МЕТОДЫ

УДК 004.94+636 *doi: 10.25687/1996-6733.prodanimbiol.2019.1.95-111*

МЕТОДЫ КОМПЬЮТЕРНОГО АНАЛИЗА ДАННЫХ В ЗАДАЧАХ ПО МОНИТОРИНГУ И СОВЕРШЕНСТВОВАНИЮ УПРАВЛЕНИЯ СТАДОМ

Михальский А.И., Новосельцева Ж.А.

Институт проблем управления им. В.А. Трапезникова РАН, Москва, Российская Федерация

Рассматриваются современные методы компьютерного анализа данных, применяемые при решении широкого круга задач, возникающих в области современного продуктивного животноводства. Некоторые задачи этого круга, рассмотренные в более ранней публикации авторов, включают в себя выявление клинического мастита при автоматическом доении, исхода лечения респираторных заболеваний, индивидуальной продуктивной ценности, фертильности, обоснование решений по выбраковке коров, прогноз различных продуктивных показателей на основании геномных оценок. Оптимизации управления стадом на основании получаемых оценок является перспективным направлением повышения эффективности современного животноводства. Настоящая статья содержит описание основных методов анализа эмпирических данных. В статье кратко описываются методы построения регрессионных зависимостей: метод наименьших квадратов, модели, учитывающие случайные эффекты, байесовские, гребневые регрессии. Описаны частичная регрессия, ядерная регрессия и адаптивная многомерная непараметрическая сплайн-регрессия. Рассмотрен классический метод дискриминантного анализа, метод логистической регрессии, методы наивного Байеса и *k* ближайших соседей, метод опорных векторов, искусственные нейронные сети, деревья решений и случайный лес. В приложении приведены основные метрики, принятые для характеристики качества решения задач классификации и построения регрессионных зависимостей: точность, полнота, метрика F_1, чувствительность, специфичность, кривая ошибок, площадь под кривой ошибок. Материал статьи будет полезен специалистам широкого профиля, интересующимся применением современных методов анализа и интерпретации экспериментальных данных.

Ключевые слова: молочный скот, анализ данных, компьютеризация, диагностика, прогнозирование, управление стадом

Проблемы биологии продуктивных животных, 2019, 1: 95-111

Введение

По мере накопления эмпирических данных в области селекции и генетики животных, воспроизводстве, кормлении и ветеринарии стало возможным ставить и решать задачи оценки и прогноза продуктивного потенциала животных, эффективности воспроизводства, оптимизации управления стадом (Кузнецов, 1995, 2018; Михайленко, 2015; Черепанов и др., 2017) Решение этих задач возможно путём создания баз данных с индивидуальными характеристиками животных и привлечения широкого набора сравнительно новых вычислительных алгоритмов, показавших свою эффективность в управлении сложными техническими системами и в медико-биологических технологиях. В более ранней публикации авторов рассмотрены некоторые проблемы продуктивного животноводства с указанием методов машинного обучения, которые используются при их решении (Михальский, Новосельцева, 2018). Всю совокупность этих методов можно условно разделить на методы классификации и методы построения регрессионных зависимостей. В первом случае

требуется сформулировать качественный ответ на поставленный вопрос. Например, будет ли потомство конкретного производителя высокопродуктивным или нет. Во втором случае ответ должен быть количественным, например – какими будут показатели продуктивности у коров с конкретным генотипом. К смешанному типу относятся методы, в которых предсказывается вероятность принадлежности изучаемого объекта к конкретному классу. Так, в методе логистической регрессии строится зависимость вероятности успешного осеменения в виде функции от выбранных факторов, принимающей значения между 0 и 1 (Hempstalk et al., 2015). Если значение этой функции при заданном наборе факторов больше заданной величины, то прогнозируется успешный результат осеменения.

Цель данной работы – дать описание методов анализа эмпирических данных по разным направлениям зоотехнии, разведения, генетики и ветеринарии, показавших свою эффективность в исследованиях, проведенных в последнее десятилетие.

В таблице дана сводка этих методов и решаемых ими задач. В приложении приведены основные метрики, принятые для характеристики качества решения задач классификации и построения регрессионных зависимостей.

Основные классы методов компьютерного анализа, применяемых при решении задач в области продуктивного животноводства

Метод	Решаемая проблема
Регрессии	Прогноз величины надоя молока, суточного выхода молочного белка и процентного содержания жира (Brugemann et al., 2011; Zhang et al., 2016). Прогноз эффективности искусственного осеменения (Hempstalk et al., 2015). Расчёты по геномному прогнозу продуктивных показателей (Savagnago et al., 2013). Прогноз эффективности конверсии корма в продукцию (Manafiazar et al., 2013). Генетическое предсказание интенсивности выбытия коров из стада (Sasaki et al., 2015).
BLUP	Генетический прогноз продуктивных показателей (Кузнецов, 1995: Pintus et al., 2012; Jimenez-Montero et al., 2013; Colombani et al., 2013; Lourenco et al., 2014; Ma et al., 2015). Геномный прогноз эффективности конверсии корма в продукцию (Pryce et al., 2012). Генетический прогноз эффективности осеменения (Aguilar et al., 2011).
Байесовские методы и сети	Классификация поведения коров на выпасе (Dutta et al., 2015). Обоснование решений по выбраковке коров (Dhakal et al., 2015). Обнаружение клинического мастита при автоматическом доении (Steeneveld et al., 2010). Предсказание исхода лечения респираторных заболеваний (Amrineab et al., 2014). Прогноз эффективности искусственного осеменения (Shahinfar et al., 2014; Hempstalk et al., 2015). Геномный прогноз продуктивных показателей (Pintus et al., 2012; Jimenez-Montero et al., 2013; Colombani et al., 2013; Lourenco et al., 2014; Ferragina et al., 2015; Ma et al., 2015). Геномный прогноз эффективности конверсии корма в продукцию (Pryce et al., 2012).
Ядерные регрессии	Предсказание фертильности быков (Abdoliahi-Arpanahi et al., 2017).
Сплайны адаптивной регрессии	Прогноз эффективности искусственного осеменения (Grzesiak et al., 2010).
Дискриминантный анализ	Автоматическая идентификации дойных коров (Li et al., 2017) Классификация поведения коров на выпасе (Dutta et al., 2015). Предсказание времени отёла (Borchers et al., 2017). Обоснование решений по выбраковке коров (Adamczyk et al., 2016).

Продолжение таблицы

Метод *k* ближайших соседей	Классификация поведения коров на выпасе (Dutta et al., 2015).
Метод опорных векторов	Автоматическая идентификации дойных коров (Li et al., 2017) Предсказание динамики живой массы бычков на откорме (Alonso et al., 2015). Прогноз эффективности искусственного осеменения (Hempstalk et al., 2015),
Искусственные нейронные сети (в том числе с использованием нечеткой логики)	Автоматическая идентификации дойных коров (Li et al., 2017) Классификация поведения коров на выпасе (Nadimi et al., 2012; Dutta et al., 2015). Обнаружение теплового стресса у бычков на откорме (de Sousa et al., 2018). Предсказание параметров рубцовой ферментации (Craninx et al., 2008). Предсказание времени отёла (Borchers et al., 2017). Предсказание величины надоя (Salehi et al., 2000; Sanzogni, Kerr, 2001; Grzesiak et al., 2006). Обоснование решений по выбраковке коров (Adamczyk et al., 2016). Обнаружение ранних стадий мастита (Ankinakattea et al., 2013). Предсказание исхода лечения респираторных заболеваний (Amrineab et al., 2014). Прогноз эффективности искусственного осеменения (Grzesiak et al., 2010).
Деревья решений	Классификация поведения коров на выпасе (Gonzalez et al., 2015, Dutta et al., 2015). Обнаружение течки по отклонениям суточного удоя (Mitchell et al., 1996). Обнаружение и фильтрация выбросов данных при регистрации лактационной кривой (Pietersma et al., 2003). Обоснование решений по выбраковке коров (McQueen et al., 1995). Обнаружение клинического мастита при автоматическом доении (Kamphuis et al., 2010). Оценка экономических аспектов различных стратегий лечения мастита (Pinzon-Sanchez et al., 2011). Предсказание исхода лечения респираторных заболеваний. (Amrineab et al., 2014). Прогноз эффективности искусственного осеменения (Shahinfar et al., 2014; Hempstalk et al., 2015).
Случайный лес, вращающий лес	Совершенствование технологий кормления скота (Flores et.al., 2017). Предсказание времени отёла (Borchers et al., 2017). Прогноз эффективности искусственного осеменения (Shahinfar et al., 2014; Hempstalk et al., 2015). Расчет геномных оценок продуктивных показателей (Li et al., 2018) Геномный прогноз эффективности конверсии корма в продукцию (Yao et al., 2013).

1. Методы построения регрессионных зависимостей

При установлении связи между количественными показателями пользуются методами восстановления регрессионных зависимостей (построением эмпирической регрессии). Будучи старейшим методом анализа данных, восходящим к временам Гаусса и Гальтона, этот метод не теряет своей популярности, модифицируясь и приобретая новые черты и свойства.

1.1 Метод наименьших квадратов (МНК)

В классической формулировке метода МНК зависимость между набором факторов $X_1,...,X_k$ и показателем *Y* определяют как такую функцию от факторов $\varphi(X_1,...,X_k)$, которая минимизирует среднеквадратичное отклонение значений этой функции от наблюдаемых значений показателя *Y*. Математически это требование записывается в виде

$$J(\varphi)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\varphi(X_{i1},...,X_{ik})\right)^2 \xrightarrow[\varphi]{} \min \qquad (1)$$

где Y_i (i=1,…,n) - значение показателя для набора значений факторов $X_{i1},...,X_{ik}$, зафиксированных в i-ом эксперименте в серии из n экспериментов. Вид функции φ задаёт исследователь, исходя из решаемой задачи и представлений своей предметной области. В

настоящее время принято рассматривать не одну, а ряд функций, выбирая из них на основании статистического критерия качества ту, которая в наибольшей степени согласована с данными.

Базовая модель

Самая простая функция, связывающая показатель Y с набором k факторов $X_1,...,X_k$, - линейная, которая записывается в виде

$$Y_i = \sum_{j=1}^{k} X_{ij} a_j + \varepsilon_i ,$$

т.е. в формуле (1) функция регрессии имеет вид

$$\varphi(X_1,...,X_k) = \sum_{j=1}^{k} X_{ij} a_j$$

здесь a_j - вес, с которым фактор X_j влияет на показатель Y, ε_i - величина рассогласования значения показателя Y_i с линейной моделью. В матричной форме эта модель записывается так:

$$Y = Xa + \varepsilon , \qquad (2)$$

где Y и ε – вектор-столбцы, состоящие из n элементов, обозначенных Y_i и ε_i соответственно, a - вектор-столбец, состоящий из k элементов a_j, матрица X состоит из n строк и k столбцов, её элементами являются значения факторов X_{ij}.

Значение вектора a, при котором для модели (2) достигается минимум среднеквадратичной ошибки, вычисляется по формуле

$$\hat{a} = \left(X^T X\right)^{-1} X^T Y \qquad (3)$$

Значение целевой переменной для нового объекта, описываемого вектором-строкой u значений факторов $u_1,...,u_k$, вычисляется по формуле $\hat{y} = u\hat{a} = u\left(X^T X\right)^{-1} X^T Y$.

Модель, учитывающая случайные эффекты (REM)

В реальности, предположение о том, что показатель Y описывается линейной комбинацией факторов $X_1,...,X_k$, может оказаться слишком упрощённым. В животноводстве такая ситуация возникает, например, если данные не однородны, а собираются от представителей групп, различающихся генетически. В различных группах степень влияния факторов может быть различной. Это можно учесть, считая, что вектор ε в выражении (2) состоит из двух случайных членов: первый зависит от конкретной породы, а второй учитывает индивидуальные особенности. Именно такая интерпретация позволяет говорить, что модель учитывает случайные эффекты. Если рассмотреть в базовой модели (2) матрицу факторов X и вектор весов a как составные, положив $X=(T,Z)$, $a=(\gamma,\beta)$, то модель (2) преобразуется к виду

$$\boldsymbol{Y} = \boldsymbol{T\gamma} + \boldsymbol{Z\beta} + \boldsymbol{\varepsilon} , \qquad (4)$$

где T – матрица значений факторов, характеризующих группу, Z – матрица значений факторов, характеризующих индивидуальные различия в группе. Для оценки весов γ и β можно воспользоваться формулой (3). Использование таких регрессионных моделей оказалось полезным, в частности, при генетическом предсказании интенсивности выбытия коров из стада.

1.2. Байесовские регрессии

Более сложные модели строятся на основании байесовского подхода, при котором оцениваемые параметры считаются случайными величинами, имеющими распределение, не зависящее от наблюдений – априорное распределение. Такой подход позволяет получать более точные оценки, чем обычный МНК, и даже вычислять оценки в ситуации, когда МНК неприменим.

Наилучший линейный несмещённый прогноз (BLUP)

В моделях, учитывающих случайные эффекты (4), в том числе в генетических исследованиях, основной интерес представляет оценка влияния эффектов, связанных с генетическими особенностями различных групп, в то время, как оценка влияния детерминированных эффектов носит второстепенный характер. В этом случае более точный результат можно получить, строя не оценку вектора весов β в модели (4), а прогнозируя его возможное значение при имеющихся данных.

Для построения наилучшего несмещённого прогноза делается предположение, что в модели (4) случайный вектор весов β и случайный вектор ε распределены по нормальному закону, независимы между собой и имеют нулевое математическое ожидание. Это априорное распределение. В математической статистике доказано, что в этом случае наиболее точным прогнозом вектора β является вектор условного математического ожидания $E(\beta|Y)$, который оказывается линейной комбинацией данных, имеет нулевое смещение относительно истинных значений параметров и минимальную дисперсию. Такой прогноз носит название BLUP и вычисляется по формуле, учитывающей корреляционные зависимости элементов векторов β и ε (Кузнецов, 1995).

В перечень задач молочного животноводства, решаемых с применением BLUP, входят расчет геномных оценок продуктивных показателей, геномный прогноз показателя конверсии корма в продукцию и генетическая оценка эффективности осеменения.

Модели BayesA, BayesB, BayesCπ, BayesDπ, BayesMulti

Прогноз влияния генетических факторов на фенотипические характеристики, построенный с помощью модели BLUP, недостаточно точен из-за того, что дисперсии элементов вектора β считаются фиксированными. Чтобы смягчить это ограничение и лучше адаптировать модель к данным, в работе (Meuwissen et al., 2001) предложены две модификации модели BLUP: BayesA и BayesB. В этих моделях дисперсии элементов вектора β также считаются случайными величинами и их значения прогнозируются на основании эмпирических данных. Различие между BayesA и BayesB заключается в разнице априорных распределений дисперсий. В модели BayesA априорное распределение дисперсий задаётся так, что величины, обратные дисперсиям, имеют многомерное распределение хи-квадрат. Параметры распределения определяются по эмпирическим данным. В модели BayesB допускается, что некоторые из дисперсий с некоторой вероятностью π равны нулю, а с вероятностью 1-π распределены, как в модели BayesA. Дисперсии остальных элементов имеют априорное распределение, как в модели BayesA.

В работе (Habier et al., 2011) расширен список байесовских моделей, а именно – введены модели BayesCπ и BayesDπ. Эти модели являются модификациями моделей BayesA и BayesB. В модели BayesCπ предполагается, что дисперсии всех элементов вектора β одинаковы, а вероятность π того, что дисперсия равна нулю, имеет равномерное априорное распределение на отрезке [0,1]. В модели BayesDπ дисперсии элементов вектора β различны, величины, обратные дисперсиям, имеют многомерное распределение хи-квадрат, а вероятность π того, что дисперсия равна нулю, имеет равномерное априорное распределение на отрезке [0,1]. В работе (Pryce et al., 2012) предложена модификация модели BayesA под названием BayesMulti, в которой предварительно производится отбор информативных элементов вектора β.

Наиболее широко эти методы используются при вычислении геномных оценок продуктивных показателей и геномном прогнозе показателя конверсии корма в продукцию.

1.3. Модификации метода наименьших квадратов

Гребневая регрессия (*Ridge Regression*). Вычисление вектора весов влияния факторов по формуле (3) возможно только в случае, если измеренные в эксперименте значения факторов независимы и число наблюдений n больше числа факторов k. В настоящее время возникают задачи, например связанные с генетическим анализом, в которых изучается

влияние на целевую переменную факторов, число которых больше числа наблюдений. В обозначениях базовой модели это означает, что матрица $(X^T X)^{-1}$ не существует. Для оценки влияния факторов в такой ситуации используется байесовский подход, то есть привлечение априорной информации о распределении элементов вектора a.

Если допустить, что элементы вектора ε_i, $i=1,...,n$, и элементы вектора a_j, $j=1,...,k$, являются независимыми случайными величинами, имеющими n-мерное и k-мерное нормальные распределения соответственно с нулевыми математическими ожиданиями, то задача построения регрессионной зависимости сведётся к задаче минимизации *регуляризованного функционала* (Мерков, 2011; Hoerl, Kennard, 1970):

$$J(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k} X_{ij} a_j\right)^2 + \lambda \sum_{j=1}^{k} a_j^2 \xrightarrow[a]{} \min ,$$

решение которой вычисляется по формуле

$$\hat{a}_\lambda = \left(X^T X + \lambda I\right)^{-1} X^T Y ,$$

где матрица I состоит из k столбцов и k строк и все её элементы, кроме диагональных, равны нулю, а диагональные элементы равны 1. Величина λ является параметром настройки модели и определяется в процессе решения задачи.

Регрессия LASSO (Least Absolute Shrinkage and Selection Operator). Если принять, что элементы вектора a_j, $j=1,...,k$ являются случайными независимыми величинами, имеющими распределение Лапласа с нулевым математическим ожиданием, то задача построения регрессионной зависимости сведётся к задаче минимизации другого регуляризованного функционала (Hastie et al., 2016):

$$J(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k} X_{ij} a_j\right)^2 + \lambda \sum_{j=1}^{k} \left|a_j\right| \xrightarrow[a]{} \min .$$

Решение этой задачи строится итерационными методами, величина λ является параметром настройки модели и определяется в процессе решения задачи.

Важным свойством метода является то, что в зависимости от величины параметра λ некоторые элементы полученного решения оказываются нулевыми. Это означает, что некоторые факторы не влияют существенно на значение целевой переменной и могут быть исключены из рассмотрения. Часто именно регрессия LASSO используется для выбора наиболее значимых в рассматриваемой задаче факторов.

Частичная регрессия PLS и *sPLS*. В работе (Colombani et al., 2013) рассмотрен метод уменьшения числа факторов в регрессионной зависимости (PLS, Partial Least Squares – частичный МНК), основанный на построении методом наименьших квадратов зависимости не от исходных факторов, а от фиктивных, латентных переменных, которые строятся методом главных компонент из условия максимизации ковариации между целевой переменной и заданным числом латентных переменных. Метод PLS предусматривает построение набора частичных регрессий для различного числа латентных переменных, из которого выбирается оптимальная регрессия. Показано, что такой подход в ряде задач по точности не уступает как методу гребневой регрессии, так и методу LASSO.

Было предложено проводить дополнительный отбор информативных переменных методом, близким к методу LASSO (Chun, Keles., 2009). Поскольку при этом отсеиваются некоторые переменные, то метод получил название sPLS (sparse Partial Least Squares – прореженный PLS).

Описанные модификации методов гребневой регрессии и LASSO применяются при построении различных геномных оценок.

1.4. Ядерная регрессия

Интересное обобщение линейной модели (4) со случайными эффектами описано в работе (Gianola, van Kaam, 2008). Если ввести обозначение $u=Z\beta$, то модель (4) запишется в виде:

$$Y = Xa + u + \varepsilon .$$

Из гипотезы о нормальности распределения элементов вектора β получим, что вектор u тоже распределён нормально с нулевым математическим ожиданием и ковариационной матрицей $K = ZAZ^T$, где A – ковариационная матрица распределения вектора β. Представим вектор u в виде $u=Kf$ и запишем:

$$Y = Xa + Kf + \varepsilon , \quad (5)$$

где f - вектор весов в другом пространстве, задаваемом матрицей K, которая называется *воспроизводящее ядро* (Pérez-Elizalde et al., 2015). Смысл такого перехода заключается в том, что задавая различные K, имеется возможность использовать линейную модель (5) для оценки нелинейных и взаимозависимых эффектов, когда основные положения линейной модели (4) нарушаются. Векторы a и f можно найти, используя, например, метод гребневой регрессии. В качестве ядра может выступать любая неотрицательно определённая матрица. В литературе по геномному анализу популярно гауссовское ядро, элементы которого определяются соотношением

$$K_{ij} = \exp\left(-h\sum_{r=1}^{k}\left(z_{ir} - z_{jr}\right)^2 \right),$$

где $i=1,...,n$, $j=1,...,n$, z_{ir} и z_{jr} – элементы матрицы Z в модели (4), h - параметр, задающий "ширину" гауссовского ядра, который определяется на основании имеющихся данных.

Метод ядерной регрессии использовался, например, при предсказании фертильности быков (см. табл.).

1.5. Адаптивная многомерная непараметрическая сплайн регрессия (MARS)

Предложен метод построения многомерной регрессионной зависимости, не требующий задания модели (Friedman, 1991). В этом методе многомерная функция регрессии $\varphi(X_1,..., X_k)$, присутствующая в выражении (1), представляется в виде:

$$\varphi(X_1,..., X_k) = \sum_{j=0}^{N} c_j B_j(X_1,..., X_k) \quad (6)$$

Функции B_j – многомерные базисные функции, которые формируются в процессе решения задачи из одномерных базисных сплайн-функций. Результирующее число базисных функций N также определяется в процессе решения. Коэффициенты c_j находятся по методу наименьших квадратов.

В качестве одномерных базисных сплайн-функций используются ломанные линии вида $S_t^+(x) = \max(0, x-t)$, $S_t^-(x) = \max(0, t-x)$. Величина t является узлом одномерного сплайна.

На первом этапе построения многомерной регрессионной зависимости (forward stage), начиная с постоянной по величине базисной функции, в формуле (6) последовательно добавляются базисные функции, чтобы минимизировать среднеквадратичную ошибку приближения данных. Процесс продолжается, пока либо не будет достигнута необходимая величина среднеквадратичной ошибки, либо число базисных функций не достигнет заданного предела. На втором этапе (backward stage) из построенной модели исключаются базисные функции, удаление которых в наименьшей степени увеличивает среднеквадратичную ошибку приближения данных. Таким образом, построенная модель адаптируется к сложным

многомерным эмпирическим данным.

С применением этого метода проводилась, в частности, оценка эффективности искусственного осеменения (см. табл.).

2. Методы классификации

Задача классификации отличается от задачи регрессии тем, что в качестве выходной величины ищется не количественное значение зависимости для исследуемого объекта, а качественное суждение, например, будет ли потомство данного производителя высоко продуктивным, или будет ли у носителя данного генотипа продолжительность продуктивной жизни выше среднестатистической. Формально это означает, что целевой переменной Y является величина, принимающая два (больше двух в случае многоклассовой классификации) значений. Часто они обозначаются как 0 и 1 или -1 и 1. Эти значения называются *метки классов.*

Ряд методов классификации, по сути, основан на построении регрессионных зависимостей, и при классификации производится сравнение значения регрессионной оценки с неким порогом. При превышении этого порога заключают о принадлежности объекта к классу с меткой 0, в противном случае – к классу с меткой 1. Так устроен метод логистической регрессии. Однако, большинство методов ориентировано на построение решающего правила, которое с минимальной ошибкой будет правильно классифицировать новые данные, даже если правило не соответствует модели, заложенной в данных. Так, например, в методе SVM строится линейное решающее правило вне зависимости от физической модели данных.

Задачи классификации не менее востребованы в продуктивном животноводстве, чем задачи регрессии. При этом разнообразие используемых методов покрывает весь диапазон современных методов классификации. Ниже кратко представлены основные из них. Данные, на основании которых строится правило классификации называют *обучающей выборкой* $\left\{X_1^1,...,X_k^1,Y^1,...,X_1^n,...,X_k^n,Y^n\right\}$. Совокупность $\left(X_1^i,...,X_k^i,Y^i\right)$ i=1,...,n обозначает значения k факторов и соответствующей им метки класса в i-ом эксперименте.

2.1. Дискриминантный анализ

Одним из старейших методов классификации является дискриминантный анализ, разработанный Р.Фишером (Fisher, 1936). Предполагается, что факторы, характеризующие объекты из класса 0 и объекты из класса 1, имеют вероятностную природу и описываются распределениями Гаусса, различающимися средними значениями M_y и ковариационными матрицами K_y при y=0,1

$$p\left(X \mid Y=y\right)=\frac{1}{\left(2\pi\right)^{k/2}\sqrt{\det\left(K_y\right)}}\exp\left(-\frac{1}{2}\left(X-M_y\right)^T K_y^{-1}\left(X-M_y\right)\right)$$

Правило классификации запишется в следующем виде: если выполняется неравенство $\varphi\left(X_1,...,X_k\right)\geq d$, то объект относится к классу 0, иначе – к классу 1. Здесь дискриминантная функция $\varphi\left(X_1,...,X_k\right)=\ln p\left(X \mid Y=0\right)-\ln p\left(X \mid Y=1\right)$ оказывается либо линейной в случае равенства ковариационных матриц K_0=K_1, либо квадратичной.

Векторы средних значений M_y и ковариационные матрицы K_y либо задаются априори, либо оцениваются по данным. В этом заключается обучение метода. Для определения величины порога d пользуются априорными сведениями о соотношении вероятности встретить объект из конкретного класса, либо методом перекрестной проверки (Мерков, 2011). Метод дискриминантного анализа популярен в качестве метода классификации; он использовался, например, при решении задач индивидуальной идентификации животных, классификации поведения на выпасе и обосновании решений по выбраковке (см. табл.).

2.2. Логистическая регрессия

В методе классификации, использующем логистическую регрессию, в качестве целевой переменной принимают вероятность принадлежности объекта к конкретному классу, а в качестве эмпирических значений этого показателя – частоту встречаемости показателя в экспериментальных данных. Поскольку восстанавливаемый показатель лежит между нулем и единицей, то используется специальный вид зависимости – логистическая кривая вида

$$\varphi(X_1,...,X_k)=\frac{1}{1+\exp\left(\sum_{j=1}^{k}a_jX_j\right)}.$$

Параметры a_j характеризуют веса различных факторов при решении задачи классификации. Величина этих параметров определяется методом максимального правдоподобия по обучающей выборке (Мерков, 2011). При классификации на два класса объект относится к классу 0, если для него значение логистической регрессии превосходит заданный порог d, в противном случае – к классу 1. Величина порога d задаётся либо из априорных соображений, ориентируясь на соотношение представителей двух классов в обучающей выборке, либо определяется из условий достижения компромисса между вероятностями ошибочно принять объект из класса 0 за объект из класса 1, и наоборот. Кроме того, широко распространён способ определения величины порога d методом перекрёстной проверки (Мерков, 2011). Метод логистической регрессии использовался, например, при предсказании индекса осеменения коров (см. табл.).

2.3. Метод наивного Байеса

Тенденция в развитии методов классификации заключается либо в упрощении предположений относительно характера распределения факторов внутри классов, либо в упрощении вида дискриминантной функции $\varphi(X_1,...,X_k)$. К первому случаю относится метод, в котором факторы внутри каждого класса считаются независимыми. В результате распределение факторов внутри классов принимает вид:

$$p(X\mid Y=y)=\prod_{j=1}^{k}p(X_j\mid Y=y),$$

где $p(X_j\mid Y=y)$ - плотность распределения фактора X_j в классе y. Параметры этих плотностей определяются методом максимального правдоподобия (Мерков, 2011), а решающее правило имеет следующий вид: если справедливо неравенство $\ln p(X\mid Y=0)-\ln p(X\mid Y=1)\geq d$, то объект относится к классу 0, в противном случае – к классу 1. Величина порога d, как и в других методах, определяется либо исходя из априорных сведений о соотношении вероятности встретить объект из конкретного класса, либо методом перекрестной проверки. Из-за простоты реализации метод наивного Байеса широко используется на практике. Например, он использовался при классификации режимов поведения пасущихся коров, диагностике клинического мастита, предсказания индекса осеменения коров.

2.4. Метод k ближайших соседей

Популярным методом классификации является метод k ближайших соседей (Мерков, 2011). В простейшей реализации этого метода выбираются k точек, ближайших к классифицируемой точке X^*. Эта точка относится к тому классу, представителей которого больше среди её k соседей. Модификации этого метода заключаются в том, как определять близость точек в пространстве признаков и как задавать число соседей k. Этот метод применялся при классификации поведения я коров на выпасе.

2.5. Метод опорных векторов SVM (Support Vector Machine)

Популярный в настоящее время метод опорных векторов, ориентирован на вычисление простой линейной дискриминантной функции $\varphi(X_1,...,X_k)$ (Мерков, 2011). Суть метода заключается в построении линейной дискриминантной функции, удовлетворяющей требованиям минимизации ошибки на обучающей выборке и минимизации нормы направляющего вектора соответствующей гиперплоскости. Линейная дискриминантная функция записывается в виде

$$\varphi(X_1,...,X_k)=\sum_{j=1}^{k}a_jX_j-a_{k+1},$$

а выражение $\varphi(X_1,...,X_k)=0$ описывает гиперплоскость в *k*-мерном пространстве.

Направляющим называется вектор $w=(a_1,...,a_k)$, который задаёт ориентацию гиперплоскости в пространстве, параметр a_{k+1} определяет величину смещения гиперплоскости относительно начала координат. Объект, характеризующийся набором факторов $X_1,...X_k$,, для которого справедливо неравенство $\varphi(X_1,...,X_k)>0$, относится к классу с меткой 0, в противном случае – к классу с меткой 1. Направляющий вектор гиперплоскости *w* и величина смещения вычисляются методами квадратичного программирования. При этом оказывается, что достаточно использовать не все элементы обучающей выборки, а лишь часть их, наименее удаленных от оптимальной гиперплоскости. Эти элементы получили название "опорные векторы", а оптимальная гиперплоскость записывается с помощью опорных векторов X^{*i} в виде

$$\sum_{i=1}^{n^*}b_i\left(X,X^{*i}\right)-d=0,$$

где n^* - число опорных векторов, параметры $b_1,...,b_{n^*}$ и *d* определяются в процессе решения задачи квадратичного программирования. Запись $\left(X,X^{*i}\right)$ обозначает скалярное произведение вектора *X* с опорным вектором X^{*i}. Использование опорных векторов существенно улучшает решение многих практических задач как по времени обработки данных, так и по статистической надежности получаемого результата классификации.

Существенное повышение эффективности метода SVM связано с использованием ядерных функций. При этом скалярное произведение $\left(X,X^{*i}\right)$ заменяется на неотрицательную симметричную функцию $K\left(X,X^{*i}\right)$, которая называется ядерной или просто ядром. Такая замена известна под названием "ядерный трюк". Объект *X* относится к классу с меткой 0, если $\sum_{i=1}^{n^*}\hat{b}_iK\left(X,X^{*i}\right)-\hat{d}\geq 0$, и к классу с меткой 1 в противном случае. Параметры $\hat{b}_i$ и $\hat{d}$ определяются в процессе решения задачи квадратичного программирования с используемым ядром. Популярными на практике являются радиальные базисные ядра $K\left(X,X^{*i}\right)=\exp\left(-\alpha\sum_{j=1}^{k}\left(X_j-X_j^{*i}\right)\right)^2$. Метод опорных векторов использовался при прогнозе эффективности осеменения и предсказании динамики живой массы по данным за предшествующие периоды (см. табл.).

2.6. Нейронные сети

Вид решающего правила, похожего на линейное, возникает при обучении искусственной нейронной сети, состоящей из нескольких слоёв нейронов – элементов, находящихся в состоянии покоя или возбуждения, которым соответствуют сигналы 0 или 1 на выходе элемента (Мерков, 2011). Состояние нейрона зависит от суммы сигналов на выходе нейронов предыдущего слоя, которые суммируются с определёнными весами. Эти веса настраиваются в процессе обучения так, чтобы минимизировать ошибку распознавания

объектов из обучающей выборки при решении задачи классификации, либо суммарную квадратичную ошибку при построении регрессионной зависимости. Таким образом, в каждом слое происходит линейное преобразование сигналов, поступающих от нейронов предыдущего слоя, а всё решающее правило описывается кусочно-линейной функцией от многих переменных.

Нейронные сети использовались при решении задач автоматической идентификации дойных коров, классификации поведения их на выпасе, выявления теплового стресса, оценки параметров рубцовой ферментации, прогнозирования времени отёла и величины надоя, обоснования выбраковки коров, обнаружения ранних стадий мастита, предсказания исхода лечения респираторных заболеваний, оценки эффективности искусственного осеменения.

2.7. Деревья решений

Кроме алгоритмов классификации, учитывающих вероятностную природу данных, широко распространены методы построения логических классификаторов (Мерков, 2011). Простейшим примером является двоичное дерево принятия решения. В этом методе факторы-признаки ранжируются по значимости для принятия решения, и значения факторов последовательно разбиваются на два диапазона в зависимости от значений факторов предыдущего уровня. Границы разбиений определяются в процессе обучения, а новый объект классифицируется в зависимости от того, в какую конечную вершину – лист получившегося дерева – он попадёт.

Этот метод применялся для классификации поведения пасущихся коров, обнаружения течки по отклонениям суточного удоя, обнаружения и фильтрации выбросов данных при регистрации лактационных кривых, обоснования решений по выбраковке коров, обнаружения клинического мастита при автоматическом доении, оценки экономических аспектов различных стратегий лечения мастита, прогноза исхода лечения респираторных заболеваний, оценки эффективности искусственного осеменения.

2.8. Случайный лес (random forest)

Обобщением метода построения решающих деревьев является метод под названием "случайный лес", реализующий идею многократного решения задачи классификации для выбора лучшего результата (Мерков, 2011). При реализации этого метода исходная обучающая выборка случайным образом многократно разбивается на подвыборки, содержащие меньшее число элементов. Некоторые элементы исходной выборки могут попадать в несколько сгенерированных подвыборок, а некоторые могут не попасть ни в одну. На каждой из подвыборок строится дерево решений, полученные деревья используются для классификации. Объект относится к тому классу, к которому его отнесло наибольшее число деревьев.

В вышеприведенной таблице приведены ссылки на работы, выполненные с использованием этого метода для предсказания динамики живой массы коров, времени отёла, оценки эффективности искусственного осеменения, расчёта геномного прогноза продуктивных показателей и эффективности конверсии корма в продукцию.

2.9. Вращающий лес (rotation forest)

Развитием метода "случайный лес" является метод "вращающий лес" (Rodriguez J.J. et al., 2006), в котором при построении каждого из деревьев исходное пространство признаков подвергается преобразованию путём случайного разбиения на K групп (возможно пересекающихся). В каждой группе вычисляются главные компоненты. Совокупность этих компонент образует базис, в который преобразуются данные обучающей выборки. По этим преобразованным данным строится дерево решения. Матрица преобразования носит название "матрица вращения". Новый объект относится к тому классу, к которому его отнесло наибольшее число деревьев. Метод использовался в задаче оценки эффективности искусственного осеменения.

Заключение

В настоящем обзоре и в предыдущей статье (Михальский, Новосельцева, 2018) предпринята попытка описать круг задач в области продуктивного животноводства и методы машинного обучения, успешно применяемые для решения этих задач. Число публикаций, посвящённых этой теме, неуклонно растёт. Только в области продуктивного скотоводства ежегодно публикуется около 6 тысяч научных статей на английском языке, посвящённых использованию методов анализа данных и машинного обучения (https://scholar.google.com). Соответственно увеличивается и число новых методов, применяемых для повышения эффективности результатов. К ним относятся *нейронные сети* глубокого обучения, в которых выделение значимых признаков передаётся нейронной сети, *бэггинг*, в котором результат определяется по результату голосования работы нескольких алгоритмов, *бустинг*, заключающийся в построении последовательности алгоритмов, в которой каждый последующий компенсирует недостатки всех предыдущих. В обзоре не затронут большой класс алгоритмов кластеризации, используемых в случае, когда в данных отсутствует признак класса, задаваемый до начала анализа, и классы формируются в процессе работы алгоритма.

Современным методам машинного обучения посвящено большое число книг и научных публикаций, в основном, на английском языке. На русском языке для начального ознакомление с областью машинного обучения можно рекомендовать цитировавшуюся в настоящем обзоре книгу (Мерков, 2011), а также издание (Мерков, 2014), в котором описываются методы построения по данным вероятностных моделей и, в частности, моделей дожития с учётом факторов риска и цензурирования данных. Глубокое описание основ машинного обучения дано в работах (Вапник, 1979) и (Vapnik, 2000). Подготовленному читателю можно рекомендовать издания (Вьюгин, 2013) и (Shalev-Shwartz., Ben-David, 2014).

В настоящее время использование методов машинного обучения в прикладных исследованиях является жизненной необходимостью. Разработано много готовых к использованию компьютерных систем, пакетов и программ для реализации этих методов. Однако, чтобы выбрать подход, адекватный решаемой задаче, учитывающий специфику экспериментальных данных, важно знать основные идеи, заложенные в конкретном методе, понимать его ограничения. Машинное обучение доказало свою эффективность в решении сложных задач генной инженерии, искусственного интеллекта, автоматического перевода, анализа текста, написанного на естественном языке. Молодым исследователям необходимо иметь в виду широкие перспективы применения новых методов анализа данных и стимулировать их изучение.

Приложение. Метрики качества построения регрессии и классификации

Коэффициент детерминации R^2 равен доле дисперсии целевой переменной, объяснённой регрессионной зависимостью

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}\left(y_i - n^{-1}\sum_{j=1}^{n} y_j\right)^2},$$

$f(x_i)$ – значение функции регрессии при значении фактора x_i, которому в обучающей выборке соответствует значение целевой переменной y_i, n – число элементов в обучающей выборке.

Корень из среднеквадратичнй ошибки (RMSE) равен квадратному корню из усреднённой по выборке величине квадрата разности значения функции регрессии от соответствующего значения целевой переменной

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

Относительная ошибка аппроксимации (RAE) характеризует величину средней ошибки апроксимации относительно усреднённых значений квадрата целевой переменной

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n} y_i^2}}$$

Метрики качества классификации

Accuracy – доля верно классифицированных объектов экзаменационной выборки. Если в экзаменационной выборке число объектов в разных классах существенно различаются, то эта метрика приводит к ошибочным выводам.

Точность (Precision) – доля верно классифицированных объектов класса 1 в экзаменационной выборке относительно всех объектов экзаменационной выборки, отнесённых к классу 1.

Полнота (Recall) – доля верно классифицированных объектов класса 1 в экзаменационной выборке относительно всех объектов экзаменационной выборки из класса 1.

Метрика F_1 (иногда просто показатель *F*) комбинирует точность и полноту по формуле

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Смысл рассмотрения такой метрики в том, что она уменьшается при уменьшении либо точности, либо полноты и приближается к максимальному значению 1 при идеальном правиле классификации, при котором и точность и полнота близки к значению 1.

Метрика F_β позволяет смещать акцент в пользу точности при $\beta<1$, либо в пользу полноты при $\beta>1$

$$F_\beta = (1+\beta^2)\frac{\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}}.$$

Чувствительность (Sensitivity) – синоним метрики *полнота*.

Специфичность (Specificity) – доля верно классифицированных объектов класса 0 в экзаменационной выборке относительно всех объектов экзаменационной выборки из класса 0.

Кривая ошибок (Receiver operating characteristics, ROC) показывает изменение доли верно классифицированных объектов из класса 1 и доли неверно классифицированных объектов из класса 0 в зависимости от параметров алгоритма, например, величины порога в алгоритме логистической регрессии. Графически кривая ошибок представляет собой выпуклую кривую. Чем ближе она к прямоугольной, тем лучше алгоритм.

Площадь под кривой ошибок (Area under an ROC curve, AUC). Числовой характеристикой кривой ошибок служит площадь под ней. Она вычисляется численным интегрированием и характеризует вероятность того, что случайно взятый из класса 1 объект с большей веороятностью будет отнесён к классу 1, чем случайно взятый объект из класса 0**.**

REFERENCES

1. Abdollahi-Arpanahi R., Morota G., Peñagaricano F. Predicting bull fertility using genomic data and biological information. *J. Dairy Sci*. 2017, 100(1): 9656-9666.
2. Adamczyk K., Zaborski D., Grzesiak W., Makulska J., Jagusiak W. Recognition of culling reasons in Polish dairy cows using data mining methods. *Computers and electronics in agriculture*. 2016, 127: 26-37.
3. Aguilar I., Misztal I., Tsuruta S., Wiggans G.R., Lawlor T.J. Multiple trait genomic evaluation of conception rate in Holsteins. *J. Dairy Sci*. 2011, 94(5): 2621-2624.
4. Alonso J., Villa A., Bahamonde A. Improved estimation of bovine weight trajectories using Support Vector Machine Classification. *Computers and electronics in agriculture*. 2015, 110: 36-41.
5. Amrineab D.E., Whiteb B.J., Larsonb R.L. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. *Computers and electronics in agriculture*. 2014, 105: 9-19.

6. Ankinakattea S., Norberga E., Løvendahla P., Edwardsa D., Højsgaardb S. Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison. *Computers and electronics in agriculture*. 2013, 99: 1-6.
7. Borchers M.R., Chang Y.M., Proudfoot K.L., Wadsworth B.A., Stone A.E., Bewley J.M. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *J. Dairy Sci*. 2017, 100(7): 5664-5674.
8. Brügemann K., Gernand E., von Borstel U.U., König S. Genetic analyses of protein yield in dairy cows applying random regression models with time-dependent and temperature x humidity-dependent covariate. *J. Dairy Sci.* 2011, 94(8): 4129-4139.
9. Cao K.L., Rossouw D., Robert-granié C. A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. 2008, 7(1): 35.
10. Cherepanov G.G., Kharitonov E.L., Makar Z.N., Mikhal'skii A.I., Novosel'tseva Zh.A. [An analysis of possible approaches to overcome the antagonism between the level of productivity and the viability of the breeding stock by using intensive technologies]. *Problemy biologii productivnykh zhivotnykh - Problems of Productive Animal Biology*. 2017, 1: 5-27. (In Russian)
11. Chun H., Keleş S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics.* 2009, 182: 79-90.
12. Colombani C., Legarra A., Fritz S., Guillaume F., Croiseau P., Ducrocq V., Robert-Granié C. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. *J. Dairy Sci*. 2013, 96(1): 575-591.
13. Craninx M., Fievez V., Vlaeminck B., De Baets B. Artificial neural network models of the rumen fermentation pattern in dairy cattle. *Computers and electronics in agriculture*. 2008, 60(2): 226-238.
14. De Sousa R.V., da Silva Rodrigues A.V., de Abreu M.G., Tabile R.A.,Martello L.S. Predictive model based on artificial neural network for assessing beef cattle thermal stress using weather and physiological variables. *Computers and electronics in agriculture*. 2018, 144: 37-43.
15. Dhakal K., Tiezzi F., Clay J.S., Maltecca C. Inferring causal relationships between reproductive and metabolic health disorders and production traits in first-lactation US Holsteins using recursive models. *J. Dairy Sci*. 2015, 98(4): 2713-2726.
16. Dutta R., Smith D., Rawnsley R., Bishop-Hurley G., Hills J., Timms G., Heanry D. Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and electronics in agriculture*. 2015, 111: 18-28.
17. Ferragina A., de los Campos G., VazquezA.I., Cecchinato A., Bittante G. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. *J. Dairy Sci*. 2015, 98(11): 8133-8151.
18. Fisher R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, 7: 179-188.
19. Flores H., Meneses C., Villalobos J.R., Sanchez O. Improvement of feedlot operations through statistical learning and business analytics tools. *Computers and electronics in agriculture*. 2017, 143: 273-285.
20. Friedman J.H. Multivariate adaptive regression splines. *The Annals of Statistics*. 1991, 19 (1): 1-141.
21. Gianola D., van Kaam J.B. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics.* 2008, 178(4): 2289-2303.
22. Gianola, D. Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics.* 2013, 194: 573-596.
23. González L.A., Bishop-Hurley G.J., Handcock R.N., Crossman C. Behavioral classification of data from collars containing motion sensors in grazing cattle. *Computers and electronics in agriculture*. 2015, 110: 91-102.
24. Grzesiak W., Błaszczyk P., Lacroix R. Methods of predicting milk yield in dairy cows –Predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). *Computers and electronics in agriculture*. 2006, 54(2): 69-83.
25. Grzesiak W., Zaborski D., Sablik P., Żukiewicz A., Dybus A., Szatkowska I. Detection of cows with insemination problems using selected classification models. *Computers and electronics in agriculture*. 2010, 74(2): 265-273.
26. Habier D., Fernando R.L., Kizilkaya K and Garrick D.J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics.* 2011, 12: 186.
27. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Series in Statistics. 2016, 764 pp.

28. Hempstalk K., McParland S., Berry D.P. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *J. Dairy Sci.* 2015, 98(8): 5262-5273.
29. Jiménez-Montero J.A., González-Recio O., Alenda R. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. *J. Dairy Sci.* 2013, 96(1): 625-634.
30. Kamphuis C., Mollenhorst H., Feelders A., Pietersma D., Hogeveen H. Decision-tree induction to detect clinical mastitis with automatic milking. *Computers and electronics in agriculture*. 2010, 70(1): 60-68.
31. Kuznetsov V.M. [BLUP genetic assessment of dairy cattle]. *Zootekhniya - Zootechnics*. 1995, 11: 8-15. (In Russian).
32. Kuznetsov V.M. [*In silico* study of expanded reproduction in a closed dairy cattle breeding]. *Problemy biologii productivnykh zhivotnykh - Problems of Productive Animal Biology*, 2018, 3: 54-86
33. Li B., Zhang N., Wang Y.-G., George A.W., Reverter A., Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in Genetics*. 2018, 9: 237. doi: 10.3389/fgene.2018.00237.
34. Li W., Ji Z., Wang L., Sun C., Yang X. Automatic individual identification of Holstein dairy cows using tailhead images. *Computers and electronics in agriculture*. 2017, 142(B): 622-631.
35. Lourenco D.A.L., Misztal I., Tsuruta S., Aguilar I., Ezra E., Ron M., Shirak A., Weller J.I. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *J. Dairy Sci.* 2014, 97(3): 1742-1752.
36. Ma P., Lund M.S., Nielsen U.S., Aamand G.P., Su G. Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. *J. Dairy Sci.* 2015, 98(12): 9026-9034.
37. Manafiazar G.G., McFadden T.T., Goonewardene L.L., Okine E.E., Basarab J.J., Li P.P., Wang Z. Z.. Prediction of residual feed intake for first-lactation dairy cows using orthogonal polynomial random regression. *J. Dairy Sci.* 2013, 96(12): 7991-8001.
38. McQueen R.J., Garner S.R., Nevill-Manning C.G., Witten I.H. Applying machine learning to agricultural data. *Computers and electronics in agriculture*. 1995, 12(4): 275-293.
39. Merkov A.B. *Raspoznavanie obrazov. Vvedenie v metody mashinnogo obucheniya* (Pattern recognition. Introduction to statistical learning methods). Moscow: Editorial URSS, 2011, 250 p. (In Russian)
40. Merkov A.B. *Raspoznavanie obrazov. Obuchenie i postroenie stokhasticheskikh modelei* [Pattern recognition. Training and building stochastic models]. Moscow: LENAND, 2014, 240 p. (In Russian).
41. Meuwissen T.H.E., Hayes B.J., Goddard M.E.: Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 2001, 157(4): 1819-1829.
42. Mikhailenko I.M. [Life cycle management of lactating cows on the basis of probabilistic-statistical and dynamic models]. *Sel'skokhozyaistvennaya biologiya - Agricultural Biology*. 2015, 50(4): 467-475. (In Russian).
43. Mikhalskii A.I., Novoseltseva Zh.A. [Application of machine learning methods in solving problems of productive animal husbandry]. *Problemy biologii productivnykh zhivotnykh - Problems of Productive Animal Biology*, 2018, 4: 98-109. (In Russian).
44. Mitchell R.S., Sherlock R.A., Smith L.A. An investigation into the use of machine learning for determining oestrus in cows. *Computers and electronics in agriculture*. 1996, 15(3): 195-213.
45. Nadimi E.S., .Jørgensen R.N., Blanes-Vidal V., Christensen S. Monitoring and classifying animal behavior using ZigBee-based mobile ad hoc wireless sensor networks and artificial neural networks. *Computers and electronics in agriculture*. 2012, 82: 44-54.
46. Pérez-Elizalde S., Cuevas J., Pérez-Rodríguez P. and Crossa J. Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. *Journal of Agricultural, Biological, and Environmental Statistics*. 2015, 20(4): 512–532.
47. Pietersma D., Lacroix R., Lefebvre R.D., Mwade K. Induction and evaluation of decision trees for lactation curve analysis. *Computers and electronics in agriculture*. 2003, 38(1): 19-32.
48. Pintus M.A., Gaspa G., Nicolazzi E.L., Vicario D., Rossoni A., Ajmone-Marsan P., Nardone A., Dimauro C., Macciotta N.P.P. Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach. *J. Dairy Sci.* 2012, 95(6): 3390-3400.
49. Pinzón-Sánchez C., Cabrera V.E., Ruegg P.L. Decision tree analysis of treatment strategies for mild and moderate cases of clinical mastitis occurring in early lactation. *J. Dairy Sci.* 2011, 94(4): 1873-1892.
50. Pryce J.E., Arias J., Bowman P.J., Davis S.R., Macdonald K.A., Waghorn G.C., Wales W.J., Williams Y.J., Spelman R.J., Hayes B.J. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers *J. Dairy Sci.* 2012, 95(4): 2108-2119.

51. Rodriguez J.J., Kuncheva L.I., and Alonso C.J. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* 2006, 28: 1619–1630.
52. Salehi F., Lacroix R., Wade K.M. Development of neuro-fuzzifiers for qualitative analyses of milk yield. *Computers and electronics in agriculture*. 2000, 28(3): 171-186.
53. Sanzogni L., Kerr D. Milk production estimates using feed forward artificial neural networks. *Computers and electronics in agriculture*. 2001, 32(1): 21-30.
54. Sasaki O., Aihara M., Nishiura A., Takeda H., Satoh M. Genetic analysis of the cumulative pseudo-survival rate during lactation of Holstein cattle in Japan by using random regression models. *J. Dairy Sci.* 2015, 98(8): 5781-5795.
55. Savegnago R.P., Rosa G.J.M., Valente B.D., Herrera L.G.G., Carneiro R.L.R., Sesana R.C., Faro L.E., Munari D.P. Estimates of genetic parameters and eigenvector indices for milk production of Holstein cows. *J. Dairy Sci.* 2013, 96(11): 7284-7293.
56. Shahinfar S., Page D., Guenther J., Cabrera V., Fricke P., Weigel K. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *J. Dairy Sci.* 2014, 97(2): 731-742.
57. Shalev-Shwartz S., Ben-David S. Understanding machine learning: from theory to algorithms. CambridgeUniversity Press, 2014, 449 p.
58. Steeneveld W., van der Gaag L.C., Barkema H.W., Hogeveen H. Simplify the interpretation of alert lists for clinical mastitis in automatic milking systems. *Computers and electronics in agriculture*. 2010, 71(1): 50-56.
59. van Pelt M.L., Meuwissen T.H.E., de Jong G., Veerkamp R.F. Genetic analysis of longevity in Dutch dairy cattle using random regression. *J. Dairy Sci.* 2015, 98(6): 4117-4130.
60. Vapnik V.N. *Vosstanovlenie zavisimostei na osnove empiricheskikh dannykh* [Dependencies reconstruction based on empirical data], Moscow: Nauka, 1979, 449 p. (in Russian).
61. Vapnik V.N. *The nature of statistical learning theory*. Springer, 2000. 311 p.
62. Viugin V.V. *Matematicheskie osnovy teorii mashinnogo obucheniya i prognozirovaniya* (Mathematical foundations of machine learning and prediction theory). Moscow: MCNMO, 2013, 390 p. (in Russian)
63. Yao C., Spurlock D.M., Armentano L.E., Page Jr C.D., VandeHaar M.J., Bickhart D.M., Weigel K.A. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *J. Dairy Sci.* 2013, 96(10): 6716-6729.
64. Zhang F., Murphy M.D., Shalloo L., Ruelle E., Upton J. An automatic model configuration and optimization system for milk production forecasting. *Computers and electronics in agriculture*. 2016, 128: 100-111.

Methods of computer analysis of data in tasks of monitoring and improving herd management

Mikhalskii A.I., Novoseltseva Zh.A.

Trapeznikov Institute of Control Sciences RAS, Russian Federation

ABSTRACT. The modern methods of computer data analysis, used in solving a wide range of tasks arising in productive animal husbandry are reviewed. Some tasks from this area, reviewed also alsow in the author's previous publication, includes a forecast of milk yield, daily yield of milk protein and fat percentage, automatic identification of dairy cows, grazing behavior classification, detection of clinical mastitis during automatic milking, prediction of the outcome of respiratory disease treatment, prognosis of individual performance traits, fertility, substantiation of decisions on cows culling, evaluation of the effectiveness of artificial insemination, the forecast of various product indicators on the basis of genomic evaluations. Optimization of herd management based on the computer estimates is a promising direction for improving the efficiency of modern productive animal husbandry. This article describes the basic methods for analyzing empirical data, including the methods for design of regressions, including the least squares method, Bayesian, ridge regressions, partial regression, nuclear regression and adaptive multidimensional non-parametric spline regression. For classification tasks, the classical discriminant analysis method, the logistic regression method, the naive Bayesian method and k nearest neighbors, the support vector machine, artificial neural networks, decision trees, and a random forest are considered. The appendix contains the main metrics adopted to characterize the quality of the solution of the problems of classification and the design of regressions: accuracy, completeness, metric F1, sensitivity, specificity; ROC and AUC are considered as well. The article material will be useful to specialists of a wide profile, who are interested in applying modern methods of analysis and interpretation of experimental data in the field of animal husbandry

Keywords: dairy cattle, data analysis, computerization, diagnostics, forecasting, herd management

Problemy biologii productivnykh zhivotnykh – Problems of Productive Animal Biology, 2019, 1: 95-111

Поступило в редакцию: 10.12.2018 Получено после доработки: 21.01.2019

Михальский Анатолий Иванович, г.н.с., д.б.н., к.т.н., т. (915)1995526, ipuran@yandex.ru, mpoctok@yandex.ru;
Новосельцева Жанна Анатольевна, с.н.с., к.т.н., т. (495)33488