УДК 551.510.535
DOI: 10.62955/0135-3748-2026-407

АППРОКСИМАЦИЯ МЕТОДИКИ ВЫДЕЛЕНИЯ ИОНОСФЕРНЫХ СИГНАЛОВ НА РАДАРАХ SUPERDARN/СЕКИРА ФИЗИЧЕСКИ-ИНФОРМИРОВАННОЙ МОДЕЛЬЮ МАШИННОГО ОБУЧЕНИЯ

В.Д. Басов[1], О.И. Бернгардт[1,2]

1Иркутский государственный университет, Иркутск, Россия
2Институт солнечно-земной физики Сибирского отделения Российской академии наук, Иркутск, Россия

APPROXIMATION OF THE SUPERDARN IONOSPHERIC SCATTER IDENTIFICATION TECHNIQUE BY A PHYSICALLY-INFORMED MACHINE LEARNING MODEL

V.D.Basov[1], O.I. Berngardt[1,2]

1 Irkutsk State University, Irkutsk, Russia
2 Institute of Solar-Terrestrial Physics SB RAS, Irkutsk, Russia

Аннотация. В работе представлен гибридный подход к классификации типов рассеяния сигналов радаров SuperDARN, являющийся адаптацией известного ранее экспертного алгоритма пространственно-временной кластеризации ансамблевой моделью машинного обучения на расширенном наборе физических признаков. Предложенный подход позволяет имитировать логику предыдущего экспертного алгоритма, используя параметры распространения радиоволн. Экспериментальная проверка с использованием качественного физического критерия (дальности до 500 км) показала, что качество разработанной модели превосходит качество предыдущего алгоритма на 3.6 % выше. Обученная модель оформлена в виде независимого программного модуля для использования на других радарах сетей SuperDARN/СЕКИРА.

Ключевые слова: SuperDARN, ионосферное рассеяние, машинное обучение, классификация.

Abstract. A hybrid approach to the classification of SuperDARN radar scatter types is presented, which adapts a previously known expert spatial-temporal clustering algorithm using an ensemble machine learning model on an extended set of physical features. The proposed approach imitates the logic of the previous expert algorithm by utilizing radio wave propagation parameters. Experimental verification using a qualitative physical criterion (ranges up to 500 km) showed that the developed model outperforms the previous algorithm by 3.6 %. The trained model is packaged as an independent software module for use with other SuperDARN radars.

Keywords: SuperDARN, ionospheric scatter, machine learning, classification.

ВВЕДЕНИЕ

Сеть высокочастотных когерентных радаров SuperDARN является одним из основных инструментов для диагностики ионосферы и верхней атмосферы [Nishitani et al., 2019]. Данные радаров содержат информацию о доплеровской скорости и спектральной ширине обратного рассеяния радиоволн, и для их корректной интерпретации требуется уверенное автоматическое разделение ионосферного рассеяния (IS) и наземного/морского рассеяния (GS).

Пороговые методы, основанные на скорости и ширине спектра [Blanchard et al., 2009], недостаточно эффективны на средних широтах, где характеристики двух типов рассеяния перекрываются. Алгоритм пространственно-временной кластеризации, предложенный Ribeiro et al. [2011], позволяет улучшить качество классификации за счет учета пространственно-временной структуры сигналов. Однако его существующие реализации, разработанные около восьми лет назад для операционной системы Linux, требуют адаптации к современным вычислительным средам и форматам данных. Кроме того, алгоритм является эмпирическим и не использует всю доступную физическую информацию о сигнале — в частности, параметры распространения радиоволн в ионосфере.

В данной работе предлагается гибридный (физически-информированный) подход, имитирующий алгоритм Рибейро ансамблевыми моделями машинного обучения на расширенном наборе признаков. Целью является создание инструмента, имитирующего логику экспертного алгоритма, но использующего более полную физическую информацию и работающего при произвольной постановке эксперимента.

АДАПТАЦИЯ АЛГОРИТМА РИБЕЙРО

Исходная реализация алгоритма Рибейро [Ribeiro et al., 2011] была написана на языке Python с использованием библиотек scikit-learn, numpy, pandas и matplotlib. Код разрабатывался под Linux и использовал устаревшие версии библиотек, что приводило к ряду проблем при запуске в среде Windows с актуальными версиями.

Первая проблема была связана с нестандартным для радаров форматом входных данных pickle/sqlite, в то время как обычные исходные данные радаров предоставляются в стандартном для радаров формате FITACF. Для решения этой проблемы был разработан модуль-конвертер FITACF-2-PICKLE.py, выполняющий чтение FITACF-файлов через библиотеку PyDARN, группировку записей в сканы, преобразование форматов полей и сериализацию в pickle.

Ключевой задачей конвертера стала корректная группировка записей в сканы. В формате FITACF каждая запись соответствует одному лучу радара, тогда как фреймворк ожидает данные, сгруппированные по сканам (один скан — один полный проход по всем лучам). Первоначальная версия конвертера использовала временные разрывы между записями для определения границ сканов, однако этот подход оказался неработоспособным: радар работает непрерывно и реальные разрывы между записями составляют доли секунды. В исправленной нами версии группировка выполняется через отслеживание изменений флага scan в данных FITACF, который принимает значение 1 для первого луча каждого нового скана и 0 для всех остальных. После исправления количество сканов для суточного объема данных увеличилось с 1 до 900, что соответствует регулярному режиму работы радара.

Вторая проблема была связана с изменением API-библиотек. Современные версии используемых библиотек (numpy) требуют строгого соответствия размерностей массивов, что вызывало ошибки в данных с неоднородными размерами, поскольку реальные данные радара содержат разное количество точек рассеяния на разных лучах. Эта проблема была исправлена.

Третья проблема касалась визуализации, которая на нерегулярных разреженных данных давала нечитаемые результаты, что было решено заменой используемых алгоритмов рендеринга на алгоритмы непосредственного отображения всех данных.

Полученные модернизации позволили запустить и проверить алгоритм на реальных данных радаров.

Адаптированный алгоритм был применен к данным радара CVE (Christmas Valley East) за 2021 г. Тестирование подтвердило эффективность метода Рибейро по сравнению с пороговыми методами Бланшарда: оригинальный метод Бланшарда практически полностью отнес все сигналы к GS, модифицированная версия выделила IS с заметным шумом, тогда как метод Рибейро дал четкую границу между двумя типами рассеяния.

РАСШИРЕННЫЙ НАБОР ПРИЗНАКОВ

Алгоритм Рибейро использует пространственно-временную структуру скорости и спектральной ширины для определения областей с ионосферным и земным рассеянием. Для улучшения классификации были привлечены дополнительные физические признаки, вычисляемые численным моделированием распространения радиоволн в ионосфере с использованием измеряемого радаром угла места прихода рассеянного сигнала. В результате для каждой измеренной точки вычислялись дополнительно: высота отражения, количество скачков и направление луча радиоволны на различных участках траектории.

Для обучения модели машинного обучения, имитирующей алгоритм Рибейро, был создан обучающий датасет. Обработка проводилась для нечетных дней (1, 3, 5, 7, 9, 11) каждого месяца 2021 г., за исключением летних (июнь–август), данные за которые отсутствовали. Все выходные файлы были объединены в единый CSV-файл, содержащий 6.4 миллиона записей.

Формирование обучающей выборки потребовало слияния двух источников данных: плоской таблицы с признаками и CSV-файлов с разметкой алгоритмом Рибейро, где данные хранились в свернутом виде (одна строка — один скан). Слияние выполнялось по составному ключу (дата, время, луч). Итоговый датасет составил 5.9 миллиона записей с распределением классов 33 % IS и 67 % GS.

МОДЕЛИ И РЕЗУЛЬТАТЫ

Для классификации были выбраны ансамблевые методы Random Forest и XGBoost, что обосновано их преимуществом над нейронными сетями на основе табличных данных [Grinsztajn et al., 2022; Shwartz-Ziv, Armon, 2022]. Оба метода работают на центральном процессоре и не требуют графического ускорителя, что делает их пригодными для использования на доступном оборудовании.

Random Forest (100 деревьев, глубина 15) достиг Accuracy 0.79 и F1-score 0.86. Анализ метрик по классам выявил проблему, характерную для несбалансированных данных: метрика Recall для IS составила 0.51, в то время как для GS — 0.90. Попытка исправить ситуацию с помощью балансировки датасета привела к перераспределению метрик: Recall для IS вырос до 0.74, но общий Accuracy снизился до 0.75.

XGBoost (200 деревьев, глубина 8, learning rate 0.1) показал Accuracy 0.78 и F1-score 0.85. AUC-ROC составил 0.87, AUC-PR — 0.91, что свидетельствует о высокой разделяющей способности.

Попиксельное сравнение разметки с помощью алгоритма Рибейро и предсказаний XGBoost для одного дня наблюдений (рис. 1) показало совпадение 82.6 %. Визуально картина очень близка: модель корректно выделяет обширную область GS и компактную область IS на ближних дистанциях.

Для оценки стабильности работы модели XGBoost в различных геофизических условиях было проведено сравнение для четырех сезонов: зима (1, 3, 5 января), весна (1, 3, 5 апреля), лето (1, 3, 5 сентября, поскольку данные за июнь—август отсутствуют) и осень (1, 3, 5 октября). Модель XGBoost продемонстрировала устойчивое качество классификации во все сезоны. В отдельных временных интервалах наблюдаются расхождения с разметкой методом Рибейро, что может быть связано как с ошибками самой разметки, так и с сезонными особенностями распространения радиоволн.

Следует отметить, что визуальное сравнение по сезонам не позволяет сделать однозначный вывод о том, какой из двух методов (Рибейро или XGBoost)

ближе к истине — оно лишь показывает степень их согласованности. Для объективной верификации необходим независимый физический критерий, не опирающийся на разметку методом Рибейро.

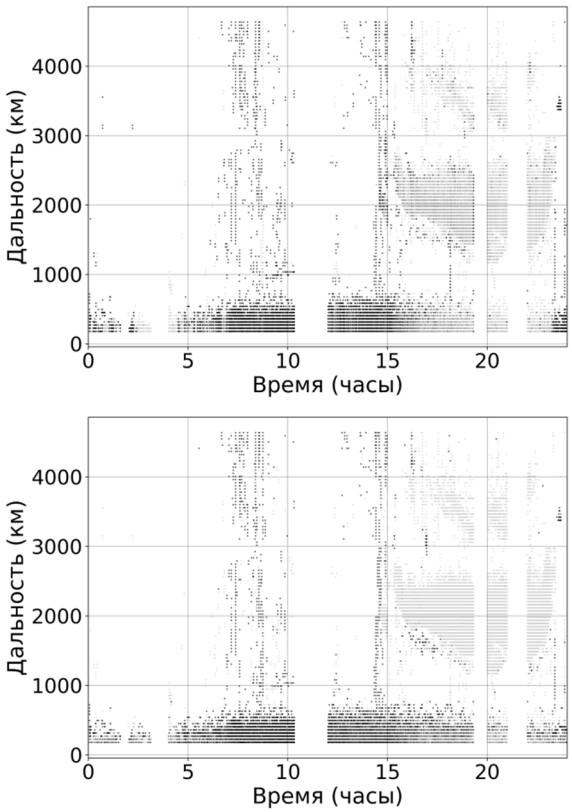

Рис. 1. Сравнение разметки методом Рибейро (сверху) с разметкой моделью XGBoost (снизу): черный цвет — IS, серый цвет — GS

Результаты экспериментов с порогом классификации для XGBoost приведены в табл. 1. При повышении порога Recall для IS растет (от 0.55 до 0.88), а Recall для GS падает (от 0.89 до 0.57). Выбор порога 0.5 обеспечивает равный Recall для обоих классов (0.76), что является разумным компромиссом для большинства научных задач. При необходимости приоритизации одного из типов рассеяния порог может быть сдвинут в соответствующую сторону без переобучения модели, что делает ее гибким инструментом для различных приложений.

Таблица 1. Зависимость метрик от порога

Порог	Recall IS	Recall GS	F1-score
0.3	0.55	0.89	0.86
0.5	0.76	0.76	0.82
0.7	0.88	0.57	0.71

Анализ важности признаков по XGBoost (табл. 2) показал, что наибольший вклад в классификацию вносит доплеровская скорость (46 %), что согласуется с физическими представлениями о различии IS и GS. Вторым по значимости оказался косинус вертикального угла на трех четвертях траектории (16 %). Группа угловых параметров в совокупности обеспечивает около 30 % важности, что подтверждает обоснованность их включения в набор признаков. Спектральная ширина, являющаяся ключевым параметром в пороговых методах Бланшарда, оказалась лишь на шестом месте (3.7 %), что может свидетельствовать о ее меньшей информативности для среднеширотных данных по сравнению с полярными.

Таблица 2. Важность признаков по XGBoost

Признак	Важность
Доплеровское смещение сигнала (скорость)	0.458
Наклон траектории на 3/4 пути	0.155
Число скачков траектории	0.110
Высота рассеяния	0.097
Наклон траектории на 2/4 пути	0.048
Спектральное уширение сигнала	0.037
Наклон траектории на 1/4 пути	0.034
Угол траектории с магнитным полем Земли в точке рассеяния	0.031
Наклон траектории в точке рассеяния	0.030

ОБЪЕКТИВНЫЙ КРИТЕРИЙ

Для проверки, не зависящей от субъективной разметки, был использован физический критерий: на дальностях до 500 км рассеяние от земной поверхности не может наблюдаться из-за геометрии распространения радиоволн, поэтому ожидаемая доля IS должна быть близка к 100 %. Анализ показал, что алгоритм Рибейро классифицирует как IS лишь 50.9 % точек на этих дальностях, а модель XGBoost — 54.5 %, что на 3.6 % выше. Это указывает на систематическую ошибку в разметке алгоритмом Рибейро и демонстрирует способность модели XGBoost частично ее компенсировать за счет использования дополнительных физических признаков.

Следует подчеркнуть, что данный результат не означает, что модель XGBoost «лучше» алгоритма Рибейро в абсолютном смысле. Алгоритм Рибейро является признанным экспертным методом, и модель обучалась на его разметке. Однако использование дополнительных физических признаков позволяет модели XGBoost частично исправлять те ошибки, которые сам алгоритм Рибейро допускает из-за ограниченности исходного набора параметров (только скорость и ширина спектра).

Для сравнения с нейросетевым подходом был обучен многослойный персептрон (MLP) с архитектурой 64-32-16, что привело к практически идентичным метрикам качества (Accuracy 0.7932 против 0.7929). Это согласуется с выводами [Grinsztajn et al., 2022] о конкурентоспособности tree-based моделей.

ЗАКЛЮЧЕНИЕ

Предложенный гибридный подход позволил создать инструмент, имитирующий логику экспертного алгоритма Рибейро с F1-score до 0.86. Модель XGBoost превосходит экспертный алгоритм на объективном физическом критерии, обеспечивает гибкое управление порогом

классификации и работает на порядок быстрее. Нейросетевой подход (MLP) не показал преимуществ. Обученная модель оформлена как независимый программный модуль, что позволяет использовать ее на других радарах сети SuperDARN/СЕКИРА.

Авторы выражают благодарность за возможность использования данных SuperDARN. SuperDARN — это сеть радаров, финансируемых национальными научными фондами Австралии, Канады, Китая, Франции, Италии, Японии, Норвегии, Южной Африки, Великобритании и Соединенных Штатов Америки.

Работа О.И. Бернгардта выполнена при финансовой поддержке Минобрнауки России.

СПИСОК ЛИТЕРАТУРЫ

Blanchard G.T., Sundeen S., Baker K.B. Probabilistic identification of high-frequency radar backscatter from the ground and ionosphere based on spectral characteristics // Radio Sci. 2009. V. 44, RS5012. DOI: 10.1029/2009RS004141.

Grinsztajn L., Oyallon E., Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? // Advances in Neural Information Processing Systems. 2022. V. 35. P. 507–520.

Nishitani N., Ruohoniemi J.M., Lester M. et al. Review of the accomplishments of mid-latitude Super Dual Auroral Radar Network (SuperDARN) HF radars // Prog. Earth Planet. Sci. 2019. V. 6, N 27. DOI: 10.1186/s40645-019-0270-5.

Ribeiro A.J., Ruohoniemi J.M., Baker J.B.H. et al. A new approach for identifying ionospheric backscatter in midlatitude SuperDARN HF radar observations // Radio Sci. 2011. V. 46, RS4011. DOI: 10.1029/2011RS004676.

Shwartz-Ziv R., Armon A. Tabular data: Deep learning is not all you need // Information Fusion. 2022. V. 81. P. 84–90.