museum of Moscow State University, Moscow, 125009 Russia*
**e-mail: alesol@rgau-msha.ru*

The article analyses basic methods of population genetics and animal breeding, as well as mathematical methods of machine learning used in animal breeding. The training of cat boost library models was carried out on the example of two domesticated species – domestic horse (*Equus caballus*) and reindeer (*Rangifer tarandus*). Data from microsatellite panels of 16 and 17 loci, respectively, were used to train the model using data on domesticated and wild reindeer, European and Russian horse breeds. The standard indicators: accuracy, precision, recall and f1 were calculated to determine the success of the model. Confusion matrices were constructed. New possibilities of identification of animal breed affiliation were shown.

Keywords: machine learning, neural networks, animal husbandry, horse, *Equus caballus*, reindeer, *Rangifer tarandus*, microsatellite analysis, genetic diversi