Разработка алгоритма классификации данных спутникового зондирования на основе машинного обучения на примере гранулометрического состава почв агроландшафтов Западной Сибири

В. В. Чурсин [1, 2], И. В. Кужевская [1], О. Э. Мерзляков [1], Т. О. Валевич [1], К. В. Ручкина [1]

[1] *Национальный исследовательский Томский государственный университет Томск, 634050, Россия*

[2] *Сибирский центр НИЦ «Планета», Новосибирск, 630099, Россия*

E-mail: ivk@ggf.tsu.ru

В статье рассмотрена возможность использования данных дистанционного зондирования Земли с космического аппарата Sentinel-2 и алгоритмов машинного обучения для классификации при создании карт пространственной неоднородности почвенного покрова по гранулометрическому составу земель сельскохозяйственного назначения, а также использования этих карт в точном земледелии. На основе полевых исследований был собран массив данных, включающий спутниковые снимки со значением NDVI менее 0,3 и дополнительно рассчитанные индексы, в том числе связанные со спектральной яркостью (чувствительные к гранулометрическому составу), для обучения и тестирования моделей бинарной классификации, основанных на деревьях решений. Были обучены четыре модели бинарной классификации, использующие алгоритм XGBoost. Подобраны оптимальные значения гиперпараметров для этих моделей и определены наиболее значимые переменные для классификации каждого типа почв. Предложена архитектура нейронной сети, которая в качестве входных данных использует значения спектральной отражаемости, расчётных индексов и результатов бинарной классификации. Точность методики на валидационной выборке составила 76 %.

Ключевые слова: гранулометрический состав почвы, физическая глина, Sentinel, градиентный бустинг, точное земледелие, пространственная неоднородность

Одобрена к печати: 25.02.2021
DOI: 10.21046/2070-7401-2021-18-2-39-50

Введение

Основная проблема при использовании данных спутникового зондирования для определения свойств почвы состоит в сложности компонентов почвы и почвенных спектров. В работе (Ben-Dor, 2002) подробно доказывается, что раз почва содержит много химических компонентов, включая глинистые минералы, карбонаты, органический углерод, воду в различных состояниях, соли и т. д., то некоторые из этих компонентов имеют сильные и чёткие спектральные сигнатуры (например, глинистый минерал монтмориллонит), а некоторые демонстрируют слабые или однородные сигнатуры (например, кварц и полевой шпат). Более того, многие из этих спектральных сигнатур перекрывают друг друга. Например, полосы поглощения в 1,4 и 1,9 мкм являются общими в почвенных спектрах и могут быть вызваны многими химическими компонентами почвы. Кроме того, поглощение в области перекрытия может не быть линейно аддитивным процессом.

Привлечение данных спутникового зондирования как к частным вопросам определения свойств почвы, так и к общим вопросам сельского хозяйства происходит практически с момента использования первых мультиспектральных сенсоров в 1970-е гг. (Ge et al., 2011; Mulla, 2013). Появление данных с разрешением 30 м в 1984 г. (Landsat-5 TM), менее 10 м в 1999 г. (IKONOS) и менее 1 м в 2009 г. (WorldView-2) позволило уже выявлять дефицит содержания азота, оценивать состояние крон плодовых деревьев и оценивать пространственные закономерности в биомассе или урожайности культур и т. д. (Ge et al., 2011). В 2015 г. миссия

Copernicus на базе Sentinel-1 и Sentinel-2 открыла новые возможности, предоставив свободно доступные спутниковые данные с высоким пространственным и временным разрешением.

По мере того как пространственное и спектральное разрешение спутниковых изображений улучшалось, возрастала пригодность данных для их использования в многокомпонентном статистическом анализе и машинном обучении.

Множество авторов приводят различные подходы к обработке спутниковых и наземных данных. На первых этапах разработки методов обработки спутниковых данных использовался регрессионный анализ и множественный регрессионный анализ (MLR — Multiple Linear Regression), регрессия главных компонентов (PCR — Principal Component Regression) и регрессия частичных наименьших квадратов (PLS — Partial Least Squares). Затем в работе (Ge, Thomasson, 2006) был предложен новый метод включения вейвлет-анализа в регрессионный анализ для определения свойств почвы. Получившиеся модели вейвлет-регрессии для свойств почвы имели возможности прогнозирования, аналогичные традиционным методам, при этом вейвлет-модели включали меньше регрессоров. Параллельно в работе (Brown et al., 2006) использовали метод так называемых деревьев ускоренной регрессии (BRT — Boosted Regression Tree) в спектрах отражения почвы от образцов, собранных со всего мира. Было обнаружено, что BRT превосходил PLS при оценке глины, органического углерода в почве, неорганического углерода, железа и ёмкости катионного обмена. Авторы исследования объяснили прогностическую способность BRT способностью учитывать множественные взаимодействия высокого уровня, а также линейные и нелинейные корреляции.

В целом нужно отметить, что большинство литературных источников по теме использования спутниковых данных в классификации свойств почв делится на три типа (Ge et al., 2011): первая категория авторов исследует сами свойства почвы, вторая категория рассматривает методы зондирования и третья — методы анализа данных.

К настоящему моменту исследования по созданию карт свойств почвы приобрели относительно высокую точность в обнаружении химических, физических и биологических свойств почвы (Yuzugullu et al., 2020), таких как показатель кислотности pH (Ballabio et al., 2019), органический углерод (Yigini, Panagos, 2016) и содержание глины (Ballabio et al., 2016), а также в комбинировании географической, спутниковой и наземной информации для интерполяции значений, характеризующих свойства почвы (Shi et al., 2012). Подобные карты достаточно широко используются при управлении сельским хозяйством в странах Европейского союза (ЕС) и Китае (Li et al., 2019).

Сельскохозяйственные почвы часто подвергаются воздействиям, например, таким как движение транспортных средств, которые приводят к уплотнению; обработке почвы и ирригации. Каждое из них влияет на содержание влаги в почве и общий размер частиц почвы, что может оказать серьёзное воздействие на спектр её излучения.

На сегодняшний день научных работ, посвящённых созданию карт свойств почв именно центральных районов Евразийского континента, насчитывается крайне мало.

Цель представленной работы состоит в создании модели классификации по гранулометрическому составу антропогенно-преобразованных почв (Anthrosols) Южной Сибири на основе данных мультиспектральных радиометров. В связи с этим была собрана библиотека полевых данных для обучения компьютерной модели, которая проводит цифровой анализ полей и определяет их структурную неоднородность.

Материалы и методы

Спутниковые данные

Согласно географическим координатам точек полевых исследований, а также смежных пикселей на удалении, не превышающем 15 м, на ESA Sentinels Scientific Data Hub (система Европейских центров морских прогнозов) были подобраны изображения космического аппарата (КА) Sentinel-2 за 2019 г. (уровня обработки Level-2A) в соответствии с заданными условиями: точка не должна быть закрыта облачностью или тенью облака; точка не должна быть

закрыта снегом; снимок должен быть сделан не ранее чем через три дня после последнего выпадения осадков; индекс NDVI (Normalized Difference Vegetation Index — нормализованный разностный вегетационный индекс) не должен превышать значение 0,3, поскольку значения больше 0,3, согласно работе (El-Gammal et al., 2014), соответствуют здоровой растительности с большой сомкнутостью травостоя. Дополнительно каждый снимок прошёл процедуру визуального контроля выявления артефактов. После процедуры отбраковки было отобрано 50 из 74 модельных площадок, которые обеспечены спутниковыми данными.

В итоге для каждой из точек было подобрано от 1 до 7 снимков. Затем для точки были сняты значения спектральной отражательной способности из каналов Sentinel-2 (Gatti, Bertolini, 2015).

Дополнительно были рассчитаны вегетационные и почвенные индексы: NDVI, TVI (Transformed Vegetation Index — трансформированный вегетационный индекс), EVI (Enhanced Vegetation Index — усовершенствованный вегетационный индекс), SATVI (Soil Adjusted Total Vegetation Index — общий индекс растительности с коррекцией по почве), SAVI (Soil-Adjusted Vegetation Index — индекс растительности с коррекцией по почве), MSI (Moisture Stress Index — индекс стресса влажности), GNDVI (Green Normalized Difference Vegetation Index — нормализованный дифференцированный вегетационный индекс зелени), BI (Brightness Index — индекс яркости) и CI (Colour Index — индекс цвета) (Gholizadeha et al., 2018), после чего собранный массив был сопоставлен с данными лабораторных исследований о содержании физической глины в поверхностном слое.

Итоговый массив состоит из 863 строк для обучения и создания мультиклассификационной модели и включает в себя значения спектрального коэффициента отражения (12 переменных), индексов (9 переменных) и процентного содержания физической глины.

Методы машинного обучения

Для решения задачи классификации поверхностного слоя почвы в зависимости от процентного содержания физической глины было решено использовать машинное обучение. Применение машинного обучения направлено на поиск связи между входом $X = [x_1, x_2, ..., x_N]$ и выходом Y. В нашем случае предполагается связь значений спектрального коэффициента отражения, рассчитанных индексов и процентного содержания физической глины в поверхностном слое. Для этой цели мы использовали алгоритм XGBoost — реализацию деревьев решений с градиентным бустингом (*англ*. Gradient boosting). XGBoost относится к группе широко используемых алгоритмов обучения деревьев (He et al., 2014). Дерево решений позволяет осуществлять прогнозирование переменной Y на основе подобранного в процессе обучения ряда условий, расположенных в древовидной структуре, состоящей из ряда узлов ветвления. Последний узел является листом и даёт нам конкретное значение переменной Y (выходной переменной). Основной алгоритм строится путём создания и объединения большого количества слабых моделей, которые дополняют друг друга, что позволяет получить на выходе более точную модель. Это сочетание может быть сделано двумя способами: BAGGING (*англ*. Bootstrap AGGregatING) и бустинг. Градиентный бустинг в отличие от BAGGING строится последовательно. Каждое последующее дерево строится так, чтобы быть максимально коррелированным с градиентом функции потерь, связанной со всей моделью на каждом шаге (Natekin, Knoll, 2013).

Алгоритмы, основанные на деревьях решений, не требуют линейных взаимосвязей между объектами. Они являются значительно лучшими классификаторами, чем другие алгоритмы (Caruana, Niculescu-Mizil, 2006). Кроме того, XGBoost имеет существенное преимущество, связанное с ускорением построения дерева.

Каждый объект в настоящем исследовании обладает набором из 21 признака, упомянутых выше. Выборка была разделена на обучающую и тестовую (объёмом 30 % от всей выборки). Использована бинарная классификация для каждого класса с применением циклической функции, внутри которой происходила бинаризация.

Таким образом, на выходе получаем шесть моделей, которые определяют, соответствует ли данный набор параметров определённому классу. Внутри цикла данные подвергались предварительной обработке. На первом этапе проводилась нормализация данных, затем балансировка выборки путём дублирования наименьшего класса до уровня соответствия количеству наибольшего по числу значений класса. После этого переходили уже непосредственно к обучению модели. Обучение заключается в циклическом переборе гиперпараметров (количество деревьев и их глубина) с параллельной оценкой качества модели. Для оценки использовалась метрика Accuracy (точность). Глубина дерева подбиралась в диапазоне от 2 до 8, а количество деревьев — от 5 до 150 с шагом в 5.

Полевые данные. Площадки исследования

В качестве модельных площадок полевых исследований были выбраны сельскохозяйственные угодья, которые приурочены к зональным и интрозональным типам почв подтайги (*рис. 1а*), лесостепи и степи (*рис. 1б*) Южной Сибири. Основную площадь на территории исследований, согласно «Классификации и диагностики почв СССР» (1977), занимают дерново-подзолистые, серые лесные, аллювиально-дерновые почвы и чернозёмы выщелоченные. Модельные площадки отбирались по правилам заложения, изложенным в ГОСТ 28168-89. Элементарные участки внутри модельных площадок разделялись на секции размером 10×10 м, из которых собирался квадрат 3×3 секции, отбор для смешанной пробы проводился с пяти из них по принципу: один участок — в центре и четыре — по углам диагонали. По описанной схеме в ходе полевых исследований 2017–2019 гг. из пахотных горизонтов почв как нарушенного, так и ненарушенного сложения были отобраны 74 образца почвы. Положение каждой точки отбора проб было записано приёмником GARMIN-64 GPSMap с точностью 3–8 м.

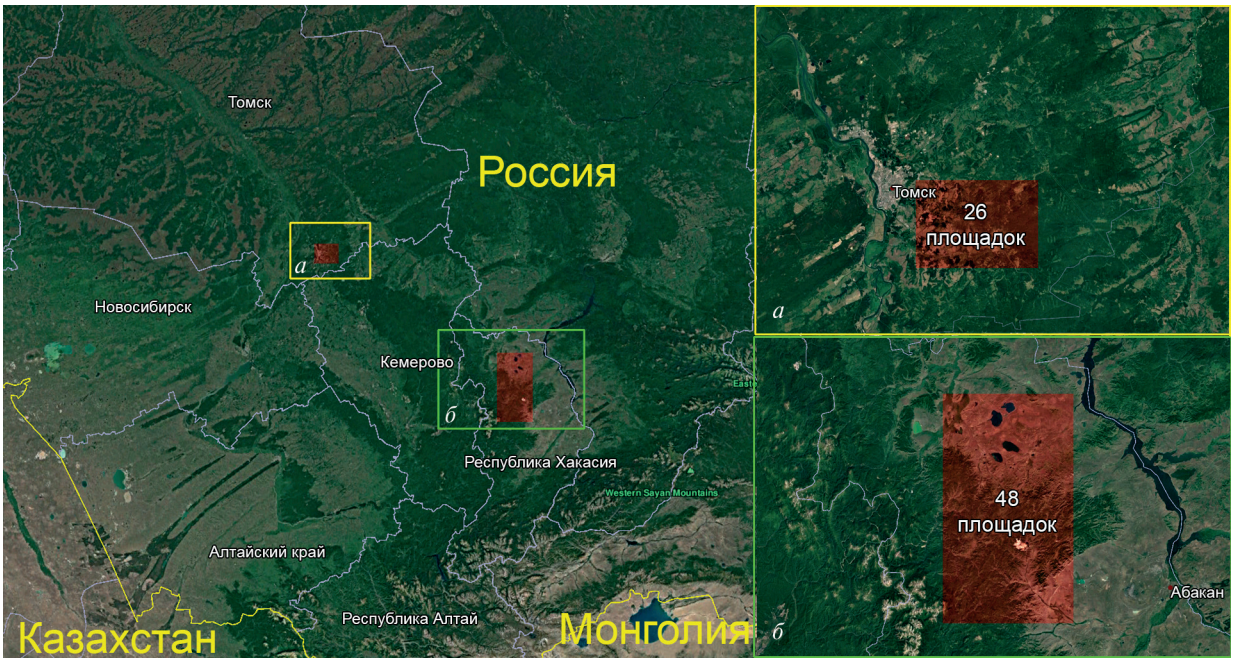

Рис. 1. Территория исследования: *а* —Томская обл.; *б* — Республика Хакасия

Образцы почвы были взяты на глубине 0–10 см, затем высушены на воздухе, измельчены, просеяны (1 мм) и тщательно перемешаны перед анализом в соответствии с требованиями ISO 11464: 2006 и ГОСТ 12071-2014 п. 4.2.3.2.2. Определение гранулометрического состава проводилось на основании рекомендации ГОСТ 12536-2014 п. 4.4. и ISO 11277: 2009.

Результаты и обсуждение

Предварительная обработка

На начальном этапе все данные были разделены на классы в соответствии с Классификацией частиц по размеру и классы гранулометрического состава почв (Jahn et al., 2006). Было получено четыре класса (0–3): супесь, лёгкий суглинок, средний суглинок и тяжёлый суглинок по международной классификации ФАО (Продовольственная и сельскохозяйственная организация ООН, *англ*. FAO — Food and Agriculture Organization), что соответствует понятиям “sandy”, “loam sandy”, “sandy clay loam” и “clay loam”. Так как значения спектрального коэффициента отражения и рассчитанных индексов имеют различные шкалы, то все данные прошли процедуру нормализации. Далее при подготовке к бинарной классификации массив был продублирован четыре раза, согласно количеству выделенных классов, после чего идентификатор класса был заменён на 0 и 1 в зависимости от того, для классификации какого типа почвы предназначен массив. Например, для создания модели бинарной классификации для класса sandy в первом массиве идентификатор этого типа почвы 0 заменён на 1, а все остальные — на 0; во втором массиве для класса loam sandy, идентификатор этого типа (1) не изменился, а все остальные типы почв получили идентификатор 0 и т. д. В итоге получаем четыре идентичных по X массива с нормированными значениями, но с различиями в колонке идентификации (0 или 1).

Спектральный коэффициент отражения различного типа почв в зависимости от длины волны

Интенсивность потока солнечной энергии к поверхности Земли и величина спектрального сигнала почвы в определённой ландшафтной зоне зависят от времени года и суток. Кроме того, ориентация поверхности по отношению к сенсорам радиометра, рассеяние в атмосфере и на окружающих объектах, взаимодействие отражённого от поверхности почвы импульса потока энергии с атмосферой оказывают влияние на величину регистрируемого спектрального сигнала.

Авторы работы (Дробыш и др., 2013) обобщают, что в процессе реализации светопреобразующей функции почвенным покровом ведущая роль принадлежит гумусу, при этом зависимость между содержанием органического вещества в почвах и их спектральной отражательной способностью близка к прямолинейной и носит обратный характер. К тому же рядом исследователей было установлено, что гранулометрический состав почв оказывает весьма существенное влияние на их отражательную способность. Наиболее часто в публикациях упоминаются длины волн видимого спектра (VIS — *от англ*. visible) ($\lambda = 440–690$ нм) и ближнего инфракрасного (ИК) $\lambda = 750$ нм. Вся ближняя ИК-область для определения содержания глины в почве использовалась в работе (Ogen et al., 2017). Чаще всего используют спектральные характеристики VIS (400–700 нм), NIR (*англ*. near infrared, ближний инфракрасный) (700–1100 нм) и SWIR (*англ*. short wave infrared, коротковолновый инфракрасный) (1100–2500 нм).

В настоящей работе внутри класса по гранулометрическому составу при осреднении значений спектральной отражательной способности каждого канала Sentinel-2 была получена спектрограмма (*рис. 2*, см. с. 44). Наибольшая дифференциация значений спектральной отражаемости для классов sandy, loam sandy, sandy clay loam и clay loam отмечается в длинах волн 842 нм (канал B8) и 1610 нм (канал B11). Эти каналы используются при расчёте индексов SATVI, SAVI и MSI.

Для подбора оптимальных значений гиперпараметров моделей общий массив был разделён на обучающую и валидационную выборку в соотношении 70:30 % с сохранением в обоих выборках пропорции присутствия классов. В представленной работе проводился подбор двух гиперпараметров: максимальной глубины деревьев решений (МГД) и количества деревьев в ансамбле (КД); остальные гиперпараметры установлены по умолчанию. Подбор производился

путём оценки абсолютного значения разности (|Accuracy|) между обучающей и валидационной выборкой, при этом соблюдалось требование: абсолютная разность должна стремится к 0, но не должна быть ему равна. То есть выбираются значения гиперпараметров, при которых метрика $|\text{Accuracy}| \to \min$. Затем из ранжированного списка полученных значений гиперпараметров выбирается тот, который имеет максимальную точность на валидационной выборке. Таким образом, можно получить точную модель, которая не имеет склонности к переобучению и в дальнейшем позволит получать стабильные результаты согласно заявленной точности.

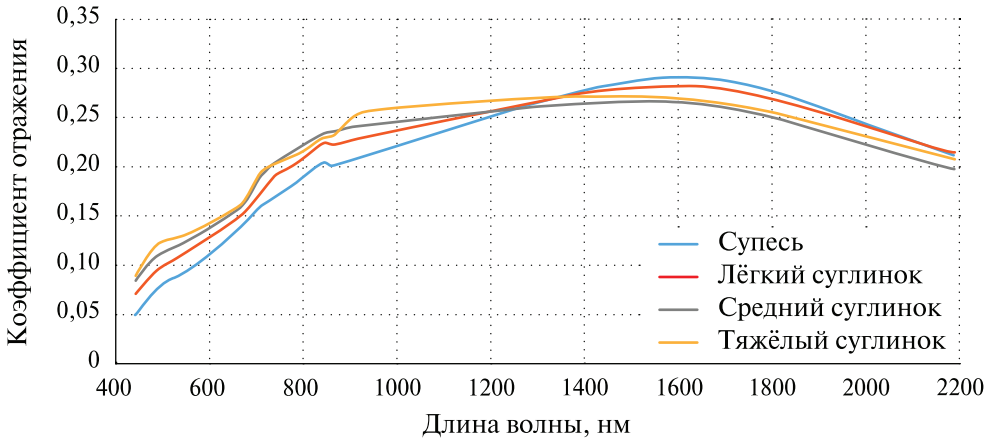

Рис. 2. Средний спектральный коэффициент отражения различных типов почвы в зависимости от длины волны

Итоговые значения выборных гиперпараметров и точность на валидационной выборке для каждого класса представлены в *таблице*.

Значения гиперпараметров и достигнутой метрики для моделей классификации различного гранулометрического состава

Класс	Содержание глины, %	Физическое содержание глины	Гиперпараметры		Точность на валидационной выборке, %
			МГД	КД	
0	0–20	Супесь (sandy)	3	100	88,8
1	20–30	Лёгкий суглинок (loam sandy)	3	25	74,4
2	30–40	Средний суглинок (sandy clay loam)	3	55	76,3
3	40–55	Тяжёлый суглинок (clay loam)	3	75	88,4

Как можно заметить, наибольшую точность удалось достичь при определении классов sandy и clay loam. Наименьшие значения точности отмечаются при выделении классов loam sandy и sandy clay loam. Это можно объяснить тем, что почвы loam sandy и sandy clay loam имеют в своём составе близкое процентное содержание физического песка и физической глины. При этом для всех классов оптимальная глубина дерева равна 3, а количество деревьев варьируется от 25 до 100.

Далее были рассмотрены наиболее значимые переменные. Оценка проводилась путём расчёта частоты использования переменной при делении на ветви. На *рис. 3* приведены первые три переменные и их частота использования в зависимости от того, для какого типа почвы предназначена модель. Нужно отметить, что в качестве первых переменных для 3-й из 4-й моделей выступили непосредственно значения отражательной способности каналов 2 (490 нм), 5 (705 нм), 8 (842 нм) и 9 (940 нм). Авторы работы (Шевырногов и др., 2019)

также отмечают, что использование значений отражательной способности каналов может быть более информативным по сравнению с индексными величинами.

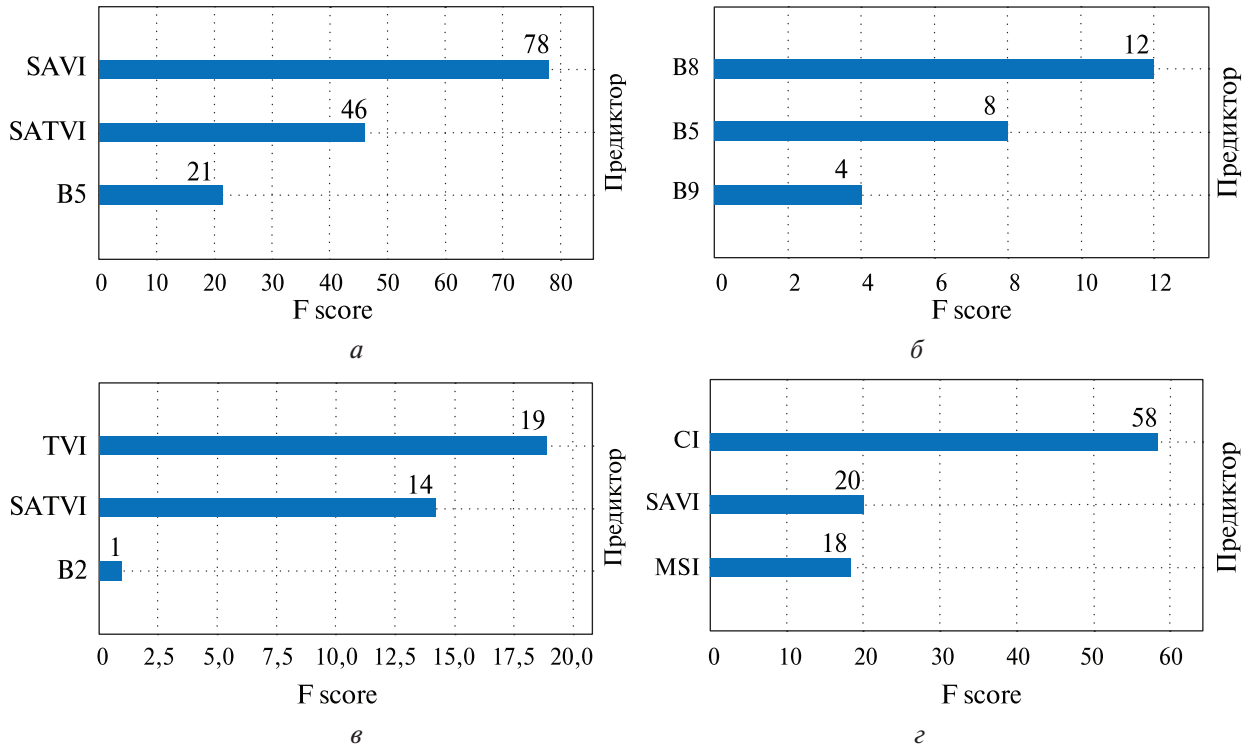

Рис. 3. Повторяемость использования переменных при делении на ветви в зависимости от гранулометрического состава: *а* — супесь (sandy); *б* — лёгкий суглинок (loam sandy); *в* — средний суглинок (sandy clay loam); *г* — тяжёлый суглинок (clay loam)

Подобный подход к оценке переменных позволяет выделить наиболее значимые из них и за счёт этого уменьшить зашумлённость массива и оптимизировать нагрузку на расчётный узел.

Нейронная сеть

Ввиду того, что к каждой записи применяется четыре модели классификации, причём ни одна не имеет гарантированной 100%-й оправдываемости, возможны случаи, когда точка после прохождения через каждую модель может соответствовать двум и более разновидностям почв. Кроме того, если принять допущение, что исходная вероятность принадлежности точки какому-то из классов составляет около 25 % (она может принадлежать только одному из четырёх классов), то в пересчёте результатов по формуле полной вероятности Байеса получим среднее значение оправдываемости мультиклассовой модели, основанной на бинарных классификациях, всего 61 %. Применение этого подхода при составлении карт пространственной неоднородности почвенного покрова неперспективно, однако результаты бинарной классификации могут быть использованы как дополнительные предикторы в более сложных моделях, в частности в нейронных сетях.

Модель нейронной сети принимает на вход значения спектральной отражаемости, индексов и результатов бинарной классификации, основанной на деревьях решения, обученных методом градиентного бустинга. Она состоит из входного слоя, четырёх скрытых слоёв, состоящих из 42 нейронов, одного скрытого слоя, состоящего из 21 нейрона, четырёх слоёв dropout после 2–4-го скрытых слоёв и выходного слоя из четырёх нейронов согласно количеству классов.

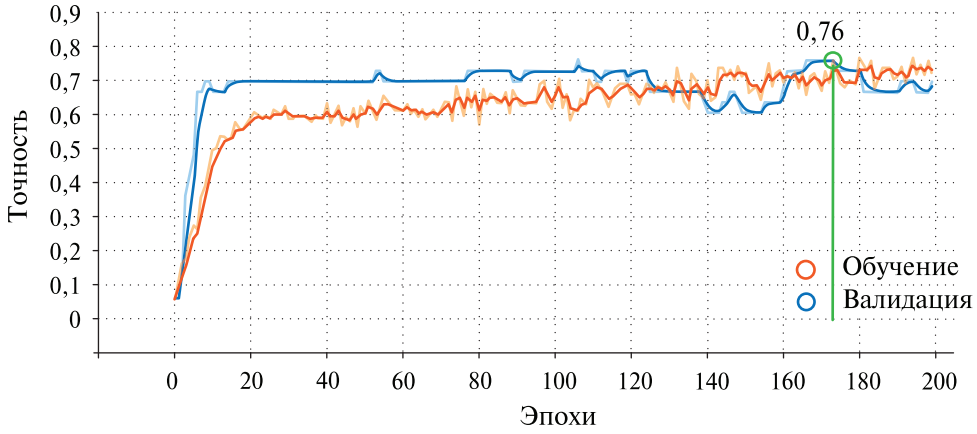

Рис. 4. Точность модели в зависимости от эпохи обучения

Функция активации на скрытых слоях — Rectified linear unit (ReLU, выпрямленная линейная единица), функция активации на выходном слое — Softmax. Оптимизатором выступает Adaptive moment estimation (Adam, метод адаптивной оценки моментов). Коэффициент скорости обучения (*англ.* learning rate) нейронной сети равен 0,0003. Согласно *рис. 4* на первых эпохах точность нейросети быстро возрастает, но затем выходит на плато; благодаря низкому значению коэффициента скорости обучения получается определить, что плато является ложным и точность на нём не максимальная. Наилучшей точности на валидационной выборке при описанной архитектуре удалось достигнуть к 174-й эпохе обучения со значением 0,76, или 76 %.

Пример использования разработанной модели

Рассмотрим применение разработанного алгоритма классификации (общая блок-схема представлена на *рис. 5*) непосредственно на спутниковом снимке для сельскохозяйственных угодий экспериментальных участков Томской обл. Ранее на этих полях производился отбор проб, которые были обработаны и проанализированы в лаборатории Томского государственного университета (ТГУ).

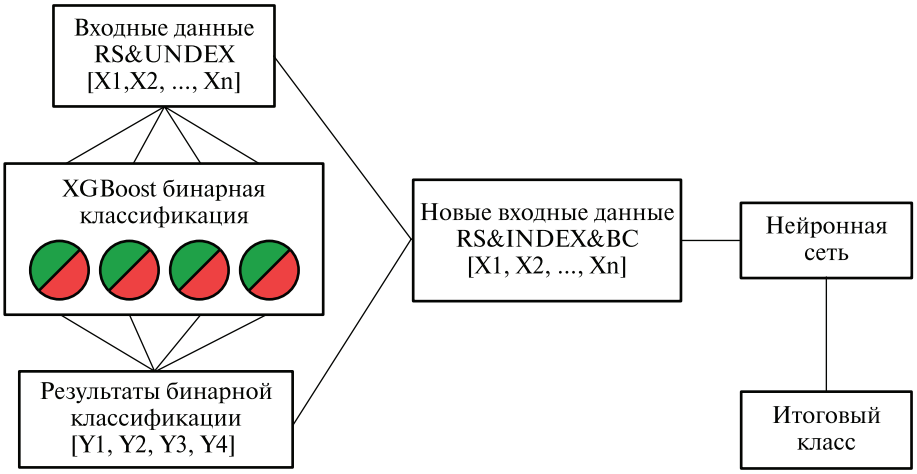

Рис. 5. Блок-схема алгоритма классификации по гранулометрическому составу

Критерии отбора данных описаны в разд. «Спутниковые данные». Данные поступают на вход в модель попиксельно. На выходе получаем карту, отражающую гранулометрический состав почв.

Для тестирования алгоритма был подобран снимок Sentinel-2 за 27.04.2020. С этого изображения были сняты значения спектральной отражаемости для 9 контрольных точек, расположенных на экспериментальных полях территории исследования по координатам отбора проб для лабораторных исследований, и рассчитаны значения спектральных индексов, описанных выше. Из 9 точек верно определены 7, что составляет 78 %. Эти значения незначительно расходятся с представленными выше, что говорит о стабильности работы алгоритма и возможности использования его при создании карт пространственной неоднородности по гранулометрическому составу почв земель сельскохозяйственного назначения.

Известно, что почвы лёгкого гранулометрического состава в большей степени подвержены деградации, от гранулометрического состава зависит качество почв и уровень их плодородия. Такие карты, как на *рис. 6*, позволяют проводить мониторинг полей и выявлять участки, наиболее подверженные неблагоприятным процессам эрозии, подкисления, загрязнения или засоления.

Рис. 6. Визуализация алгоритма классификации гранулометрического состава для двух экспериментальных полей на территории Томской обл.

Учёт неоднородности гранулометрического состава почвы необходим при определении структуры угодий и посевов сельскохозяйственных культур, их размещения, что особенно важно при применении методов точного земледелия.

Заключение

Рассмотрена возможность использования алгоритмов машинного обучения к данным КА Sentinel-2 для их применения в точном земледелии на примере классификации и картографирования антропогенно-преобразованных почв земель сельскохозяйственного назначения Южной Сибири. Получены следующие результаты:

1. Использование возможностей алгоритма XGBoost для анализа весовых коэффициентов входных переменных позволяет оптимизировать алгоритмы, сократить время расчёта и объём вычислительных ресурсов.
2. Методика подбора оптимальных значений гиперпараметров для моделей, включая процесс формирования массива для обучения и тестирования моделей бинарной классификации XGBoost, основанных на деревьях решений, позволяет обеспечить значение точности модели классификации по гранулометрическому составу от 74 до 88 % для отдельного класса и 61 % общей вероятности. Методика перспективна для подбора оптимальных значений гиперпараметров для подобных моделей.
3. Представлена возможность использования результатов бинарной классификации как дополнительных предикторов и интерпретация их как входных данных в модель на основе нейронной сети. При таком подходе точность разработанного алгоритма

составляет 76 % на валидационной выборке, а также обеспечивается стабильность результатов.

4. Анализ точности разделения на специально собранной выборке позволил использовать заявленный алгоритм определения структурной неоднородности почвы при создании карт земель сельскохозяйственного назначения Южной Сибири.

В статье использованы результаты, полученные в ходе выполнения проекта в рамках Программы повышения конкурентоспособности ТГУ.

Литература

1. *Дробыш С. В.*, *Бубнова Т. В.*, *Матыченкова О. В.* Спектральная отражательная способность агродерново-подзолистых почв в зависимости от гранулометрического состава // Почвоведение и агрохимия. 2013. № 1(50). 126–132.
2. *Шевырногов А. П.*, *Ботвич И. Ю.*, *Емельянов Д. В.*, *Ларько А. А.*, *Высоцкая Г. С.*, *Ивченко В. К.*, *Демьяненко Т. Н.* Возможность распознавания почвенного покрова опытного поля с использованием наземных и спутниковых данных // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 4. С. 150–160.
3. *Ballabio C.*, *Panagos P.*, *Monatanarella L.* Mapping topsoil physical properties at European scale using the LUCAS database // Geoderma. 2016. V. 261. P. 110–123.
4. *Ballabio C.*, *Lugato E.*, *Fernández-Ugalde O.*, *Orgiazzi A.*, *Jones A.*, *Borrelli P.*, *Montanarella L.*, *Panagos P.* Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression // Geoderma. 2019. V. 355. Art. No. 113912.
5. *Ben-Dor E.* Quantitative remote sensing of soil properties // Advances in Agronomy. 2002. V. 75. P. 173–243.
6. *Brown D. J.*, *Shepherd K. D.*, *Walsh M. G.*, *Mays M. D.*, *Reinsch T. G.* Global soil characterization with VNIR diffuse reflectance spectroscopy // Geoderma. 2006. V. 132(3–4). P. 273–290.
7. *Caruana R.*, *Niculescu-Mizil A.* An empirical comparison of supervised learning algorithms // Proc. 23rd Intern. Conf. Machine Learning. 2006. P. 161–168. DOI: 10.1145/1143844.1143865.
8. *El-Gammal M. I.*, *Ali R. R.*, *Samra R. A.* NDVI threshold classification for detecting vegetation cover in Damietta governorate, Egypt // J. American Science. 2014. V. 10(8). P. 108–113.
9. *Gatti A.*, *Bertolini A.* Sentinel-2 Products Specification Document. ESA, 2015. Iss. 13.1. 496 p.
10. *Ge Y.*, *Thomasson J. A.* Wavelet incorporated spectral analysis for soil property determination // Trans. ASABE. 2006. V. 49(4). P. 1193–1201. DOI: 10.13031/2013.21719.
11. *Ge Y.*, *Thomasson J. A.*, *Sui R.* Remote sensing of soil properties in precision agriculture: A review // Frontiers of Earth Science. 2011. V. 5. P. 229–238.
12. *Gholizadeha A.*, *Žižalaa D.*, *Saberioonc M.*, *Borůvka L.* Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging // Remote Sensing of Environment. 2018. V. 218. P. 89–103.
13. *He X.*, *Pan J.*, *Jin O.*, *Xu T.*, *Liu B.*, *Xu T.*, *Shi Y.*, *Atallah A.*, *Herbrich R.*, *Bowers S.*, *Candela J. Q.* Practical lessons from predicting clicks on ads at Facebook // Proc. 8th Intern. Workshop on Data Mining for Online Advertising. ACM, 2014. 9 p. URL: https://doi.org/10.1145/2648584.2648589.
14. *Jahn R.*, *Blume H. P.*, *Asio V. B.*, *Spaargaren O.*, *Schad P.* Guidelines for Soil Description. FAO, 2006. 109 p.
15. *Li Z.*, *Taylor J.*, *Frewer L.*, *Zhao C.*, *Yang G.*, *Cheng X.* A Comparative Review of the State and Advancement of Site-Specific Crop Management in the UK and China // Frontiers of Agricultural Science and Engineering. 2019. V. 6(2). P. 116–136.
16. *Mulla D. J.* Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps // Biosystems Engineering. 2013. V. 114(4). P. 358–371.
17. *Natekin A.*, *Knoll A.* Gradient boosting machines, a tutorial // Front Neurorobotics. 2013. V. 7(21). P. 1–21.
18. *Ogen Y.*, *Goldshleger N.*, *Ben-Dor E.* 3D spectral analysis in the VNIR–SWIR spectral region as a tool for soil classification // Geoderma. 2017. V. 302. P. 100–110.
19. *Shi W.*, *Liu J.*, *Du Z.*, *Yue T.* Development of a surface modeling method for mapping soil properties // J. Geographical Sciences. 2012. V. 22. P. 752–760.
20. *Yigini Y.*, *Panagos P.* Assessment of soil organic carbon stocks under future climate and land cover changes in Europe // Science of the Total Environment. 2016. V. 557. P. 838–850.
21. *Yuzugullu O.*, *Lorenz F.*, *Fröhlich P.*, *Liebisch F.* Understanding Fields by Remote Sensing: Soil Zoning and Property Mapping // Remote Sensing. 2020. V. 12. P. 11–16.

Design of satellite sensing data classification algorithm based on machine learning using the example of granulometric composition of soils in agricultural landscapes of Western Siberia

V. V. Chursin [1, 2], I. V. Kuzhevskaya [1], O. E. Merzlyakov [1], T. O. Valevich [1], K. V. Ruchkina [1]

[1] *National Research Tomsk State University*, *Tomsk 634050*, *Russia*
[2] *Siberian Center of SRC Planeta*, *Novosibirsk 630099*, *Russia*
E-mail: ivk@ggf.tsu.ru

The possibility of using Sentinel-2 images and machine learning algorithms to identify and map the spatial heterogeneity of ground cover from the PSD (particle size distribution) of agricultural land, along with the use of precise farming data is discussed. An array was obtained on the basis of field data comprising satellite images with NDVI values <0.3 and additionally computed indices, including those related to spectral brightness (sensitive to PSD), for training and evaluating binary classification models based on solution trees. The XGBoost algorithm was used to train four binary classification models. For these models, the optimum hyperparameter values were chosen and the most important variables for the classification of each type of soil were determined. The architecture of the neural network, including spectral reflectivity values, calculated indices and effects of binary classification, was suggested as input data. The precision of the designed procedure on the validation set reached 76 %.

Keywords: soil texture, clay content, Copernicus mission, Sentinel, multispectral imagery, gradient boosting, mapping of soils

Accepted: 25.02.2021
DOI: 10.21046/2070-7401-2021-18-2-39-50

References

1. Drobysh S. V., Bubnova T. V., Matychenkova O. V., Impact granulometric composition on the spectral reflectivity of the agro-sod-podzolic soils, *Pochvovedenie i agrokhimiya*, 2013, Vol. 50, No. 1, pp. 126–132 (in Russian).
2. Shevyrnogov A. P., Botvich I. Yu., Emelyanov D. V., Larko A. A., Vysotskaya G. S., Ivchenko V. K., Demyanenko T. N., Possibilities of experimental field soil cover recognition using ground and satellite data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, Vol. 16, No. 4, pp. 150–160 (in Russian).
3. Ballabio C., Panagos P., Monatanarella L., Mapping topsoil physical properties at European scale using the LUCAS database, *Geoderma*, 2016, Vol. 261, pp. 110–123.
4. Ballabio C., Lugato E., Fernández-Ugalde O., Orgiazzi A., Jones A., Borrelli P., Montanarella L., Panagos P., Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression, *Geoderma*, 2019, Vol. 355, Art. No. 113912.
5. Ben-Dor E., Quantitative remote sensing of soil properties, *Advances in Agronomy*, 2002, Vol. 75, pp. 173–243.
6. Brown D. J., Shepherd K. D., Walsh M. G., Mays M. D., Reinsch T. G., Global soil characterization with VNIR diffuse reflectance spectroscopy, *Geoderma*, 2006, Vol. 132, No. 3–4, pp. 273–290.
7. Caruana R., Niculescu-Mizil A., An empirical comparison of supervised learning algorithms, *Proc. 23rd Intern. Conf. Machine Learning*, 2006, pp. 161–168, DOI: 10.1145/1143844.1143865.
8. El-Gammal M. I., Ali R. R., Samra R. A., NDVI threshold classification for detecting vegetation cover in Damietta governorate, Egypt, *J. American Science*, 2014, Vol. 10, No. 8, pp. 108–113.
9. Gatti A., Bertolini A., *Sentinel-2 Products Specification Document*, ESA, 2015, Issue 13.1, 496 p.
10. Ge Y., Thomasson J. A., Wavelet incorporated spectral analysis for soil property determination, *Trans. ASABE*, 2006, Vol. 49, No. 4, pp. 1193–1201, DOI: 10.13031/2013.21719.
11. Ge Y., Thomasson J. A., Sui R., Remote sensing of soil properties in precision agriculture: A review, *Frontiers of Earth Science*, 2011, Vol. 5, pp. 229–238.
12. Gholizadeha A., Žižalaa D., Saberioonc M., Borůvka L., Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging, *Remote Sensing of Environment*, 2018, Vol. 218, pp. 89–103.

13. He X., Pan J., Jin O., Xu T., Liu B., Xu T., Shi Y., Atallah A., Herbrich R., Bowers S., Candela J. Q., Practical lessons from predicting clicks on ads at Facebook, *Proc. 8th Intern. Workshop on Data Mining for Online Advertising*, ACM, 2014, 9 p., available at: https://doi.org/10.1145/2648584.2648589.
14. Jahn R., Blume H. P., Asio V. B., Spaargaren O., Schad P., *Guidelines for Soil Description*, FAO, 2006, 109 p.
15. Li Z., Taylor J., Frewer L., Zhao C., Yang G., Cheng X. A., A Comparative Review of the State and Advancement of Site-Specific Crop Management in the UK and China, *Frontiers of Agricultural Science and Engineering*, 2019, Vol. 6, No. 2, pp. 116–136.
16. Mulla D. J., Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps, *Biosystems Engineering*, 2013, Vol. 114, No. 4, pp. 358–371.
17. Natekin A., Knoll A., Gradient boosting machines, a tutorial, *Front neurorobotics*, 2013, Vol. 7, No. 21, pp. 1–21.
18. Ogen Y., Goldshleger N., Ben-Dor E., 3D spectral analysis in the VNIR–SWIR spectral region as a tool for soil classification, *Geoderma*, 2017, Vol. 302, pp. 100–110.
19. Shi W., Liu J., Du Z., Yue T., Development of a surface modeling method for mapping soil properties, *J. Geographical Sciences*, 2012, Vol. 22, pp. 752–760.
20. Yigini Y., Panagos P., Assessment of soil organic carbon stocks under future climate and land cover changes in Europe, *Science of the Total Environment*, 2016, Vol. 557, pp. 838–850.
21. Yuzugullu O., Lorenz F., Fröhlich P., Liebisch F., Understanding Fields by Remote Sensing: Soil Zoning and Property Mapping, *Remote Sensing*, 2020, Vol. 12, pp. 11–16.