Научная статья
УДК 004.89
DOI: 10.14529/ctcr220305

ПРОГНОЗИРОВАНИЕ КОЛИЧЕСТВЕННЫХ ХАРАКТЕРИСТИК МОЛОКА НА ОСНОВЕ ИНФРАКРАСНОЙ СПЕКТРОСКОПИИ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Л.В. Легашев[1], *silentgir@gmail.com, https://orcid.org/0000-0001-6351-404X*
И.П. Болодурина[1, 2], *prmat@mail.osu.ru, https://orcid.org/0000-0003-0096-2587*
Л.С. Гришина[1], *grishina_ls@inbox.ru, https://orcid.org/0000-0003-2752-7198*
И.А. Лашнева[3], *lashnevaira@gmail.com*
А.А. Сермягин[3], *alex_sermyagin85@mail.ru, https://orcid.org/0000-0002-1799-6014*
[1] *Оренбургский государственный университет, Оренбург, Россия*
[2] *Федеральный научный центр биологических систем и агротехнологий Российской академии наук, Оренбург, Россия*
[3] *Федеральный исследовательский центр животноводства – ВИЖ имени академика Л.К. Эрнста, Москва, Россия*

Аннотация. Спектроскопия среднего инфракрасного диапазона с преобразованием Фурье представляет собой быстрый и дешевый способ анализа проб коровьего молока для определения содержания жира, белка, лактозы и других количественных и качественных показателей молока. Современные инструменты анализа данных позволяют выявить наиболее значимые зависимости между различными парами количественных и качественных признаков состава молока. **Цель исследования.** Выполнить прогнозирование ряда ключевых признаков состава молока коров с использованием данных инфракрасной спектроскопии для изучения точности разработанной математической модели. **Методы.** Работу проводили в зимний период 2022 года на базе экспериментального стада голштинизированного черно-пестрого скота (Краснодарский край). Анализ компонентов молока осуществляли с использованием автоматического анализатора MilkoScan (FOSS) с применением метода инфракрасной спектроскопии путем выгрузки полученных спектров при анализе состава сырого молока. Исследованы 23 показателя количественного состава молока: массовая доля жира, белка (истинного и общего), лактозы, СОМО (сухого обезжиренного молочного остатка), сухого вещества, казеина, следы ацетона и бета-гидроксибутирата, мочевина, точка замерзания, кислотность молока, миристиновая, пальмитиновая, стеариновая, олеиновая жирные кислоты (ЖК), длинноцепочечные ЖК, среднецепочечные ЖК, короткоцепочечные ЖК, мононенасыщенные и полиненасыщенные ЖК, насыщенные ЖК, трансизомеры жирных кислот. Рассмотрены методы на основе линейной регрессии (Linear Regression), подходы к регуляризации модели линейной регрессии (Ridge, Lasso и ElasticNet), а также полиномиальная регрессия, метод частичной регрессии (PLSRegression) и метод Байесовской регрессии для задачи прогнозирования ключевых признаков состава молока. Реализован метод снижения размерности данных инфракрасной спектроскопии на основе алгоритма случайного перебора считывания по длине окна и выделены наиболее значимые признаки. **Результаты.** Разработаны модели прогнозирования шести основных показателей качества молока – массовая доля жира ('Fat'), массовая доля казеина ('Cas.B'), жирных кислот – миристиновой ('C14:0') и олеиновой ('C18:1'), мононенасыщенных ('MUFA') и полиненасыщенные жирных кислот ('PUFA') – со средней абсолютной ошибкой, не превышающей 0,016. **Заключение.** Результаты, полученные в ходе проведенного исследования, позволят в дальнейшем улучшить предиктивную способность уравнения для определения качества, состава молока по новым селекционным признакам молочной продуктивности, снизить издержки анализа и проводить контроль за состоянием здоровья животных на ранних стадиях.

Ключевые слова: прогнозирование, показатели состава молока, инфракрасная спектроскопия, машинное обучение, регрессия

Благодарности: Исследование выполнено за счет гранта Российского научного фонда (проект № 21-76-20046) в части исследования расширенного компонентного состава молока коров.

Для цитирования: Прогнозирование количественных характеристик молока на основе инфракрасной спектроскопии с применением методов машинного обучения / Л.В. Легашев, И.П. Болодурина, Л.С. Гришина и др. // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2022. Т. 22, № 3. С. 47–56. DOI: 10.14529/ctcr220305

Original article
DOI: 10.14529/ctcr220305

PREDICTION OF MILK QUANTITATIVE TRAITS BASED ON INFRARED SPECTROSCOPY USING MACHINE-LEARNING METHODS

L.V. Legashev[1], *silentgir@gmail.com, https://orcid.org/0000-0001-6351-404X*
I.P. Bolodurina[1, 2], *prmat@mail.osu.ru, https://orcid.org/0000-0003-0096-2587*
L.S. Grishina[1], *grishina_ls@inbox.ru, https://orcid.org/0000-0003-2752-7198*
I.A. Lashneva[3], *lashnevaira@gmail.com*
A.A. Sermyagin[3], *alex_sermyagin85@mail.ru, https://orcid.org/0000-0002-1799-6014*
[1] *Orenburg State University, Orenburg, Russia*
[2] *Federal Research Centre of Biological Systems and Agrotechnologies of the Russian Academy of Sciences, Orenburg, Russia*
[3] *Federal Research Center for Animal Husbandry named after Academy Member L.K. Ernst, Moscow, Russia*

Abstract. Fourier transform mid-infrared spectroscopy is a fast and cheap way to analyze cow's milk samples to determine fat, protein, lactose and other quantitative and qualitative indicators of milk quality. Modern tools for data analysis will reveal the relationship between different pairs of quantitative and qualitative characteristics of milk. **Purpose of the study.** Perform predictions on some key milk quality traits based on infrared spectroscopy data to study the accuracy of the developed mathematical model. **Methods.** The work was carried out in the winter period of 2022 on the basis of an experimental herd of Holsteinized black-and-white cattle (Krasnodar Territory). The analysis of milk traits was carried out with an automatic analyzer MilkoScan (FOSS) using the method of infrared spectroscopy by unloading the obtained spectra when analyzing the composition of raw milk. 23 indicators of the quantitative milk traits were studied: mass fraction of fat, protein (true and total), lactose, DSMR (dry skimmed milk residue), dry matter, casein, traces of acetone and beta-hydroxybutyrate, urea, freezing point, acidity of milk, myristic, palmitic, stearic, oleic fatty acids (FA), long-chain fatty acids, medium-chain fatty acids, short-chain fatty acids, monounsaturated and polyunsaturated fatty acids, saturated fatty acids, trans fatty acids. Methods based on linear regression, approaches to the regularization of the linear regression model (Ridge, Lasso and ElasticNet), as well as polynomial regression, the partial regression method (PLSRegression) and the Bayesian regression method for the problem of predicting key features of milk traits were considered. A method for reducing the dimensionality of infrared spectroscopy data is implemented based on the algorithm of random search of readings along the length of the window, and the most significant features are identified. **Results.** Models have been developed for predicting six main indicators of milk quality – mass fraction of fat ('Fat'), mass fraction of casein ('Cas.B'), fatty acids – myristic ('C14:0') and oleic ('C18: 1'), monounsaturated ('MUFA') and polyunsaturated fatty acids ('PUFA') – with an average absolute error not exceeding 0,016. **Conclusion.** The results obtained in the course of the study will further improve the predictive ability of the equation for determining the quality and composition of milk according to new breeding traits of milk productivity, reduce analysis costs and monitor the health of animals at an early stage.

Keywords: prediction, milk quality traits, infrared spectroscopy, machine learning, regression

Acknowledgments: The study was supported by a grant from the Russian Science Foundation (project no. 21-76-20046) in terms of studying the expanded component composition of cows' milk.

For citation: Legashev L.V., Bolodurina I.P., Grishina L.S., Lashneva I.A., Sermyagin A.A. Prediction of milk quantitative traits based on infrared spectroscopy using machine-learning methods. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2022;22(3):47–56. (In Russ.) DOI: 10.14529/ctcr220305

Введение

В настоящее время модели искусственного интеллекта находят свое применение в области безопасности, медицины, образования, сельского хозяйства и других. Современные методы интеллектуального анализа данных активно используются для прогнозирования урожайности на полях, мониторинга состояния скота на фермах и т. д. Одно из достоинств методов машинного обучения в сельскохозяйственных процессах состоит в своевременной оценке качества изготавливаемой продукции. Например, автоматические анализаторы инфракрасной спектроскопии [1] активно применяются для анализа количественного состава молока по его компонентам [2, 3]. Более детальное исследование данного процесса позволит вести мониторинг здоровья животных, выявлять воспалительные процессы и контролировать качество производимого сырого молока и вырабатываемой из него молочной продукции.

В этой связи актуальной является задача прогнозирования ключевых показателей состава молока современными методами машинного обучения. Результаты обучения моделей на данных инфракрасной спектроскопии позволят описать формальную математическую зависимость между спектрограммой молока и заданным критерием качества в форме функциональной зависимости. Результаты проведенного исследования позволят в дальнейшем улучшить селекционный контроль качества молока и молочной продукции, снизить издержки анализа и прогнозировать различные потенциальные функциональные нарушения на ранних стадиях, а также в перспективе проводить анализ технологических свойств молока индивидуально для каждого животного (сыропригодность и др.).

1. Обзор исследований

Данные, полученные с помощью инфракрасной спектроскопии молока, используют для различных задач прогнозирования. В частности, в работе [4] представлен обзор использования спектроскопии в качестве инструмента фенотипирования признаков молока, например, для прогнозирования основного минерального состава, как это изложено в [5]. В работе [6] проводят исследование по прогнозированию статуса беременности коров на основе методов глубокого обучения. В исследовании [7] определяют статус туберкулеза у крупного рогатого скота на основе методов глубокого обучения. В статьях [8, 9] прогнозируют качественные характеристики молока с использованием статистических методов машинного обучения. В работе [10] на базе устройств интернета вещей исследуют фальсификацию молочной продукции с использованием машинного обучения. Аналогичные исследования фальсификации кокосового молока проводят в работе [11]. В статье [12] используется ансамблевый классификатор Stacking и многослойная нейронная сеть с прямой связью для ежедневного прогнозирования метаболического профиля крови у молочного скота. Авторы исследования [13] сравнивают метод частичной регрессии (PLSR), методы с частными коэффициентами наименьших квадратов (PLS) в сочетании с линейной и полиномиальной регрессией опорных векторов (PLS + SVR), а также метод с использованием PLS и искусственной нейронной сети с одним скрытым слоем для прогнозирования содержания лактоферрина в коровьем молоке. В работе [14] использовался метод главных компонент для снижения размерности входных данных и многослойный персептрон с двумя скрытыми слоями с целью прогнозирования параметров качества коровьего молока по спектральным данным. Метод частичной регрессии PLSR использовался в исследовании [15] для прогнозирования потребления сухого вещества корма (конверсии) на основе данных инфракрасной спектроскопии молока.

Таким образом, обзор проведенных исследований показал, что методы машинного обучения активно применяются для решения широкого спектра практических задач на основе данных спектрального анализа молока. В то же время разработка регрессионных уравнений моделей для прогнозирования функционального состояния животных по спектрам сырого молока позволит открыть новые возможности управления и улучшения биологических качеств молочного скота на примере российской популяции черно-пестрой и голштинской породы.

2. Постановка задачи прогнозирования показателей при спектроскопии молока

Задача построения модели прогнозирования основных показателей качества молока может быть отнесена к задаче множественной регрессии (обучение с учителем), которая позволяет вос-

становить зависимость между показателями качества Y и фиксируемыми признаками инфракрасной спектроскопии по волновым точкам $X = \{x_1, x_2, \dots, x_n\}$. Цель регрессионного анализа состоит в том, чтобы оценить значение непрерывной выходной переменной Y по значениям входных переменных X.

Формальная **математическая постановка задачи регрессии** имеет следующий вид.

Пусть дано X – множество признакового описания объектов ($X \in \mathbb{R}^n$), Y – множество ответов ($Y \in \mathbb{R}^m$). Задача состоит в построении на основе $X^l = (x_i, y_i)_{i=1}^l$ – обучающей выборки неизвестной зависимости $a: X \to Y$, для которой справедливо $y_i = y(x_i), i = 1, .., l$.

Исходные данные. Был проанализирован 521 образец молока от 196 коров, собранный в период регистрации (учета) молочной продуктивности коров в январе 2022 года на базе экспериментального хозяйства Краснодарского края. Количество волновых точек в средней инфракрасной области спектроскопии с различной степенью поглощения вещества составило n = 1060. Данные разбиты на две выборки: в первой содержались только результаты спектроскопии (столбцы с 240-го по 1299-й), во второй – количественные показатели молока, включая суточный удой коровы.

По изученным количественным показателям молока построена матрица корреляции (рис. 1). Первые 15 записей пар признаков с наибольшей положительной/отрицательной корреляцией представлены в табл. 1.

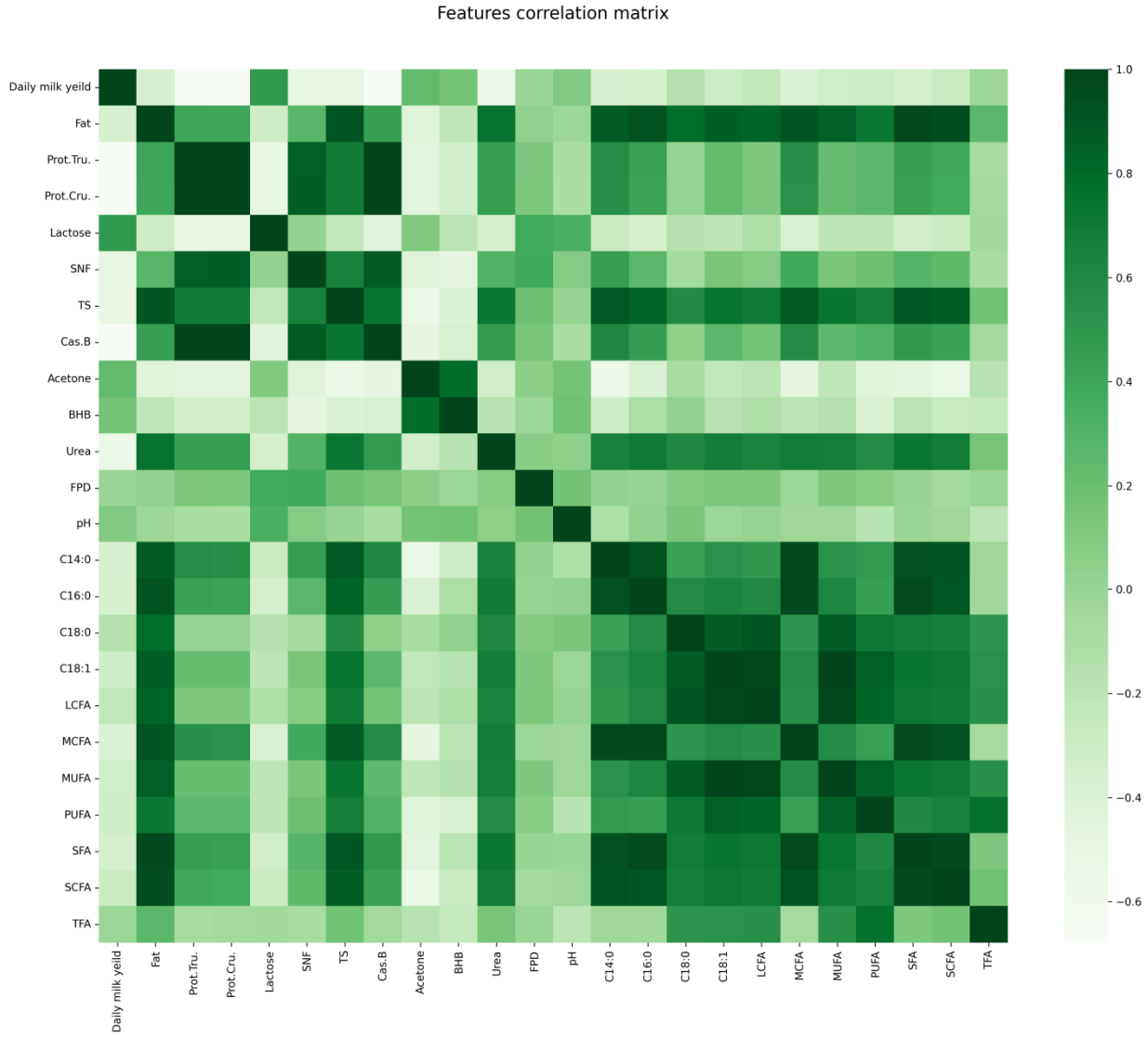

Рис. 1. Матрица корреляции признаков качественных и количественных показателей молока
Fig. 1. Correlation matrix of qualitative and quantitative milk quality traits features

Таблица 1

Первые 15 записей пар признаков с наибольшей положительной/отрицательной корреляцией

Table 1

The first 15 records of feature pairs with the highest positive/negative correlation

Первый признак	Второй признак	Корреляция	Первый признак	Второй признак	Корреляция
Prot.Tru.	Prot.Cru.	0,999	Daily milk yeild	Prot.Cru.	–0,679
C18:1	MUFA	0,997	Daily milk yeild	Prot.Tru.	–0,676
Prot.Tru.	Cas.B	0,995	Daily milk yeild	Cas.B	–0,665
Prot.Cru.	Cas.B	0,995	Daily milk yeild	Urea	–0,551
C18:1	LCFA	0,987	TS	Acetone	–0,537
LCFA	MUFA	0,984	Acetone	C14:0	–0,534
C16:0	MCFA	0,980	Acetone	SCFA	–0,511
Fat	SFA	0,979	Daily milk yeild	TS	–0,496
SFA	SCFA	0,975	Prot.Tru.	Lactose	–0,494
C14:0	MCFA	0,973	Prot.Cru.	Lactose	–0,489
MCFA	SFA	0,967	Daily milk yeild	SNF	–0,482
C16:0	SFA	0,966	Cas.B	Acetone	–0,482
Fat	SCFA	0,956	SNF	BHB	–0,481
C14:0	SFA	0,939	Acetone	MCFA	–0,463
C14:0	C16:0	0,936	Prot.Tru.	Acetone	–0,453

Корреляционный анализ позволяет сделать выводы о том, что множество показателей тесно связаны друг с другом и, например, высокая точность при прогнозировании признака по жирным кислотам – 'C14:0' (миристиновая ЖК) гарантирует аналогичный порядок точности признака для насыщенных жирных кислот 'SFA' (корреляция 0,9394). В связи с этим в рамках данного исследования выделены ключевые показатели качества молока $\bar{Y} \subset Y = \{'\mathrm{Fat}', '\mathrm{Cas.B}', '\mathrm{C14{:}0}', '\mathrm{C18{:}1}', '\mathrm{MUFA}', 'PUFA'\}$, характеризующие массовую долю жира, процент казеина, жирные кислоты – миристиновая (насыщенная ЖК) и олеиновая (ненасыщенная ЖК), мононенасыщенные и полиненасыщенные жирные кислоты соответственно. Казеин молока коров представляет особый интерес для выработки сыра и творога. Селекционный контроль у животных миристиновой ЖК и ряда насыщенных кислот молока связан с общим выходом жира, обменом веществ и также с потенциальной оценкой продукции метана, выделяемого с производными жизнедеятельности скота. Содержание олеиновой ЖК, а также моно- и полиненасыщенных жирных кислот обуславливает выход масла, связано с технологическими свойствами и органолептическими особенностями молока, продуцируемого животными, а также с их фертильностью (способностью к устойчивому воспроизводству потомства). В этой связи были выбраны именно эти показатели, как отвечающие и имеющие наибольший интерес в среде селекционеров для возможностей прогнозирования качественных характеристик животных.

Наиболее тесные корреляционные взаимосвязи ($r > 0,9$) отмечались для признаков со схожей природой синтеза (образования) в молочной железе коровы: белков молока (истинного и общего белка с казеином, а также между собой) и жирных кислот с учетом длины углеродной цепи и степени насыщения. В то же время взаимосвязь между суточным удоем как основной производной жизнедеятельности и использования коров и компонентами молока – белком, жиром, некоторыми жирными кислотами, следами метаболитов характеризовалась отрицательными значениями ($–0,453 < r < –0,679$) (см. табл. 1).

На рис. 2 представлены графики рассеяния и графики распределения зависимости между признаками жира 'Fat' (верхний ряд) и белка 'Prot.Cru.' (нижний ряд) и группы признаков из ацетона 'Acetone', мочевины 'Urea' и бета-гидроксибутирата 'BHB'. С увеличением жирности молока наблюдалось повышение концентрации мочевины в нем, тогда как следы ацетона и БГБ имели плавную отрицательную динамику. Подобная закономерность отмечена для массовой доли общего белка по мочевине и следам метаболитов в молоке коров.

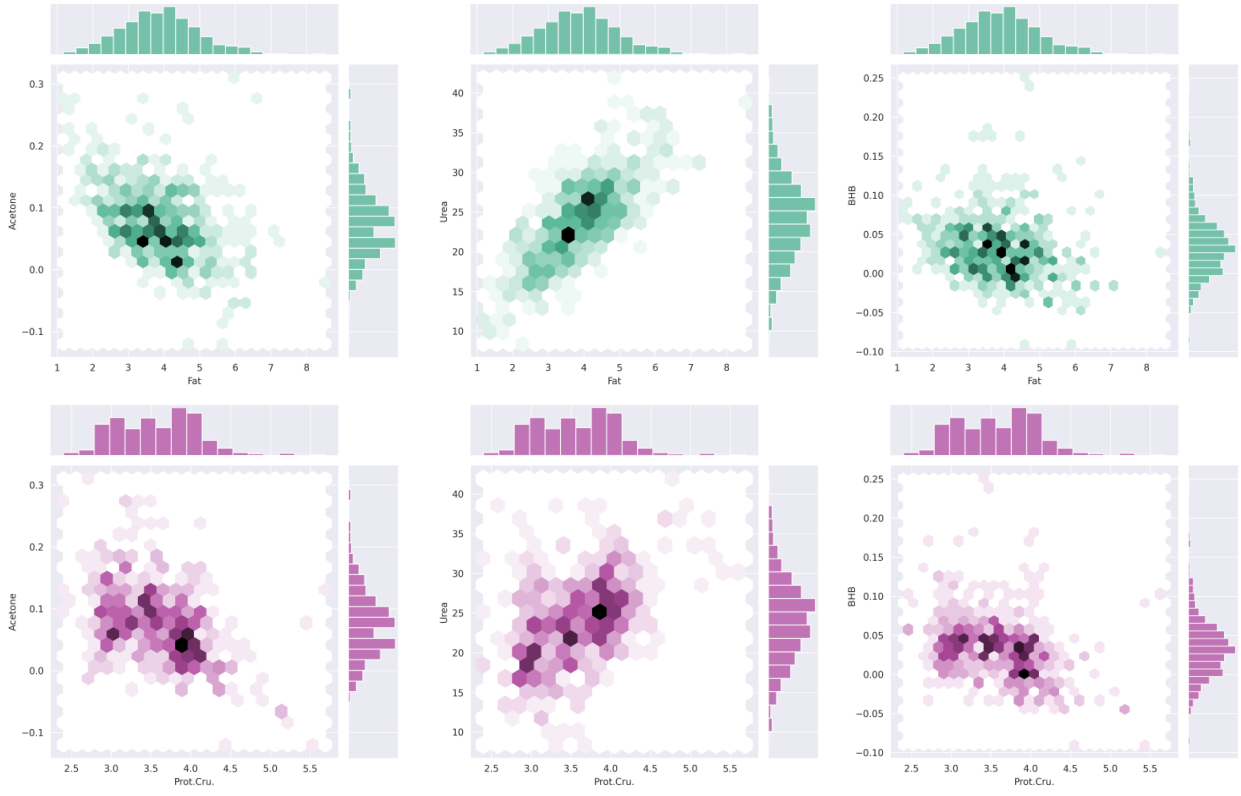

Рис. 2. Графики рассеяния и распределения между признаками жира и белка и группы признаков из ацетона, мочевины и бета-гидроксибутирата
Fig. 2. Jointplots for the features of fat and protein, and feature sets of acetone, urea and beta-hydroxybutyrate

Описательная статистика для шести прогнозируемых признаков 'Fat', 'Cas.B', 'C14:0', 'C18:1', 'MUFA' and 'PUFA' представлена на рис. 3.

	Fat	Cas.B	C14:0	C18:1	MUFA	PUFA
count	521.000000	521.000000	521.000000	521.000000	521.000000	521.000000
mean	3.906507	2.915528	0.412555	1.049572	1.009850	0.116973
std	1.101313	0.465043	0.110201	0.307287	0.287715	0.029744
min	1.180000	1.810000	0.162000	0.398000	0.428000	0.049000
25%	3.160000	2.520000	0.329000	0.852000	0.827000	0.096000
50%	3.880000	2.920000	0.414000	1.006000	0.974000	0.115000
75%	4.580000	3.250000	0.482000	1.195000	1.146000	0.132000
max	8.530000	4.650000	0.898000	2.786000	2.676000	0.261000

Рис. 3. Описательная статистика по шести прогнозируемым признакам
Fig. 3. Descriptive statistics on six predictive traits

Оценивать качество регрессии можно различными способами. Наиболее типичными мерами качества в задачах регрессии являются:

1) средняя абсолютная ошибка (англ. Mean Absolute Error, MAE):

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|\mathrm{y_i} - \widehat{\mathrm{y}_\mathrm{i}}|; \quad (1)$$

2) корень из средней квадратичной ошибки (англ. Root Mean Squared Error, RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}; \quad (2)$$

3) коэффициент детерминации R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \quad (3)$$

Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Фактически данная мера качества – это нормированная среднеквадратичная ошибка. Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным и поэтому более чувствителен к выбросам.

3. Прогнозирование показателей молока современными методами машинного обучения

Исходные данные разбиты на обучающую и тестовую выборки в соотношении 4 : 1. Выполнена линейная регрессия методом наименьших квадратов с помощью LinearRegression библиотеки *sklearn*. Результаты регрессии оценивались метриками (1)–(3): среднеквадратичная ошибка модели (Root Mean Square Error, RMSE) и средняя абсолютная ошибка модели (Mean Absolute Error, MAE), а также коэффициент детерминации R^2. Результаты прогнозирования представлены в табл. 2.

Таблица 2

Метрики модели машинного обучения для шести прогнозируемых признаков

Table 2

Machine learning model metrics for six predictive features

Прогнозируемый признак	Область определения	RMSE	MAE	R^2
Fat	[1,18; 8,53]	0,0799	0,0645	0,9946
Cas.B	[1,81; 4,65]	0,1164	0,0925	**0,9372**
C14:0	[0,16; 0,90]	0,0154	0,0124	0,9787
C18:1	[0,39; 2,79]	0,0443	0,0357	0,9795
MUFA	[0,42; 2,68]	0,0453	0,0361	0,9758
PUFA	[0,04; 0,27]	0,0121	0,0095	**0,8455**

Результаты оценки качества прогнозных моделей по различным показателям выявили, что прогнозируемые признаки 'Cas.B' и 'PUFA' имеют наиболее слабую долю дисперсии. При этом остальные показатели близки к единице, т. е. модель хорошо объясняет данные.

На следующем этапе исследования реализован алгоритм снижения размерности данных инфракрасной спектроскопии. Сравнительно небольшое количество исходных данных позволяет нам использовать для этого алгоритм случайного перебора окна. В зависимости от масштаба области определения признаков использованы различные метрики качества для подбора окна волнового спектра. Размер отрезка варьировался в интервале [50; 400], количество итераций составляло 5000, в качестве целевой функции f_{opt} использовалась одна из метрик RMSE или MAE. Результаты работы алгоритма для метода линейной регрессии представлены в табл. 3. На рис. 4 представлен пример линейного графика зависимости спектра поглощения от длины волны, а также близкие к оптимальному окна волнового спектра для всех шести прогнозируемых признаков. Использование оптимального окна спектра для признака 'Cas.B' позволило улучшить долю дисперсии до значения $R^2 = 0{,}9966$, для признака ' PUFA' позволило улучшить долю дисперсии до значения $R^2 = 0{,}9686$.

Таблица 3

Результаты работы алгоритма случайного перебора размера окна волнового спектра для метода линейной регрессии

Table 3

Results of the random search algorithm of finding the wave spectrum window length for the linear regression method

Прогнозируемый признак	Метрика	f_{opt}	Окно волнового спектра
Fat	RMSE	0,0209	[293; 453]
Cas.B	RMSE	0,0187	[344; 396]
C14:0	MAE	0,0041	[252; 462]
C18:1	MAE	0,0130	[251; 456]
MUFA	MAE	0,0119	[240; 469]
PUFA	MAE	0,0043	[246; 464]

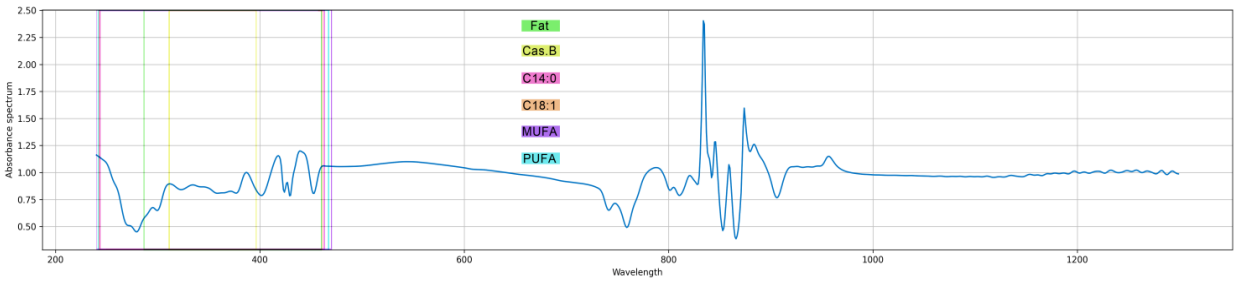

Рис. 4. Линейный график зависимости спектра поглощения от длины волны, близкие к оптимальным окна длины волны при прогнозировании признаков
Fig. 4. Linear plot of absorbance spectrum vs. wavelength, close-to-optimal wavelength windows when predicting features

Кроме того, реализовано сравнение эффективности модели LinearRegression с рядом других моделей машинного обучения. Рассмотрены подходы к регуляризации модели линейной регрессии (Ridge, Lasso и ElasticNet), а также расширение линейной модели базисными функциями (полиномиальная регрессия), методом частичной регрессии (PLSRegression) и наименьших квадратов, методом байесовской регрессии. Результаты прогнозирования представлены в табл. 4.

Таблица 4

Ошибки моделей машинного обучения для шести прогнозируемых признаков

Table 4

Machine learning model errors for six predictive features

	Fat		Cas.B		C14:0	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge	0,024186 (alpha=1e-03)	0,019668 (alpha=1e-03)	0,028024 (alpha=1e-03)	0,02276 (alpha=1e-03)	0,007418 (alpha=1e-06)	0,00595 (alpha=1e-06)
Lasso	0,027916 (alpha=1e-05)	0,021313 (alpha=1e-05)	0,02149 (alpha=1e-05)	0,016037 (alpha=1e-05)	0,016608 (alpha=1e-06)	0,013043 (alpha=1e-06)
ElasticNet	0,038695	0,029758	0,027905	0,021019	0,02793	0,02098
Полиномиальная регрессия (degree = 2)	0,031258	0,023921	0,03508	0,027013	0,035539	0,027914
BayesianRidge	0,025172	0,019832	0,027887	0,022553	0,025332	0,019092
PLSRegression (n_components = 2)	0,083546	0,068356	0,161498	0,126435	0,042118	0,033418
	C18:1		MUFA		PUFA	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Ridge	0,018545 (alpha=1e-06)	0,01403 (alpha=1e-06)	0,017819 (alpha=1e-06)	0,01379 (alpha=1e-06)	0,005859 (alpha=1e-06)	0,0045199 (alpha=1e-06)
Lasso	0,047717 (alpha=1e-06)	0,03648 (alpha=1e-06)	0,044095 (alpha=1e-06)	0,033121 (alpha=1e-06)	0,011433 (alpha=1e-06)	0,00941 (alpha=1e-06)
ElasticNet	0,10079	0,07566	0,094107	0,071053	0,0145849	0,0114307
Полиномиальная регрессия (degree = 2)	0,12155	0,09059	0,118251	0,0879466	0,02407	0,019333
BayesianRidge	0,07354	0,057898	0,0747286	0,0574054	0,020836	0,016419
PLSRegression (n_components = 2)	0,13965	0,10865	0,131281	0,1036154	0,0199687	0,015649

Линейная регрессия с регуляризацией Ridge показывает лучшие результаты по большинству прогнозируемых признаков со средней ошибкой MAE = 0,0116, за исключением казеина, где ошибка MAE составила 0,0227. Для показателя 'Cas.B' наиболее высокую точность при прогнозировании продемонстрировал метод линейной регрессии с регуляризацией Lasso (MAE = 0,0160).

В результате проведенных исследований построены шесть моделей линейной регрессии по каждому из основных показателей качества молока, включающих в среднем 200 признаков инфракрасной спектроскопии по волновым точкам на интервале [247; 463] с весовыми коэффициентами, подобранными на основе методов, описанных выше. Данные модели могут быть использованы на практике для оценки качества молока.

Заключение

Таким образом, в рамках данного исследования осуществлено прогнозирование по шести ключевым признакам состава молока на основе данных инфракрасной спектроскопии. Исходные данные о компонентах сырого молока взяты за январь 2022 года в экспериментальном стаде голштинизированного черно-пестрого скота. Анализ молока осуществляли с использованием автоматического анализатора MilkoScan (FOSS, Дания) на основе экспресс-метода инфракрасной спектроскопии. Исследована эффективность подходов по прогнозированию показателей состава молока на основе линейной регрессии (Linear Regression) и подходов к ее регуляризации (Ridge, Lasso и ElasticNet), полиномиальной регрессии, метода частичной регрессии (PLSRegression), а также метода байесовской регрессии. В качестве ключевых показателей состава молока выступали: массовая доля жира ('Fat'), массовая доля казеина ('Cas.B'), жирные кислоты – насыщенная миристиновая ('C14:0') и ненасыщенная олеиновая ('C18:1'), сумма мононенасыщенных ('MUFA') и сумма полиненасыщенных жирных кислот ('PUFA'). Кроме того, реализован метод снижения размерности данных инфракрасной спектроскопии молока на основе алгоритма случайного перебора длины окна. Построенные прогнозные модели показателей состава молока показали высокую эффективность при экспериментальном исследовании со средней абсолютной ошибкой, не превышающей 0,016.

С точки зрения применимости методов машинного обучения в сельскохозяйственной биологии они представляют собой новое направление научно-практических исследований по совершенствованию и повышению точности оценки фенотипа животных. Получаемые результаты по количественному составу молока коров (белковая и жировая фракции) уже сейчас составляют важный элемент в разведении животных для целей поиска новых селекционных критериев. Накопленный большой массив данных, как мы полагаем, позволит в ближайшее время перейти к оценке качественных биологических характеристик объектов в молочном скотоводстве для улучшения технологических свойств молока как сырья для переработки и повышения способности предсказания (прогнозирования) функционального состояния животных (признаки здоровья, долголетнего использования, воспроизводительные качества).

Список литературы/References

1. De Marchi M., Penasa M., Cecchinato A., Mele M., Secchiari P., Bittante G. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. *Animal*. 2011;5(10):1653–1658. DOI: 10.1017/S1751731111000747

2. Filipejová T., Kováčik J., Kirchnerová K., Foltýs V. Changes in milk composition as a result of metabolic disorders of dairy cows. *Potravinárstvo*. 2011;5(1):10–16.

3. Zaalberg R.M., Shetty N., Janss L., Buitenhuis A.J. Genetic analysis of Fourier transform infrared milk spectra in Danish Holstein and Danish Jersey. *Journal of Dairy Science*. 2019;102(1): 503–510. DOI: 10.3168/jds.2018-14464

4. M. De Marchi, V. Toffanin, M. Cassandro, M. Penasa Invited review: mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science*. 2014;97(3):1171–1186. DOI: 10.3168/jds.2013-6799

5. Visentin G. et al. Phenotypic characterisation of major mineral composition predicted by mid-infrared spectroscopy in cow milk. *Italian Journal of Animal Science*. 2018;17(3):549–556. DOI: 10.1080/1828051X.2017.1398055

6. Brand W. et al. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *Journal of Dairy Science*. 2021;104(4):4980–4990. DOI: 10.3168/jds.2020-18367

7. Denholm S.J. et al. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *Journal of Dairy Science*. 2020;103(10):9355–9367. DOI: 10.3168/jds.2020-18328

8. Frizzarin M. et al. Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *Journal of Dairy Science*. 2021;104(7):7438–7447. DOI: 10.3168/jds.2020-19576

9. Frizzarin M. et al. Mid infrared spectroscopy and milk quality traits: A data analysis competition at the “International workshop on spectroscopy and chemometrics 2021”. *Chemometrics and Intelligent Laboratory Systems*. 2021;219:1–9. DOI: 10.1016/j.chemolab.2021.104442

10. Sowmya N., Ponnusamy V. Development of spectroscopic sensor system for an IoT application of adulteration identification on milk using machine learning. *IEEE Access*. 2021;9:53979–53995. DOI: 10.1109/ACCESS.2021.3070558

11. Al-Awadhi M.A., Deshmukh R.R. Detection of Adulteration in Coconut Milk using Infrared Spectroscopy and Machine Learning. In: *2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI)*. IEEE; 2021. P. 1–4. DOI: 10.1109/MTICTI53925.2021.9664764

12. Giannuzzi D. et al. In-line near-infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. *Scientific Reports*. 2022;12(1):1–13. DOI: 10.1038/s41598-022-11799-0

13. Soyeurt H. et al. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. *Journal of dairy science*. 2020;103(12):11585–11596. DOI: 10.3168/jds.2020-18870

14. Muñiz R., Cuevas-Valdés M., de la Roza-Delgado B. Milk quality control requirement evaluation using a handheld near infrared reflectance spectrophotometer and a bespoke mobile application. *Journal of Food Composition and Analysis*. 2020;86:1–8. DOI: 10.1016/j.jfca.2019.103388

15. Dórea J. R. R. et al. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science*. 2018;101(7):5878–5889. DOI: 10.3168/jds.2017-13997

Информация об авторах

Легашев Леонид Вячеславович, канд. техн. наук, ведущий научный сотрудник лаборатории цифровых решений и аналитики больших данных, Оренбургский государственный университет, Оренбург, Россия; silentgir@gmail.com.

Болодурина Ирина Павловна, д-р техн. наук, проф., заведующий кафедрой прикладной математики, Оренбургский государственный университет; Федеральный научный центр биологических систем и агротехнологий Российской академии наук, Оренбург, Россия; ipbolodurina@yandex.ru.

Гришина Любовь Сергеевна, преподаватель кафедры прикладной математики, Оренбургский государственный университет, Оренбург, Россия; zabrodina97@inbox.ru.

Лашнева Ирина Алексеевна, аспирант, младший научный сотрудник отдела популяционной генетики и генетических основ разведения животных, Федеральный исследовательский центр животноводства – ВИЖ имени академика Л.К. Эрнста, Москва, Россия; lashnevaira@gmail.com.

Сермягин Александр Александрович, канд. с.-х наук, ведущий научный сотрудник, заведующий отделом популяционной генетики и генетических основ разведения животных, Федеральный исследовательский центр животноводства – ВИЖ имени академика Л.К. Эрнста, Москва, Россия; alex_sermyagin85@mail.ru.

Information about the authors

Leonid V. Legashev, Cand. Sci. (Eng.), Leading Researcher, Orenburg State University, Orenburg, Russia; silentgir@gmail.com.

Irina P. Bolodurina, Dr. Sci. (Eng.), Prof., Head of Department, Orenburg State University; Federal Research Centre of Biological Systems and Agrotechnologies of the Russian Academy of Sciences, Orenburg, Russia; ipbolodurina@yandex.ru.

Lubov S. Grishina, Lecturer, Orenburg State University, Orenburg, Russia; zabrodina97@inbox.ru.

Irina A. Lashneva, Postgraduate Student, Junior Researcher, Federal Research Center for Animal Husbandry named after Academy Member L.K. Ernst, Moscow, Russia; lashnevaira@gmail.com.

Alexander A. Sermyagin, Cand. Sci. (Agricultural sciences), Leading Researcher, Head of Department, Federal Research Center for Animal Husbandry named after Academy Member L.K. Ernst, Moscow, Russia; alex_sermyagin85@mail.ru.

Авторы заявляют об отсутствии конфликта интересов.
The authors declare no conflicts of interests.

Статья поступила в редакцию 27.06.2022
The article was submitted 27.06.2022