УДК 621.396.969, 551.501.81, 551.594.9, 51-73
DOI: 10.12737/szf-112202503

Поступила в редакцию 11.12.2024
Принята к публикации 28.04.2025

УПРАВЛЯЕМЫЙ ДАННЫМИ ПОДХОД К КЛАССИФИКАЦИИ ДАННЫХ СРЕДНЕШИРОТНЫХ РАДАРОВ КОГЕРЕНТНОГО РАССЕЯНИЯ

DATA-DRIVEN APPROACH TO MID-LATITUDE COHERENT SCATTER RADAR DATA CLASSIFICATION

О.И. Бернгардт
Институт солнечно-земной физики СО РАН,
Иркутск, Россия, berng@iszf.irk.ru

O.I. Berngardt
Institute of Solar-Terrestrial Physics SB RAS,
Irkutsk, Russia, berng@iszf.irk.ru

Аннотация. Развит самосогласованный, управляемый данными подход к классификации данных, получаемых на среднеширотных радарах когерентного рассеяния ИСЗФ СО РАН. На основе материала 2021 г. приведено решение задачи автоматической классификации данных без их разметки экспертом и без постулирования количества классов. Алгоритм самостоятельно проводит разметку, определяет оптимальное количество классов сигналов, наблюдаемых радарами, и обучает двухслойную классифицирующую нейронную сеть предельно простой структуры. При траекторных расчетах используется метод волновой оптики и международные ссылочные модели ионосферы и магнитного поля Земли. Модель обучена на сигналах, приходящих с главного лепестка диаграммы направленности. При обучении для адаптации части данных, полученных с повышенным спектральным разрешением проводится их искусственное загрубление до стандартного разрешения. Каждый класс сигнала, определенный нейронной сетью, проинтерпретирован с физической точки зрения исходя из статистических характеристик сигналов, принадлежащих ему. Показано, что количество классов в данных составляет от 23 до 35. Проведена оценка значимости различных параметров входных данных. Показано, что наиболее важными для классификации параметрами являются расчетные высота рассеяния и наклон траектории в точке рассеяния, а наименее важными — спектральная ширина принятого сигнала и расчетное количество отражений от нижележащей поверхности.

Ключевые слова: декаметровый радар, СЕКИРА, ионосфера, автоматическая классификация.

Abstract. A self-consistent, data-driven approach to classifying data obtained at the ISTP SB RAS mid-latitude coherent scatter radars has been developed. Based on 2021 data, a solution of the problem of automatic data classification is presented without their labeling by an expert and without postulating the number of classes. The algorithm automatically labels the data, determines the optimal number of signal classes observed by the radars, and trains a two-layer classifying neural network of an extremely simple structure. The trajectory calculations use the wave optics method and international reference models of the ionosphere and the geomagnetic field. The model is trained on signals coming from the main lobe of the antenna pattern. During training, to adapt part of the data obtained with improved spectral resolution, it is artificially coarsened to the standard resolution. Each signal class determined by the neural network is interpreted from a physical point of view, using statistical characteristics of the signals belonging to it. The number of classes in the data is demonstrated to range from 23 to 35. The significance of various parameters of the input data is assessed. It is shown that the most important parameters for the classification are the calculated scattering height and the elevation of the trajectory at the scattering point, and the least important are the spectral width of the received signal and the calculated number of reflections from the underlying surface.

Keywords: decameter radar, SECIRA, ionosphere, automatic classification.

ВВЕДЕНИЕ

Часто проблемой интерпретации данных являются несколько возможных сценариев их объяснения, выбор из которых может быть субъективным и зависеть от интерпретатора. Поэтому важной является независимость интерпретации результатов от интерпретатора. Задача может быть сформулирована как управляемый данными подход — построение моделей на основе объективной информации, содержащейся в данных. В работе излагается самосогласованный, управляемый данными подход к решению задачи классификации обработанных данных радаров когерентного рассеяния ИСЗФ СО РАН с точки зрения радиофизических механизмов формирования и распространения этих сигналов.

Российская сеть когерентных радаров СЕКИРА [Бернгардт и др., 2020] состоит из радаров, близких радарам международной сети SuperDARN [Greenwald et al., 1995; Chisham et al., 2007; Nishitani et al., 2019] по программному и аппаратному обеспечению. Радары СЕКИРА представляют собой программно-модифицированные стереорадары CUTLASS [Lester et al., 2004]. Интерпретация принимаемых сигналов обычно начинается с классификации данных на различные типы (классы). Основными типами рассеянных сигналов являются: ионосферное рассеяние

от магнитоориентированных неоднородностей, рассеяние от подстилающей поверхности (земной и морской), рассеяние на метеорных следах, ближнее эхо (near-range echo) на высотах E-слоя ионосферы и другие [Nishitani et al., 2019].

Задача классификации данных радаров когерентного рассеяния методами машинного обучения на два класса (ионосферное рассеяние и рассеяние от земной поверхности) рассматривалась, например, в работах [Ponomarenko et al., 2007; Blanchard et al., 2009], где было показано, что для разделения данных на два класса достаточно очень простой модели — всего несколько свободных параметров. В работе [Ribeiro et al., 2011] для кластеризации на два кластера используется интуитивный алгоритм, в основе своей близкий к алгоритму DBSCAN [Ester et al., 1996]. В работе [Kunduri et al., 2022] используется аналог DBSCAN для разделения данных на 9 кластеров, при этом данные предварительно преобразуются в вероятности различных классов нейросетевой моделью, обученной на синтетическом наборе данных, генерирующем сигналы этих 9 классов (рассеяние в ионосфере на скачках 0.5, 1, 1.5 и 2, а также рассеяние от земной/морской поверхности). В работе [Kong et al., 2024] рассматривается задача кластеризации внутренних (латентных) представлений сигналов, извлеченных из них автоэнкодером [Rumelhart et al., 1986; Goodfellow et al., 2016]. Решение более сложной задачи разбиения на 20 классов для случая самообучающейся сети, когда кластеризация обучает классификатор, было предложено в [Berngardt et al., 2022; Бернгардт, 2022].

Из сравнения алгоритмов [Ponomarenko et al., 2007; Blanchard et al., 2009] и [Berngardt et al., 2022; Бернгардт, 2022] стоило ожидать, что решение задачи разделения на 20 классов возможно моделью со сравнительно небольшим числом (несколькими сотнями) неизвестных параметров. Однако решение, приведенное в [Бернгардт, 2022], требует ~30000 свободных параметров и применения полиномиального сглаживающего пространства на входе, что делает модель избыточно сложной. Модель применяется в настоящее время на радарах ИСЗФ СО РАН, однако для ее усовершенствования необходимо ответить на следующие вопросы.

1. Как учитывать тип радара и режим его работы при обучении и использовании модели?

2. Как определить число классов рассеянных сигналов в данных без привлечения эксперта?

3. Какие характеристики принятого сигнала сильнее всего влияют на качество его классификации?

Целью работы является улучшение этого метода. Показано, что применение управляемого данными подхода позволяет ответить на указанные выше вопросы.

МОДЕЛЬ РАСПРОСТРАНЕНИЯ СИГНАЛА И ВХОДНЫЕ ПАРАМЕТРЫ МОДЕЛИ

Предложенный в работах [Berngardt et al., 2022; Бернгардт, 2022] подход заключается в использовании неразмеченных данных для создания их классификатора. Получаемая при этом нейронная сеть представляет собой схему, похожую на автоэнкодер [Rumelhart et al., 1986; Goodfellow et al., 2016], но с множеством голов-декодеров, где каждая голова (декодер) тренируется отдельно разметкой, созданной неким кластеризатором по отдельному эксперименту. Голова декодера представляет собой полносвязный слой $y_j = SoftMax\left(\sum_i A_{ij} x_i\right)$ с неотрицательными коэффициентами $A_{ij} \geq 0$ — проекцию множества скрытых классов x_i на множество кластеров y_j, по смыслу близкую к определению полной вероятности через условные. Энкодер для всех голов автоэнкодера един и является искомым классификатором данных на скрытые классы x_i. Схема нейронной сети и ее тренировки приведена на рис. 1, *а*. Считается, что автоэнкодер эффективно работает, когда количество скрытых классов не меньше, чем реальное количество переменных, необходимых для точного решения задачи [Goodfellow et al., 2016].

Предложенный в [Berngardt et al., 2022] подход состоит из двух этапов. Первый этап (кластеризация) — разделение данных на слабо пересекающиеся классы. На этом этапе используется метод смеси гауссовых функций (GM), при котором предполагается, что данные в каждом кластере подчиняются многомерному гауссову распределению с неизвестными параметрами, их число равно 20. К недостаткам этого подхода относится сложность его обоснования: существует множество различных методов кластеризации, и их выбор будет приводить к различному разбиению данных на кластеры. В работе [Ribeiro et al., 2011] для кластеризации на два кластера используется алгоритм, близкий к алгоритму DBSCAN. В работе [Kunduri et al., 2022] используется аналог DBSCAN для разделения данных на 9 кластеров (рассеяние в ионосфере на скачках 0.5, 1, 1.5 и 2, а также рассеяние от земной/морской поверхности), при этом данные предварительно преобразуются в вероятности различных классов нейросетевой моделью. В работе [Kong et al., 2024] для кластеризации на два кластера (рассеяние от земной поверхности и от ионосферы) использовался алгоритм AE-K-means, представляющий собой кластеризацию методом K-means особенностей, извлеченных из данных с помощью нейронной сети-автоэнкодера. Поэтому выбор метода кластеризации субъективен и зависит от исследователя.

Второй этап (классификация) алгоритма [Berngardt et al., 2022] — обучение классификатора на данных, размеченных на первом этапе. К недостаткам этого алгоритма можно отнести необоснованно большую нейронную сеть, которую сложно интерпретировать, и интуитивно выбранное количество классов, равное 20. Следует заметить, что и в других работах, использующих нейросети, например в [Kunduri et al., 2022; Kong et al., 2024], выбор архитектуры нейронной сети часто не обосновывается.

Для усовершенствования модели в рамках управляемого данными подхода архитектуру и параметры нейронной сети необходимо выбрать оптимальными исходя из особенностей используемого набора данных.

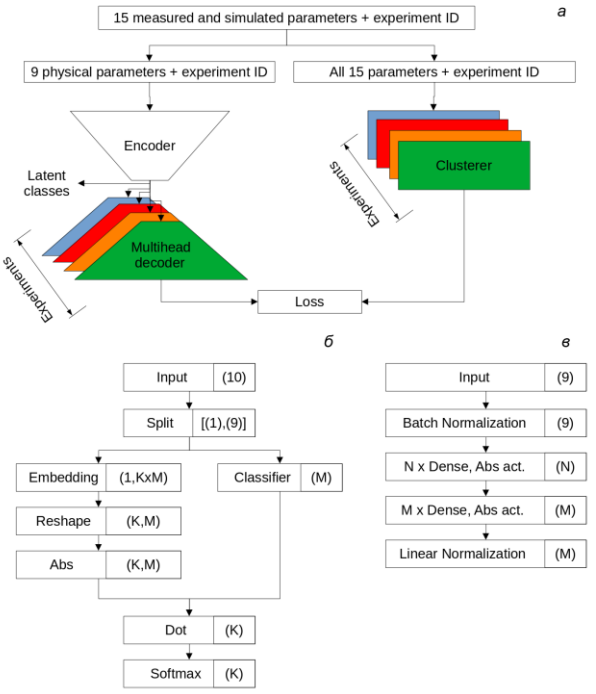

Рис. 1. Используемые в работе нейронные сети и метод их обучения: *а* — схема обучения обернутого классификатора (encoder). Каждый кластеризатор и декодер соответствуют одному эксперименту (фиксированный луч, частотный канал и день). Количество голов декодера и кластеризаторов равно числу экспериментов и составляет ~15 000. Цвета соответствуют различным экспериментам. Детальная архитектура (*б*) оболочки, используемой для тренировки классификатора. Детальная архитектура (*в*) нейронной сети-классификатора (Encoder). *K* — максимальное число кластеров после этапа 1, *M* — число скрытых классов в сигнале, *N* — размерность скрытого слоя классификатора

Необходимость интерпретируемости классов, на которые мы разделяем сигналы, требует, чтобы число скрытых классов было минимально возможным и тем не менее достаточным для уверенного описания наших данных и предсказания результатов. С математической точки зрения задача сводится к поиску нейронной сети минимальной ширины, обеспечивающей максимально возможное качество решения.

ЭТАП 1. КЛАСТЕРИЗАЦИЯ ДАННЫХ

Используемые для кластеризации данные

При построении классификатора по аналогии [Berngardt et al., 2022; Бернгардт, 2022] использовались следующие измеряемые и модельные данные. Измеряемые радаром параметры:

1. Время, дальность до рассеивателя.
2. Доплеровская скорость *V*, измеренная спектральная ширина *W* — определяются по сигналу алгоритмом FITACF [Ribeiro et al., 2013], в работе используется спектральная ширина в экспоненциальной модели корреляционной функции (*Wl*, ед. скорости).
3. Угол места — определяется алгоритмом [Berngardt et al., 2021] с калибровкой радара по метеорам.

Параметры, полученные моделированием распространения радиоволны методом волновой оптики в модельной ионосфере, описываемой международными моделями IRI и IGRF.

4. Эффективная высота рассеяния — рассчитывается как результат прямолинейного (безрефракционного) распространения радиоволны.
5. Угол наклона (синус угла места) траектории по отношению к горизонтали в четырех точках по дальности (1/4, 2/4, 3/4, 4/4 измеренной дальности).
6. Угол между направлением распространения радиоволны и магнитным полем Земли (косинус угла).
7. Мода распространения — количество отражений от нижележащего (ионосферного) слоя или от земной поверхности во время распространения до рассеивателя.
8. Высота рассеяния.

В качестве входных данных классификатора в работах [Berngardt et al., 2022; Бернгардт, 2022] использовались только 10 параметров (2, 4–8). Для кластеризации использовались все 15. В данной работе из параметров классификатора исключена эффективная высота рассеяния (пункт 4).

Учет типа радара при обучении модели

Ранее для обучения модели использовались только данные зондирования 7- и 8-импульсными зондирующими последовательностями (наиболее часто использующимися на радарах СЕКИРА и SuperDARN). Предварительный анализ показал, что при использовании полного набора данных радаров ИСЗФ СО РАН результат прогноза чувствителен к типам радаров и режимам их работы. Поэтому для построения единой модели, не зависящей от характеристик радара, при подготовке данных необходимо компенсировать различия их характеристик и режимов. К основным трем отличиям относятся форма зондирующих сигналов (тип используемой зондирующей последовательности влияет на спектральное разрешение [Berngardt et al., 2020]), расстояние между интерференционной и основной решетками (влияет на неопределенность в вычислении угла места приходящего сигнала [Milan et al., 1997]) и тип используемых антенн (диаграмма направленности влияет на зависимость мощности сигналов от азимута и угла места).

Тип используемых антенн наиболее сложен для учета, поэтому в данной работе не рассматривается.

Учет формы зондирующего сигнала в работе компенсируется искажением (аугментацией) данных и приведением всех данных к статистически похожему виду вне зависимости от типа используемого сигнала, а учет расстояния между решетками компенсируется отсевом «плохих» сигналов, приходящих не с главного лепестка диаграммы направленности [Milan et al., 1997].

Рассмотрим искажение (аугментацию) данных радаров. На когерентных радарах SuperDARN чаще всего используются зондирующие сигналы двух основных типов — стандартный 7-импульсный сигнал [Barthes et al., 1998] и 8-импульсный katscan [Ribeiro

et al., 2013], иногда — 13-импульсный tauscan [Greenwald et al., 2008]. На радарах системы СЕКИРА к ним добавляются еще 10-импульсный и 16-импульсный [Berngardt et al., 2020] сигналы. Все они обладают разной длительностью и различным спектральным разрешением. Это особенно влияет на измерение спектральной ширины W принятого сигнала: самый короткий 7-импульсный сигнал дает максимальные ошибки, самый длинный 16-импульсный — минимальные.

Существует три основных подхода к приведению этих данных к единому виду, не зависящему от типа зондирующей последовательности: решение обратной задачи, отсев нестандартных данных и преднамеренное искажение нестандартных данных.

Первый подход — решение обратной задачи — сводится математически к задаче обращения свертки и требует переделки существующего алгоритма обработки данных FITACF, поэтому не рассматривается.

Второй подход — исключение спектральной ширины из рассмотрения, что очевидно неэффективно: во всех существующих алгоритмах разделения сигналов SuperDARN/СЕКИРА спектральная ширина играет важную роль.

Используемым в работе подходом является преднамеренное искажение данных, получаемых с высоким спектральным разрешением, до состояния, в котором их сложно отличить от данных, получаемых с низким спектральным разрешением. В машинном обучении такое преднамеренное искажение называется аугментацией и широко используется [Shorten, Khoshgoftaar, 2019].

В рамках этого подхода все данные приводились к наихудшей точности: спектральная ширина в данных, полученных более длинными последовательностями, увеличивалась таким образом, чтобы получившиеся распределения спектральных ширин были близки к данным 7-импульсной последовательности. Такой подход позволяет задействовать для обучения данные, полученные с различным спектральным разрешением.

Поскольку чаще всего на радарах ИСЗФ используются 16-импульсные и 7-импульсные последовательности, была проведена оценка необходимых дополнительных искажений данных зондирования 16-импульсными сигналами по сравнению с 7-импульсными сигналами. Алгоритм оценки спектральной ширины FITACF достаточно сложен, поэтому необходимые искажения спектральной ширины были определены экспериментально согласно формуле

$$W_{16,\mathrm{augm}} = W_{16} + \delta W \approx W_7. \qquad (1)$$

Здесь W_{16} и W_7 — спектральная ширина, получаемая алгоритмом FITACF по данным измерений 16-импульсной и 7-импульсной последовательностью соответственно; $W_{16,\mathrm{augm}}$ — ее аугментированное значение; δW — искомое искажение.

Пусть δW — случайная величина с неизвестной плотностью вероятности $\mathbb{P}_{\delta W}$. Тогда плотность вероятности аугментированной спектральной ширины $\mathbb{P}_{W_{16,\mathrm{augm}}}$ представляет собой свертку плотности вероятности $\mathbb{P}_{W_{16}}$ спектральной ширины, измеренной 16-импульсной последовательностью, с плотностью вероятности искажений $\mathbb{P}_{\delta W}$ и должна быть примерно равна плотности вероятностей $\mathbb{P}_{W_7}$ спектральной ширины, измеренной 7-импульсной последовательностью:

$$\mathbb{P}_{W_{16,\mathrm{augm}}}(W) = \mathbb{P}_{W_{16}}(W) * \mathbb{P}_{\delta W}(W) \approx \mathbb{P}_{W_7}(W). \qquad (2)$$

Распределение искажений (ядро свертки $\mathbb{P}_{\delta W}$) было найдено из распределений спектральных ширин 16-импульсных последовательностей и 7-импульсных последовательностей анализом сигналов с низким доплеровским смещением, обычно характерным для сигналов, рассеянных от земной поверхности [Blanchard et al., 2009]. Для этого выбирались сигналы с доплеровским смещением не более 30 м/с с дальностей 500–1500 км в экспериментах, где зондирование проводилось с чередованием видов импульсной последовательности. Этот режим был регулярным на радаре EKB с апреля по декабрь 2021 г. Пример распределений показан на рис. 2, *а*, *б*.

Решение задачи поиска распределения $\mathbb{P}_{\delta W}$ проводилось обучением нейронной сети, состоящей из одной свертки шириной [–100, 100] м/с без использования функций активации. Коэффициенты найденного ядра свертки $\mathbb{P}_{\delta W}$ показаны на рис. 2, *г* оранжевым цветом. Поэтому в качестве модели искажений, хорошо аппроксимирующей такое распределение, была выбрана случайная величина

$$\delta W = \tan(\eta)19 - 5\,[\text{м/с}]. \qquad (3)$$

Здесь η — случайная величина, имеющая равномерное распределение в диапазоне [0, atan(6)],

$$\eta \sim \mathbb{U}([0, \operatorname{atan}(6)]). \qquad (4)$$

Параметры этой модели были подобраны вручную для обеспечения удовлетворительного совпадения кривых рис. 2, *г*. Использование δW привело к близкому распределению спектральных ширин измеряемых 7-импульсной последовательностью и аугментированных данных, получаемых 16-импульсной последовательностью (см. рис. 2, *е*).

На рис. 2 приведены распределения спектрального уширения сигналов, полученных во время регулярных измерений 16-импульсной и 7-импульсной последовательностями до ($\mathbb{P}_{W_{16}}$, рис. 2, *в*) и после ($\mathbb{P}_{W_{16,\mathrm{augm}}}$, рис. 2, *е*) компенсации спектральной ширины по методу (1, 3, 4) в сравнении с распределением спектральных ширин, измеренных 7-импульсной последовательностью $\mathbb{P}_{W_7}$.

На рис. 2, *а*, *б*, *д* приведены распределения плотностей вероятности в координатах скорость — спектральное смещение для измерений 7-импульсной, 16-импульсной последовательностями и для 16-импульсной последовательности после ее аугментации.

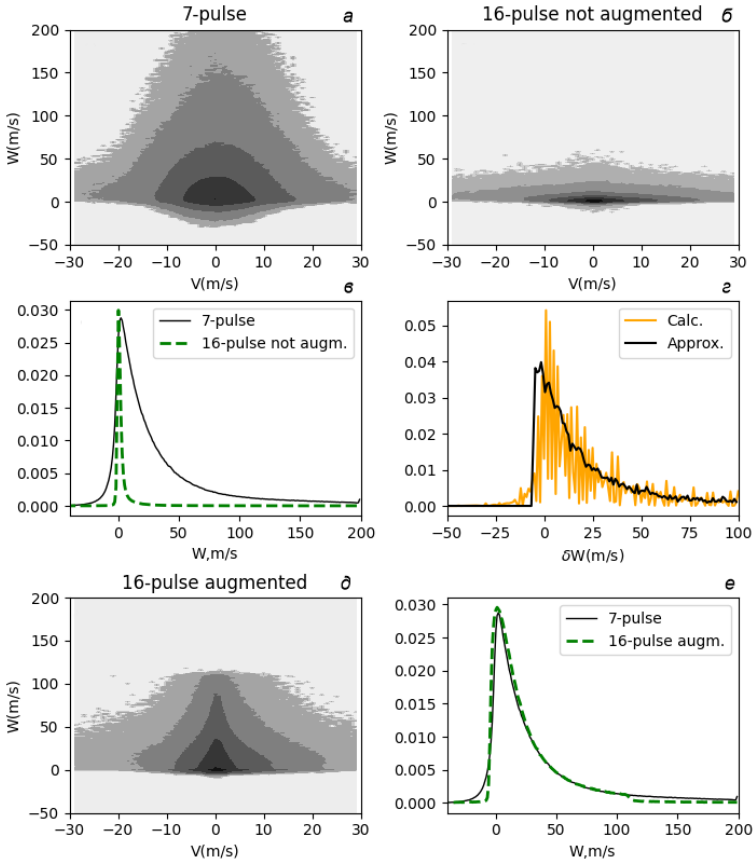

Рис. 2. Аугментация спектрального уширения 16-импульсных последовательностей по данным радара EKB за 2021 г. Распределение пар скорость — спектральное уширение для импульсов двух основных типов — 7-импульсного (*а*) и 16-импульсного (*б*); распределение данных по спектральной ширине (*в*); вычисленное и модельное распределения аугментирующей добавки $\mathbb{P}_{\delta W}$ (*г*); распределение пар скорость — спектральное уширение для 16-импульсной последовательности после аугментации (*д*); распределение данных по спектральной ширине после аугментации (*е*)

Приведенные на рис. 2, *а* распределения *V*, *W* для 7-импульсной последовательности известны и использовались, в частности, для определения условий разделения сигналов рассеяния от земной поверхности [Ponomarenko et al., 2007; Blanchard et al., 2009]. Небольшая доля негативных спектральных ширин — известная ошибка, связанная с особенностями обработки сигналов алгоритмом FITACF.

Сужение (см. рис. 2, *б*) распределения по *W* при измерении 16- импульсным сигналом по сравнению с 7-импульсным сигналом связано с более высоким спектральным разрешением 16-импульсной последовательности.

Из рис. 2 видно, что до аугментации спектральные уширения, полученные 16-импульсной последовательностью значительно уже, чем 7-импульсной, а аугментация позволяет получить распределения для спектральной ширины, близкие к результатам наблюдений 7-импульсной последовательностью.

Особенности кластеризации данных и анализ результатов

Кластеризация — это поиск разбиения данных на кластеры, хорошо интерпретируемые наблюдателем. В рамках подхода [Berngardt et al., 2022; Бернгардт, 2022] требуется не только разбить данные на кластеры, но и выбрать их количество, так чтобы полученная кластеризация была объяснима с физической точки зрения. Существуют различные методы кластеризации[Saxena et al., 2017], каждый соответствует своей модели данных и обычно требует подбора какого-либо гиперпараметра, от значения которого существенным образом зависит и разбиение, и итоговое число кластеров. Чтобы корректно решить задачу выбора метода кластеризации и значений его гиперпараметров, обычно используются два основных подхода: использование внешнего эксперта, оценивающего качество каждой конкретной кластеризации или использование некой метрики качества кластеризации, основанной на каком-либо критерии. В обоих случаях оценка субъективна: и каждый эксперт может интерпретировать данные по-своему, и каждая из используемых метрик качества дает свои оценки. Из метрик наиболее часто используются коэффициент силуэта (Silhouette) [Rousseeuw, 1987] и семейство информационных критериев, например байесовский информационный критерий (BIC) [Schwarz, 1978]. Выбор критерия часто определяет ожидаемую форму и количество кластеров. Например, в работе [Kong et al., 2024], используется сравнение нескольких методов кластеризации по нескольким критериям, включая Silhouette.

Для кластеризации по аналогии с [Berngardt et al., 2022] будем использовать 15-мерные данные,

состоящие из измеренных радарами параметров (скорости, спектральные ширины, углы места, зондирующие частоты, время, азимут/номер луча, частотный канал, и т. д.) и результатов моделирования траектории распространения сигнала (угол относительно магнитного поля Земли, угол относительно горизонта на различных участках траектории, количество скачков распространения сигнала).

Частотные каналы и номера лучей являются категориальными переменными, требующими выбора векторного представления, которое в этой задаче неочевидно. В данной работе, в отличие от [Berngardt et al., 2022; Бернгардт, 2022] используется кластеризация отдельно на каждом луче и на каждом из двух частотных каналов, что исключает необходимость поиска векторного представления, но существенно увеличивает число анализируемых независимых экспериментов.

Для проверки применимости алгоритма GM к этой задаче и определения числа оптимальных кластеров задача кластеризации была решена двумя способами. Первый способ — алгоритм GM с определением числа кластеров по байесовскому информационному критерию (BIC), далее в работе этот метод кластеризации будем называть GMBIC, он ищет кластеры в основном эллиптической формы. Второй способ — кластеризация алгоритмом GMsDB [Berngardt, 2023], который берет кластеры GMBIC и объединяет существенно пересекающиеся эллиптические кластеры в более крупные кластеры сложной формы. Близкий к GMsDB метод кластеризации DBSCAN-GM предлагался, например, в [Smiti et al., 2016], но с несколько отличающимся принципом объединения кластеров. Сравнение числа кластеров, получаемых обоими алгоритмами (GMsDB и GMBIC) позволяет определить, сколько эллиптических кластеров не пересекают друг друга, и в случае, если их число велико, косвенно доказать, что использование GMBIC в этой задаче допустимо.

На рис. 3 показаны распределения числа найденных кластеров в радарных данных двумя методами: алгоритмами GMBIC и GMsDB с уровнем статистической значимости при объединении кластеров $\alpha=0.1$ [Berngardt, 2023]. На панелях *а1*, *а2* показаны зависимости количества кластеров, определенных двумя алгоритмами. Видна пропорциональность в числе кластеров, что говорит о примерно постоянной доле пересекающихся эллиптических кластеров. На панелях *б1*, *б2* показаны распределения числа кластеров GMBIC, объединяемых алгоритмом GMsDB. Видно, что кластеры более сложной формы, которые определяются алгоритмом GMsDB, составлены не более чем из 3–4 эллиптических кластеров. Низкая доля сложных кластеров говорит о том, что в основном кластеры имеют несложную эллиптическую форму. На панелях *в1*, *в2* показаны распределения доли изолированных кластеров GMBIC, у которых нет близких соседей. Видно, что в среднем 80–83 % кластеров, определяемых алгоритмом GMBIC, изолированы. Их высокая доля говорит о допустимости использования GMBIC для первоначальной кластеризации. Это объясняет также приемлемое качество аппроксимации кластеров данных когерентных радаров, достигаемое моделями, основанными на GM [Berngardt et al., 2022; Бернгардт, 2022]. На панелях *г1*, *г2* показаны распределения количества GMBIC кластеров в данных, *д1*, *д2* — распределения количества GMsDB кластеров в данных. Видно, что число кластеров не превышает 52 и в данных радара MAGW кластеров в среднем больше, чем в данных радара EKB. На панелях *е1*, *е2* показаны доли данных, находящихся в изолированных GMBIC кластерах. Видно, что существенная часть данных радаров (от 40 до 80 %) расположена в изолированных эллиптических кластерах. Их высокая доля также говорит о допустимости использования метода GMBIC для кластеризации. На радаре MAGW доля данных в кластерах сложной формы превышает долю таких данных на радаре EKB.

Алгоритм GMsDB имеет тенденцию к объединению соприкасающихся кластеров в один [Berngardt, 2023], негативно влияющую на качество кластеризации в случае соприкасающихся кластеров. Поэтому в дальнейшем в качестве базовой использовалась кластеризация GMBIC, примерно 80 % кластеров которой являются изолированными и совпадают с кластерами GMSDB и только 20 % кластеров в данных являются кластерами сложной (неэллиптической) формы (см. панели *в1*, *в2*). Это означает, что, используя метод GMBIC, мы кластеризуем корректно от 50 до 80 % всех данных на простые эллиптические кластеры, а кластеры сложной формы разделяем на несколько простых.

ЭТАП 2. КЛАССИФИКАЦИЯ ДАННЫХ

Построение оптимального классификатора данных и его обоснование

Основная задача этого этапа — поиск минимальной полносвязной сети-классификатора, повторяющей кластеризацию с высокой точностью. Минимальная сеть важна для облегчения интерпретации получаемых ею результатов с физической точки зрения: число нейронов на выходе сети-классификатора соответствует минимальному числу независимых типов сигналов в наблюдаемых данных. Построим полносвязную сеть, состоящую из малого количества слоев (два слоя) с минимально возможным числом нейронов в каждом слое.

Для построения сети было создано два набора данных: полный и укороченный. Укороченный набор данных использовался для получения хорошего начального приближения для всех коэффициентов нейронной сети и определения ее гиперпараметров — минимального числа нейронов в каждом ее слое. Полный набор данных использовался для окончательного решения задачи и окончательного дообучения сети.

Полный набор данных был создан из ~15000 экспериментов общим объемом ~42 млн. записей: (20 млн. записей радара EKB и 22 млн. записей радара MAGW) и разбит в соотношении 4:1 на обучающую и тестовую части.

Укороченный набор данных был создан из 1000 экспериментов (объемом ~2.8 млн. записей), случайно

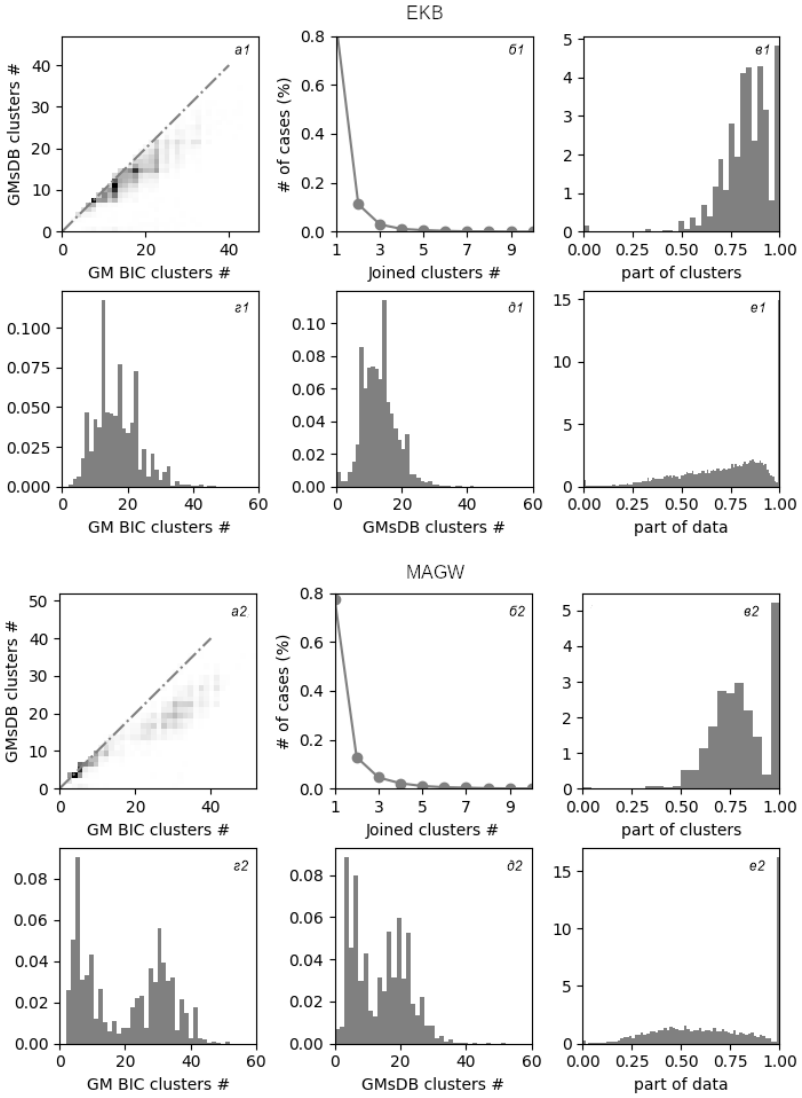

Рис. 3. Распределения количества кластеров определяемых GMBIC и GMsDB алгоритмами на радарах EKB и MAGW: распределение количеств кластеров, определенных двумя алгоритмами (*а1*, *а2*); распределение числа кластеров GMBIC, объединяемых алгоритмом GMsDB (*б1*, *б2*); распределение доли изолированных кластеров GMBIC (*в1*, *в2*); распределение количества GMBIC кластеров в данных (*г1*, *г2*); распределение количества GMsDB кластеров в данных (*д1*, *д2*); доля данных, находящихся в изолированных GMBIC кластерах (*е1*, *е2*)

выбранных из полного набора данных, и разбит в соотношении 4:1 на обучающую и тестовую части. Валидационные части набора данных в обоих случаях отсутствовали, поскольку при обучении использовалась кросс-валидация обучающего набора данных (по трем фолдам) и всегда тренировались три версии модели, что необходимо для последующего анализа.

Обоснование архитектуры сети-классификатора

Архитектура нейронных сетей (обертки и классификатора) показана на рис. 1, *б*, *в*. Здесь K — максимальное число кластеров после этапа 1; M — число скрытых (латентных) классов в данных; N — размерность скрытого слоя классификатора. Сеть является существенным упрощением варианта, предложенного в [Berngardt et al., 2022; Бернгардт, 2022], но обеспечивает лучшее качество прогноза.

Выбор архитектуры новой сети-классификатора (см. рис. 1, *в*) основывается на трех принципах: для аппроксимации непрерывных функций достаточно широкой двухслойной сети [Kolmogorov, 1957; Arnold, 1963]; в качестве функций активации нейронной сети желательно использовать вычисление абсолютных значений для сохранения связи с алгоритмами, хорошо зарекомендовавшими себя ранее [Ponomarenko et al., 2007]; в качестве преобразования выходов сети к вероятностям вместо функции *Softmax* возможно использовать нормировку неотрицательных величин на их сумму.

Обоснуем выбор абсолютной функции активации. Предлагаемая в работе модель является развитием стандартных подходов к разделению сигналов на радарах когерентного рассеяния. Известное и широко используемое на радарах SuperDARN условие разделения сигналов на два типа (рассеяние от земной поверхности и рассеяние от ионосферных неоднород-

ностей) по их спектральным характеристикам сигналов имеет вид [Ponomarenko et al., 2007]

$$A|V|+B|W|+C>0, \quad (5)$$

где A, B, C — некоторые постоянные; V, W — измеренное доплеровское смещение сигнала и его спектральная ширина. Можно предположить, что в случае с другими классами границы можно описать также суперпозицией функций модуля и поэтому выгодно использование функции абсолютного значения в качестве функции активации.

Обоснуем использование слоя линейной нормализации. Традиционно на выходе большинства классификаторов используется функция *Softmax*, позволяющая нормировать выходы нейронной сети таким образом, чтобы значения были неотрицательными, а их сумма была равна 1. Однако часто необходимо выполнять дополнительную калибровку получаемых значений [Guo et al., 2017] или видоизменять функцию *Softmax* [Sutton et al., 2018]. Поэтому в выборе функции активации на выходе классификатора допустим произвол. В данной работе мы будем использовать в качестве активации на выходе сети-классификатора слой линейной нормализации

$$LinearNormalization\left(\vec{x}\right)_i=\frac{x_i}{\sum_j x_j}. \quad (6)$$

При выполнении условия $x_i \geq 0 \; \forall i$ (оно выполнено автоматически из-за использования абсолютной активации на предыдущем слое, см. рис. 1, *в*) выходные значения слоя, как и при использовании *Softmax*, удовлетворяют аксиоматике Колмогорова в теории вероятностей [Kolmogoroff, 1933]: они неотрицательны, их сумма равна 1, а вероятность нескольких взаимоисключающих событий равна сумме их вероятностей. Поэтому выходы такого слоя могут интерпретироваться, как вероятности соответствующих классов, и это не требует изменения стандартных функций потерь при тренировке сети (кросс-энтропии).

В отличие от сети, разработанной ранее [Berngardt et al., 2022; Бернгардт, 2022], использование физически адекватных (абсолютных) функций активации позволило решить задачу:

- меньшим числом входных параметров: в новой модели классификатора не требуется знание эффективной высоты рассеяния;
- без использования методов повышения размерности данных: классификатор не использует предварительного увеличения размерности данных с помощью Polynomial Features;
- существенно уменьшив глубину сети с шести до двух полносвязных слоев.

На входе классификатора (см. рис. 1, *в*) стоит слой батч-нормализации [Ioffe, Szegedy, 2015], по смыслу являющийся адаптивным линейным масштабированием входных данных и использующийся для ускорения поиска оптимальных коэффициентов сети. При необходимости его коэффициенты могут быть внесены в коэффициенты первого слоя сети. Как показал последующий анализ, построенная нейронная сеть обеспечивает намного лучшее качество прогноза кластеров по сравнению с моделями [Berngardt et al., 2022; Бернгардт, 2022].

Определение числа классов сигналов для классификации

В рамках предложенной модели задача классификации сигналов сводится к анализу данных на выходе энкодера (см. рис. 1, *а*, *в*) как к вероятности того, что данные принадлежат к одному из нескольких классов. Важным вопросом при построении интерпретируемой сети является выбор оптимального количества нейронов этого слоя (и остальных слоев нейронной сети) при фиксированных функциях активации.

В качестве исходной широкой сети для классификатора была выбрана сеть шириной N=300, M=140 на первом и втором слое соответственно. Это можно обосновать следующим образом: количество обнаруженных кластеров в данных для радаров не превышает 52 (см. рис. 3, *г1*, *г2*), таково и максимально ожидаемое число скрытых классов M в данных и минимальное число нейронов в выходном слое классификатора. Согласно [Berngardt, 2024]), начальное количество нейронов желательно выбирать как минимум в два раза больше ожидаемого минимального числа нейронов. Поэтому число нейронов в последнем слое было выбрано примерно в три раза превышающем максимальное число кластеров, а число нейронов в первом слое — примерно в 6 раз. Как оказалось далее, такой архитектуры достаточно для поиска минимального числа нейронов и такая нейронная сеть может быть натренирована за приемлемое время на обычном персональном компьютере. Для ускорения поиска минимального числа нейронов сеть была обучена на уменьшенном наборе данных (1000 экспериментов), описанном ранее. Кросс-валидация в соответствии с алгоритмом [Berngardt, 2024] проводилась по трем фолдам, в результате были обучены модели трех вариантов.

В соответствии с алгоритмом [Berngardt, 2024] для поиска минимального числа нейронов при оценке качества работы сети требуется использовать метрики качества Q, удовлетворяющие соотношению

$$\begin{aligned} &Q\left(X_1 \vee X_2\right)= \\ &=\frac{Dim\left(X_1\right)Q\left(X_1\right)+Dim\left(X_2\right)Q\left(X_2\right)}{Dim\left(X_1\right)+Dim\left(X_2\right)}; \\ &X_1 \wedge X_2=\varnothing, \end{aligned} \quad (7)$$

где X_1, X_2 — непересекающиеся наборы данных, а $Dim(X)$ — число элементов (сэмплов) в X. Поэтому в качестве базовой метрики при поиске минимального числа нейронов была выбрана метрика точности (Accuracy).

Оценка числа независимых классов проводилась двумя способами.

Первый способ дает верхнюю оценку минимально достаточного числа классов в данных. На наборе данных (отфильтрованном по углам места и аугментированном) проводилась кластеризация каждого эксперимента методом GMBIC, обучалась нейронная сеть (300×140 нейронов), определялось минимальное

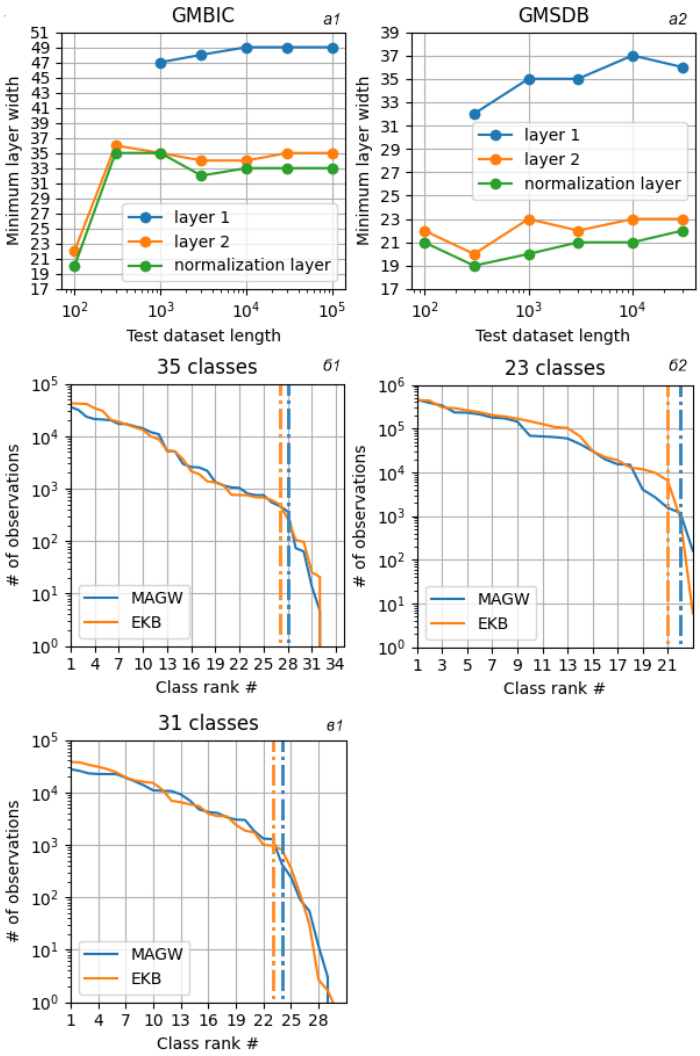

Рис. 4. Расчетное минимальное число нейронов в сети и распределение классов по числу элементов: слева — использование кластеризатора GMBIC; справа — использование кластеризатора GMsDB; *а1*, *а2* — зависимость минимального числа нейронов в слоях классификатора (первом скрытом, втором скрытом и выходном/нормализационном) от объема используемого для поиска набора данных; *б1*, *б2* — зависимость числа элементов в классе от ранга этого класса при использовании оптимальной сети (35 классов и 23 класса соответственно); *в1* — зависимость числа элементов в классе от ранга этого класса при использовании неоптимальной сети (31 класс). Вертикальные линии соответствующих цветов на *б1*, *б2*, *в*1 показывают границу между часто и редко встречающимися классами на соответствующих радарах

число нейронов в слоях (49 и 35 соответственно, рис. 4, *а1*), обучалась минимальная сеть (49×35 нейронов). При этом пример полагался принадлежащим к заданному классу, если все три обученных классификатора предсказывали для него одинаковый класс (ансамблевый способ классификации голосованием). Число элементов в ранжированном ряду классов показано на рис. 4, *б1*. По наблюдаемой точке перегиба (резкое уменьшение числа наблюдений точек в классе) можно оценить число часто наблюдаемых классов. Оно составляет 27 для радара EKB и 28 для радара MAGW (вертикальные штрих-пунктирные линии на рис. 4, *б*1). Необходимо отметить, что уменьшение числа классов меньше 35 (и обучение классификатора заново с новым количеством классов) не делает положение перегиба стабильным: например, при выборе количества скрытых классов, равного 31, оно становится уже равным 23 и 24 на радарах EKB и MAGW соответственно (см. рис. 4, *в1*, вертикальные штрих-пунктирные линии). Поэтому уменьшение числа классов ниже 35, видимо, неоправданно. Использование этого способа соответствует тому, что все найденные нами кластеры имеют форму, близкую к эллиптической, но могут существенно пересекаться. Очевидно, что, если реальные кластеры имеют более сложную форму, этим способом мы можем переоценивать число независимых классов. Поэтому такая оценка соответствует оценке числа классов сверху.

Второй способ дает нижнюю оценку минимально достаточного числа классов. На оригинальном наборе данных (отфильтрованном по углам места и аугментированном) проводилась кластеризация методом GMsDB, обучалась широкая нейронная сеть, как и в первом способе (300×140), определялось минимальное число нейронов в слое (36 и 23 соответственно, рис. 4, *а2*). После этого на наборе данных, кластеризованном методом GMBIC, обучалась итого-

вая сеть с минимальным числом нейронов (36×23). Это соответствует выделению минимально возможного числа непересекающихся классов. Следует отметить, что число классов, уверенно определяемых как положение перегиба на графике рис. 4, *б2*, лишь на 1–2 меньше, чем это значение: 21 класс для радара EKB и 22 для радара MAGW, что говорит о стабильности такого разделения. Использование этого метода соответствует тому, что найденные кластеры могут иметь сложную форму и не пересекаются. Очевидно, что, если реально существующие в данных кластеры пересекаются между собой (типы сигналов слабо различимы), этим способом мы недооцениваем число независимых классов. Поэтому такая оценка соответствует оценке числа классов снизу. Следует отметить, что такая оценка числа классов снизу близка к эмпирически используемому в работах [Berngardt et al., 2022; Бернгардт, 2022] количеству кластеров и скрытых классов (20 классов).

Таким образом, ожидаемое число классов в данных составляет от 23 до 35, причем ошибка при классификации по верхней оценке (35 классов) может приводить к разделению сложных классов на несколько частей, а ошибка при классификации по нижней оценке (23 класса) может приводить к объединению нескольких разнотипных, но близких классов в один. Очевидно, что первый вариант для исследователя предпочтительнее, поэтому и будет нами использоваться в дальнейшем.

Для повышения точности и общности модель была дообучена на всем наборе обучающих данных (более 15 тыс. экспериментов, или ~25 млн. различных сэмплов). Модель обучается в трех вариантах, с использованием кросс-валидации по трем фолдам, разбиение на фолды случайное.

Итоговый алгоритм поиска оптимального классификатора

Итоговый алгоритм поиска оптимального классификатора состоит из следующих действий.

1. Выделение из данных части, которая соответствует углам места ниже порогового [Milan et al., 1997] (28° для радара EKB и 38° для радара MAGW), соответствующим сигналам, приходящим с главного лепестка диаграммы направленности.

2. Расчет по данным радара траекторных параметров распространения радиоволны, формы траектории, угла с магнитным полем и высоты рассеяния и расширение с их помощью набора измеренных радарами параметров.

3. Аугментация вычисленной спектральной ширины принятого сигнала для экспериментов, использующих длинные импульсные последовательности (16-импульсные) в соответствии с (3), (4).

4. Кластеризация каждого эксперимента (эксперименты отличаются по датам, азимутам и частотным каналам) методом GMBIC в 15-мерном пространстве параметров, описанных ранее.

5. Выделение из экспериментов небольшого набора данных (1000 экспериментов), с которыми проводятся этапы 6–7.

6. Обучение достаточно широкой нейронной сети (архитектура рис. 1, *в*) по 9 физическим параметрам с использованием сети с 300 скрытыми нейронами в первом скрытом слое и 140 нейронами (латентными классами) во втором.

7. Определение минимального количества нейронов в слоях получившейся сети по алгоритму [Berngardt, 2024] с использованием Accuracy как метрики.

8. На всех доступных экспериментах обучение нейронной сети архитектуры (см. рис. 1, *а–в*) с найденным оптимальным количеством нейронов в сети-классификаторе.

На этапах 1–3 мы подготавливаем данные, на этапах 4–7 по небольшой части набора данных определяем количество скрытых классов в данных и оптимальное число нейронов в сети, на этапе 8 обучаем итоговый оптимальный классификатор по всему доступному набору данных.

Предложенный алгоритм автоматически определяет количество классов в данных, является полностью управляемым данными и не требует наличия эксперта на любом из этапов. Алгоритм является самосогласованным и самообучающимся: все параметры алгоритма определяются им самим автоматически, за исключением списка параметров, используемых для кластеризации и классификации, и общей архитектуры сети, основания выбора которой были изложены ранее.

Итоговая нейронная сеть имеет 49 нейронов в первом слое и 35 нейронов во втором, что означает наличие в данных 35 различных классов. Модель достигает качества повторения кластеризации 0.92 по метрике AUC-PR и существенно превышает качество 0.68 предыдущих сетей [Бернгардт, 2022; Berngardt et al., 2022].

Полученная нейронная сеть-классификатор (см. рис. 1, *в*) позволяет обеспечить минимальное число параметров нейронной сети при высоком качестве ее работы, что в последующем упрощает ее интерпретацию. С другой стороны, это число нейронов можно интерпретировать как оптимальное число различимых с радиофизической точки зрения классов в данных радаров EKB и MAGW.

Итоговая модель (см. рис. 1, *в*) для определения по параметрам сигнала его класса имеет аналитический вид

$$y_k = \left| \sum_{j=1}^{49} C_{kj} \left| \sum_{i=1}^{9} A_{ij} x_i + B_j \right| + D_k \right|, \tag{8}$$

$$k_{\text{detected}} = argmax\left(y_k\right), \tag{9}$$

где A_{ij}, B_j, C_{kj}, D_k — коэффициенты, которые ищутся в результате обучения сети; x_i — входные параметры; k_{detected} — номер скрытого класса, к которому будет принадлежать измеренный сигнал с параметрами x_i. Количество параметров модели может быть легко вычислено из формулы выше и равно 2240. Модель может быть легко реализована для быстрых расчетов и без использования специализированных библиотек нейронных сетей. Структурно полученная формула оптимальной классификации сигналов близка к результатам теоремы Колмогорова—Арнольда [Kolmogorov, 1957; Arnold, 1963]. Также видна ее структурная связь со стандартным алгоритмом раз-

деления сигналов, рассеянных от ионосферы и земной поверхности [Ponomarenko et al., 2007], приведенным в (5) и традиционно используемым на радарах SuperDARN.

ОБСУЖДЕНИЕ

Интерпретация получаемых классов

При обучении и исследовании модели использовалась кросс-валидация, что дает возможность использовать три варианта сети для построения ансамблевой модели. Как показал анализ данных 2021 г., классы, определяемые тремя версиями модели, совпадают с высокой степенью качества: скорректированный индекс Рэнда [Hubert, Arabie, 1985] попарно между результатами трех моделей лежит в пределах 0.936–0.967 для радара EKB, и 0.897–0.937 для радара MAGW, что говорит о высокой близости получаемых этими моделями классификаций и позволяет использовать для интерпретации как любую из моделей отдельно, так и три модели совместно (ансамбль).

При совместном (ансамблевом) использовании трех моделей удобно использовать механизм голосования и принимать решение о классе сигнала, когда прогнозы всех моделей совпадают. Если прогнозы сетей не совпадают, результат выделяется в отдельный класс (данные, которые нельзя однозначно интерпретировать).

Статистика параметров различных классов (95%-й доверительный интервал), определенных таким ансамблевым методом в 2021 г., показана на рис. 5 отдельно для радаров EKB и MAGW. Классы разделены на три группы, выделенные цветом: рассеяние от земной поверхности, рассеяние на ионосферных неоднородностях и сигналы которые сложно интерпретировать.

В последний класс выделялись сигналы с неправдоподобно высокими скоростями или спектральными ширинами (>1000 м/с). Сигналы с низкой средней высотой рассеяния (<100 км) интерпретировались как рассеяние от земной поверхности, оставшаяся часть классов — как рассеяние от ионосферы разных типов.

На рис. 5 показана статистика высот рассеяния (Hiri), количество скачков распространения (Mode), радиолокационная дальность (Range), доплеровская скорость V_d, спектральная ширина *Wl*, косинус угла луча радиоволны с магнитным полем Земли cos(*k*, *B*), угол места луча с горизонтом в точке рассеяния sin(*k*, *xy*) и количество случаев наблюдения данного класса (# of cases).

Необходимость использования большого числа классов для сигналов, рассеянных от ионосферы, при автоматической классификации данных уже ранее предлагалась и обосновывалась [Burrell et al., 2015]. Множественные типы сигналов, рассеянных от земной поверхности, уже предлагались и обосновывались ранее, например в [Kunduri et al., 2022]. Вследствие сложности и динамичности процессов, происходящих в ионосфере ожидаемым является превышение числа типов сигналов, рассеянных от ионосферы, над числом типов сигналов, рассеянных от земной поверхности.

Три класса из обнаруженных имеют пренебрежимо малое количество данных (1, 22, 27 класс).

Анализ поведения основных признаков сигнала, показанных на рис. 5, позволяет предварительно интерпретировать классы следующим образом.

Рассеяние ионосферных типов

К нему относятся 13 классов: 0, 2, 3, 5–7, 10, 11, 19–21, 27, 32. Общая доля таких сигналов (из главного лепестка диаграммы направленности) составляет 56 % на радаре EKB и 48 % на радаре MAGW. Их можно интерпретировать следующим образом (см. рис. 5).

1. 0 класс — ракурсное рассеяние в E/F слое на скачках 1.5-м или 2.5-м. Высоты 100–200 км, 2-й скачок, расстояния 1500–3000 км, высокие в основном положительные скорости до 800 м/с, высокие спектральные ширины до 600 м/с, близость к ортогональности к магнитному полю $\cos\left(\vec{k}, \vec{B}\right) \in [-0.25..0]$, рассеяние на восходящей ветви траектории $\sin\left(\vec{k}, \overrightarrow{xy}\right)[R] > 0$.

2. 2 класс — ракурсное рассеяние в F-слое на скачке 0.5-м. Высоты 300–450 км, 1-й скачок, расстояния 800–2500 км, высокие скорости до 250 м/с, высокие спектральные ширины до 250 м/с, близость к ортогональности к магнитному полю, рассеяние на восходящей ветви траектории.

3. 3 класс — неракурсное рассеяние в E/F-слое на скачке 1.0-м или магнитоориентированное квази-двухпозиционное рассеяние [Kravtsov, Namazov, 1980; Berngardt et al., 2016], когда траектории падающей и рассеянной волны существенно отличаются. Высоты 80–200 км, 1-й скачок, расстояния 1000–2500 км, высокие скорости до 300 м/с, средние спектральные ширины до 200 м/с, отсутствие ортогональности к магнитному полю, рассеяние на нисходящей части траектории.

4. 5 класс — ракурсное рассеяние в E/F-слое на скачке 1.5 м. Высоты 50–200 км, 2-й скачок, расстояния 1500–3000 км, высокие в основном отрицательные скорости до 500 м/с, высокие спектральные ширины до 800 м/с, выраженная ортогональность к магнитному полю, рассеяние в основном на горизонтальной или восходящей части траектории.

5. 6 класс — ракурсное рассеяние в E-слое на скачке 0.5-м. Высоты 100–200 км, 1-й скачок, расстояния 350–700 км, скорости низкие, спектральные ширины до 200 м/с, выраженная ортогональность к магнитному полю, рассеяние в основном на горизонтальной или восходящей части траектории.

6. 7 класс — предположительно рассеяние на луче Педерсена [Ponomarenko et al., 2011] или магнитоориентированное квазидвухпозиционное рассеяние [Kravtsov, Namazov., 1980; Berngardt et al., 2016]. Высоты 200–250 км, 1-й скачок, расстояния 1000–1500 км, низкие скорости до 100 м/с, положительные на EKB, отрицательные на MAGW, средние спектральные ширины до 250 м/с, отсутствие ортогональ-

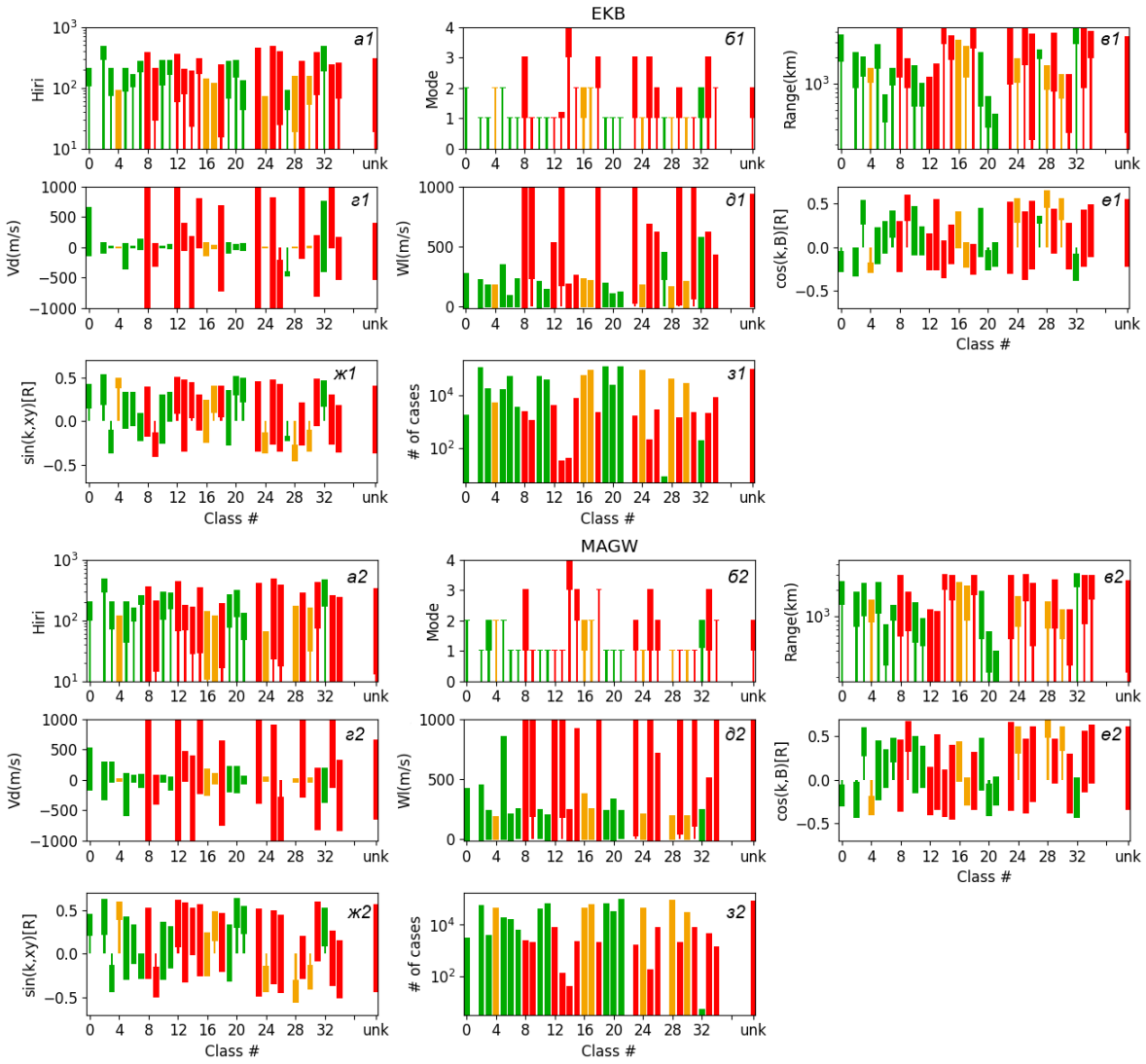

Рис. 5. Параметры различных классов на радарах EKB и MAGW согласно статистике 2021 г. (95%-й доверительный интервал) с использованием ансамблевого метода оценки. Красный цвет — классы предположительно шумовые; зеленый — ионосферное рассеяние; оранжевый — рассеяние от земной поверхности. Последний столбец (unk) — сигналы, интерпретируемые различным образом разными вариантами сети; *а1*, *а2* — высота рассеяния; *б1*, *б2* — число отражений от нижележащего слоя или земной поверхности; *в1*, *в2* — радиолокационная дальность; *г1*, *г2* — доплеровская скорость; *д1*, *д2* — спектральная ширина; *е1*, *е2* — косинус угла с магнитным полем в точке рассеяния; *ж1*, *ж2* — синус угла места в точке рассеяния; *з1*, *з2* — количество наблюдений сигналов

ности к магнитному полю, рассеяние в основном на горизонтальной или нисходящей части траектории.

7. 10 класс — рассеяние на луче Педерсена [Ponomarenko et al., 2011], или в E/F-слое скачка 0.5. Высоты 100–300 км, 1-й скачок, расстояния 500–1500 км, скорости низкие, средние спектральные ширины до 250 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние вблизи горизонтальной части траектории.

8. 11 класс — ионосферное рассеяние в E/F-слое скачка 0.5. Высоты 170–300 км, 1-й скачок, расстояния 500–1000 км, скорости низкие, средние спектральные ширины до 200 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние на горизонтальной или восходящей части траектории.

9. 19 класс — рассеяние в E/F-слое на 0.5 скачка. Высоты 70–300 км, скачок 1-й, расстояния 600–2000 км, средние скорости до 200 м/с, средние спектральные ширины до 250 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние в основном на горизонтальной части траектории.

10. 20 класс — возможный аналог near-range echo (рассеяние на высотах E-слоя на малых дальностях от радара, менее 300 км [Ponomarenko et al., 2016]), но для высот F-слоя (далее F-layer near-range echo). Высоты 150–300 км, 1-й скачок, расстояния 250–700 км, средние скорости до 200 м/с, высокие спектральные ширины до 300 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние на восходящей части траектории.

11. 21 класс — метеорное эхо [Chisham, Freeman, 2013; Бернгардт, 2022] и near-range echo [Ponomarenko et al., 2016]. Высоты 60–100 км, 1-й скачок, расстояния 220–400 км, низкие скорости до 100 м/с, средние спектральные ширины до 200 м/с, слабовы-

раженная ортогональность к магнитному полю, рассеяние на восходящей части траектории.

12. 27 класс — рассеяние в E-слое 1-го скачка, высоты 30–80 км, 1-й скачок, расстояния 2000 км, высокая скорость (~−400 м/с), высокие спектральные ширины до 500 м/с, ортогональность к магнитному полю не выражена, рассеяние на нисходящей части траектории.

13. 32 класс — возможно F-рассеяние скачков 1.5–2.5. Высоты 200–450 км, скачки 1–3-й, расстояния 2000–4500 км, высокие скорости до 800 м/с, высокие спектральные ширины высокие до 300 м/с, ортогональность к магнитному полю не выражена, рассеяние на восходящей части траектории.

Рассеяние от земной поверхности

К нему относятся 6 классов: 4, 16, 17, 24, 28, 30. Общая доля таких сигналов (из главного лепестка диаграммы направленности) составляет 31 % на радаре EKB и 37 % на радаре MAGW. Их можно интерпретировать следующим образом.

1. 4 класс — рассеяние от земной поверхности 1-го скачка. Высоты ниже 100 км, 2-й скачок, расстояния 900–1500 км, низкие скорости, средние спектральные ширины до 200 м/с, отсутствие ортогональности к магнитному полю, рассеяние на восходящей ветви траектории.

2. 16 класс — рассеяние с высокими спектральными ширинами и скоростями, возможно, от земной поверхности 1-го скачка при сильной рефракции на быстроживущих ионосферных неоднородностях. Высоты 0–100 км, скачки 1–2, расстояния 180–3000 км, средние скорости до 200 м/с, высокие спектральные ширины до 400 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние на горизонтальной части траектории.

3. 17 класс — рассеяние на земной поверхности 2-го скачка или E-рассеяние на скачке 1.5. Высоты 0–100 км, 2-й скачок, расстояния 1000–2500 км, скорости низкие, средние спектральные ширины до 200 м/с, слабовыраженная ортогональность к магнитному полю, рассеяние на восходящей части траектории.

4. 24 класс — рассеяние от земной поверхности 1-го скачка. Высоты 0–70 км, 1-й скачок, расстояния 700–2000 км, скорости низкие, средние спектральные ширины до 200 м/с, ортогональность к магнитному полю не выражена, рассеяние на нисходящей части траектории.

5. 28 класс — рассеяние от земной поверхности 1-го скачка, высоты 20–150 км, 1-й скачок, дальность 800–1500 км, скорости низкие, средние спектральные ширины до 200 м/с, отсутствие ортогональности к магнитному полю, рассеяние на нисходящей части траектории.

6. 30 класс — рассеяние на земной поверхности 1-го скачка, 1-й скачок, высоты 40–150 км, дальность 600–1200 км, скорости низкие, средние спектральные ширины до 200 м/с, отсутствие ортогональности к магнитному полю, рассеяние на нисходящей части траектории.

Доля сигналов в остальных (неинтерпретируемых) классах мала и составляет 4 % на радаре EKB и 5 % на радаре MAGW. Доля сигналов, определяемых по-разному различными моделями составляет ~10 % на каждом из радаров. Таким образом, предложенный метод позволяет автоматически классифицировать ~85 % всех данных принимаемых в главном лепестке диаграммы направленности.

Суточно-дальностный ход различных классов

На рис. 6, 7 показан суточный ход сигналов различных классов по данным радаров EKB и MAGW ИСЗФ СО РАН за 2021 г. на тестовом наборе данных, не участвовавшем в обучении. Польза такого представления данных состоит в том, что ни дальность, ни время непосредственно не участвуют в классификации данных и появление в этих координатах сгруппированных областей точек служит субъективным подтверждением хорошего качества классификации. Полученные графики позволяют в части случаев подтвердить интерпретацию этих классов, предложенную выше.

Дополнительным подтверждением корректности классификации является распределение по высотам сигналов нескольких классов (рис. 8): метеорного эхо/near-range echo (рассеяния на близких дальностях от E-слоя), F-layer near-range echo (рассеяния на близких дальностях от F-слоя), других ионосферных сигналов и сигналов, рассеянных земной поверхностью. Видно, что распределение метеоров по высотам, определенное алгоритмом, хорошо соответствует ожидаемому с максимумом в ~80–100 км [Chisham, Freeman, 2013], распределение F-layer near-range echo соответствует высотам ~180 км, сигналы, рассеянные от земной поверхности, сосредоточены на высотах 0–100 км, а ионосферное рассеяние остальных типов имеет максимум распределения в области высот 180–200 км. Распределения сигналов различных типов по высотам на радарах EKB и MAGW близки.

Степень важности различных входных параметров модели

Одним из актуальных вопросов при идентификации типов рассеянных сигналов является выбор необходимых параметров [Burrell et al., 2015; Ponomarenko, McWilliams, 2023 При использовании управляемого данными подхода мы можем сформулировать эту задачу с точки зрения важности признаков: какие параметры наиболее сильно влияют на качество определения каждого конкретного класса в рамках построенной нами модели. Похожий подход использовался ранее в работе [Kong et al., 2024]. В машинном обучении существует большое количество разных методов такой оценки [Huang et al., 2020]. Одним из универсальных методов, позволяющих это сделать, является перестановочный метод (permutation feature importance) [Breiman, 2001], при котором о важности входного параметра для прогноза судят по изменению качества прогноза при случайной перестановке значений этого параметра в наборе данных.

Поскольку нам желательно не только упорядочить входные параметры в порядке важности, но и найти оптимальную для классификации комбинацию

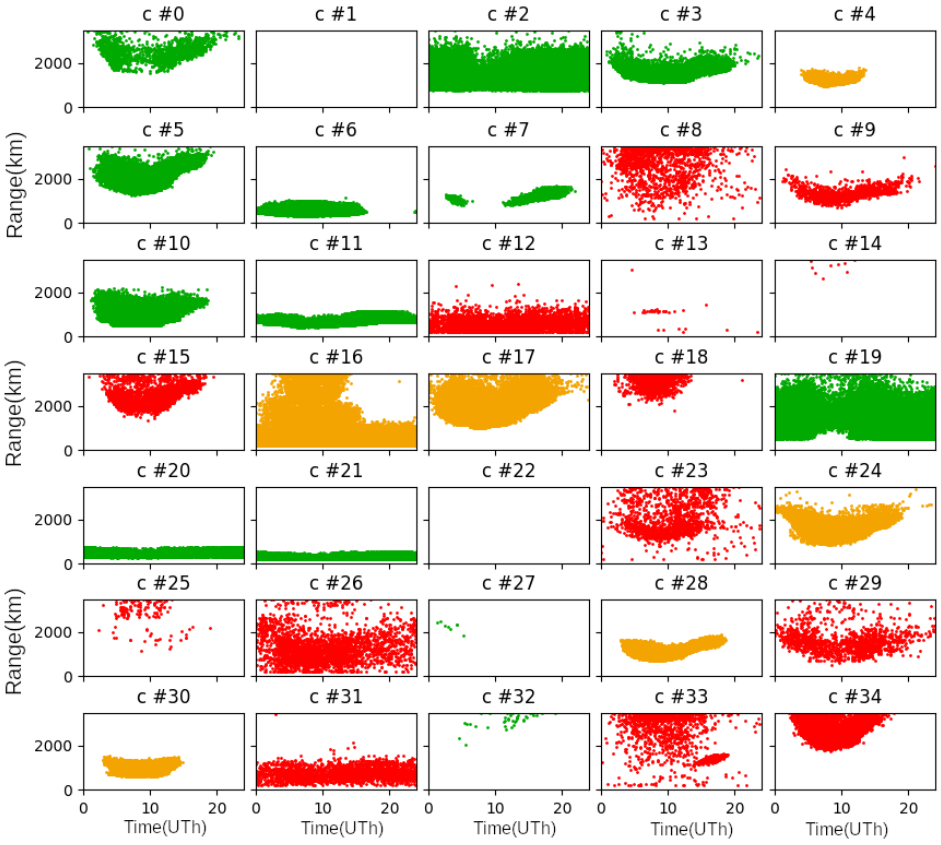

Рис. 6. Суточно-дальностные распределения сигналов на радаре EKB: результат распределения данных по классам на тестовом наборе данных голосованием по ансамблю из трех сетей. Неуверенно определяемые данные исключены. Цвета соответствуют различным типам сигнала аналогично рис. 5

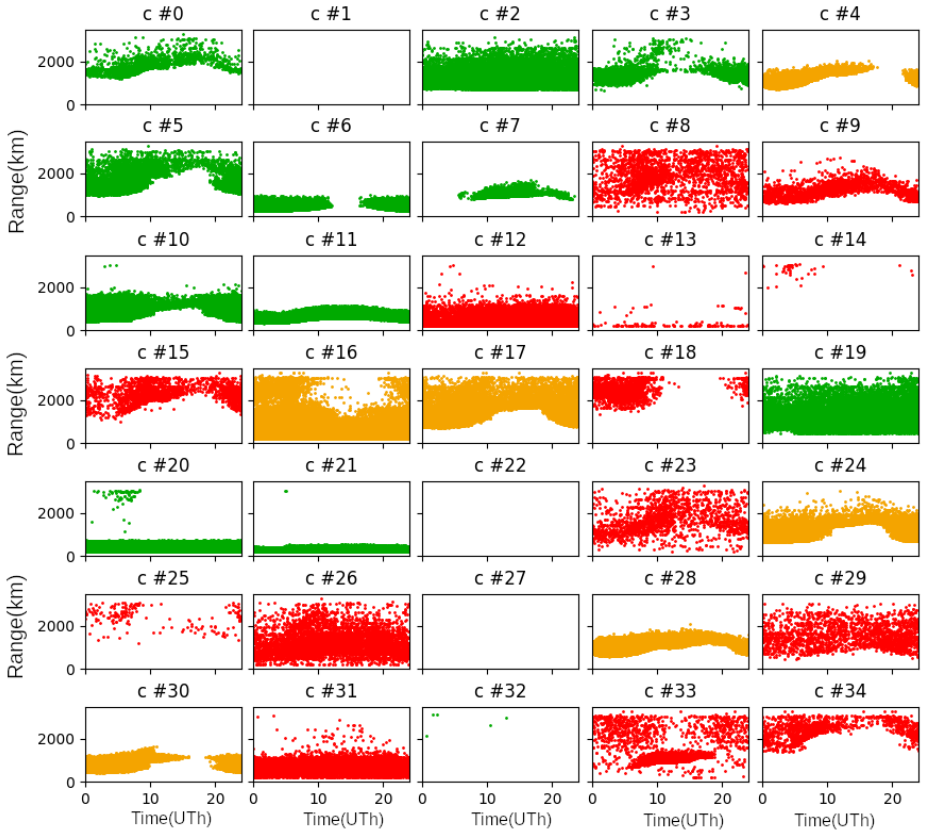

Рис. 7. Суточно-дальностные распределения сигналов на радаре MAGW: результат распределения данных по классам на тестовом наборе данных голосованием по ансамблю из трех сетей. Неуверенно определяемые данные исключены. Цвета соответствуют различным типам сигнала аналогично рис. 5

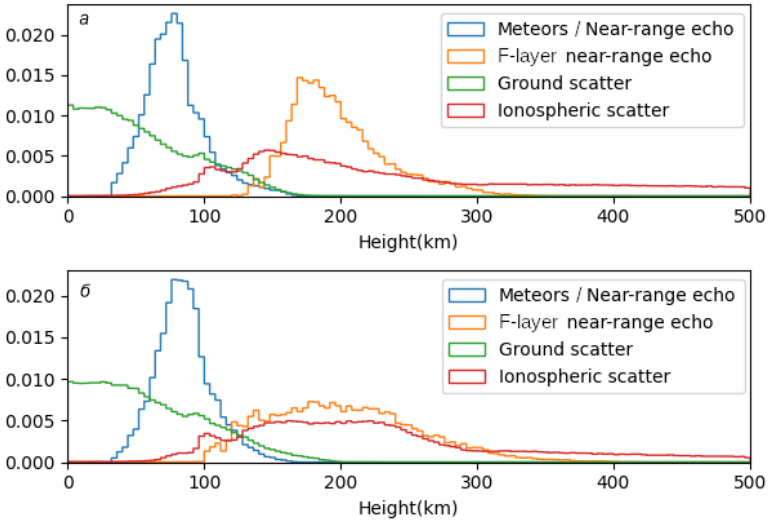

Рис. 8. Распределение рассеяния различных типов по высотам, полученным в результате трассировки лучей с использованием радарных данных и модели IRI на радарах EKB (*а*) и MAGW (*б*) за 2021 г. Приведены распределения метеоров/near-range echo (класс 21), F-layer near-range echo (рассеяния на близких дальностях от F-слоя) (класс 20), рассеяние от ионосферы остальных типов (классы 0, 2, 3, 5–7, 10, 11, 19, 27, 32), и рассеяние от земной поверхности (классы 4, 16, 17, 24, 28, 30)

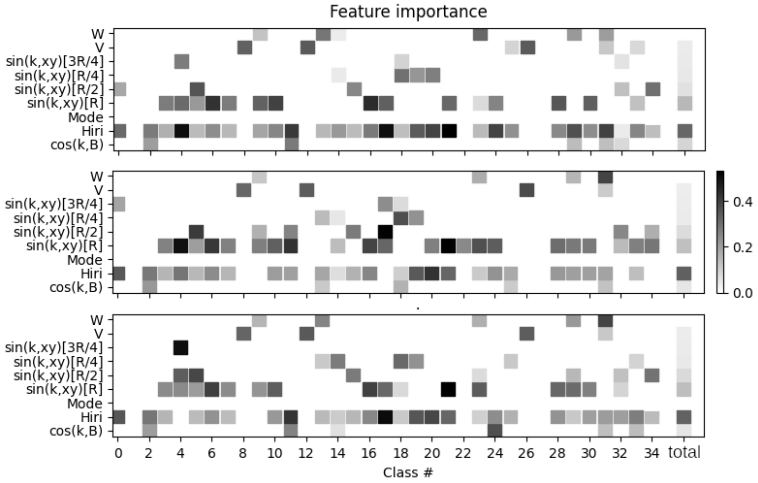

Рис. 9. Степень важности различных параметров для определения классов сети трех различных вариантов, полученных в результате кросс-валидационной тренировки, а также итоговая степень важности каждого параметра при оценке результата классификации (столбец “total”)

таких параметров, желательно использовать жадные модификации алгоритма (используемый в работе вариант дан в Приложении 1).

На рис. 9 показана степень важности различных входных параметров на определение различных классов (величина ΔQ_{lopt}), более высокое значение соответствует более важной компоненте. Приведена также степень важности итоговой классификации (столбец “total”). Результаты даны для каждого варианта сети, полученного в результате кросс-валидации. Ячейки, в которых значение отсутствует, соответствуют малозначимым параметрам.

Видно, что чаще всего важными для классификации являются высота, на которой рассеивается сигнал, наклон траектории распространения радиоволны в точке рассеяния, а также примерно в равной мере угол с магнитным полем Земли в точке рассеяния и угол места в середине траектории распространения сигнала. Наименее важными параметрами являются мода распространения сигнала и спектральная ширина принятого сигнала. Это совпадает с качественными ожиданиями: знание высоты рассеяния и траектории распространения сигнала действительно позволяет легко отличить сигналы различных типов. Для рассеяния от земной поверхности высота рассеяния должна быть ~0, для метеорного эха — ~90 км, для ионосферного рассеяния — от 100 до 400 км. Наклон траектории и угол с магнитным полем позволят отличать обычное рассеяние от ракурсного, характерного для плазменных неустойчивостей E- и F-слоев ионосферы.

Таким образом, наиболее важными при классификации рассеянных сигналов являются форма траектории распространения радиосигнала и высота рассеяния. Эти параметры не могут быть измерены непосредственно радаром и требуют моделирования процесса распространения радиоволны. Для их определения необходимо знать частоту зондирования, трехмерную диаграмму направленности антенной решетки, измеренный угол места и азимут приходящей радио-

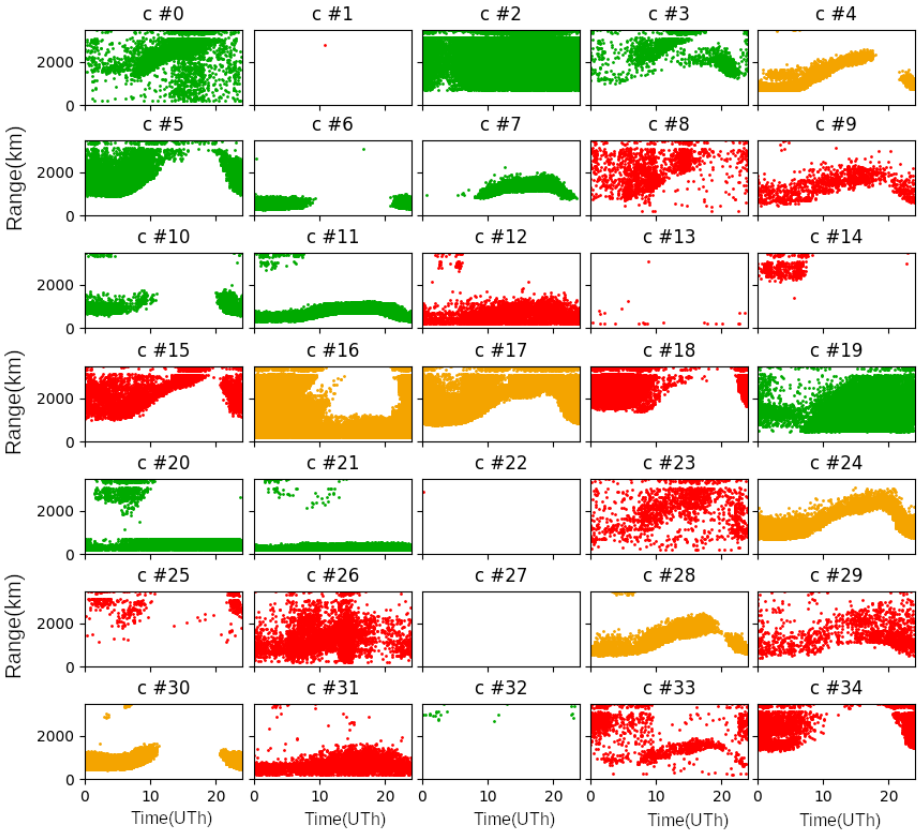

Рис. 10. Результаты классификации данных радара MAGW за январь–май 2023 г.: суточно-дальностное распределение наблюдений каждого класса. Цвета соответствуют различным типам сигнала аналогично рис. 5

волны, а также трехмерную структуру коэффициента преломления ионосферы. Очевидно, что в сложных ситуациях, когда траекторию распространения сложно предсказать (нет уверенных измерений угла места приходящей радиоволны или отсутствует достаточно точная модель ионосферы), этот метод будет давать существенные ошибки, что объясняет широкое применение более простых методов, основанных на измерении скорости и спектральной ширины [Ponomarenko et al., 2007; Blanchard et al., 2009] на высокоширотных радарах. Для калиброванных среднеширотных радаров с использованием модели IRI, как показывает работа, разработанный нами метод может быть применим.

Проверка алгоритма на наблюдениях 2023 г.

Для проверки работоспособности модели проведена обработка данных, не использовавшихся ранее при обучении и в отличающихся геофизических условиях, на данных радара MAGW за первую половину 2023 г.

Результаты обработки (суточно-дальностные распределения сигналов различных классов и распределения различных классов по характеристикам) показаны на рис. 10, 11. Видно хорошее качественное согласие с результатами обработки исходных (обучающих) данных 2021 г. (см. рис. 6, 7). При моделировании распространения радиоволны использовалась модель IRI-2020 [Bilitza et al., 2022].

Основной особенностью данных 2023 г. является существенное отличие в уровне возмущенности ионосферы. Согласно данным Королевской обсерватории Бельгии, 2021 г. характеризовался среднегодовым числом солнечных пятен 30, в то время как для первой половины 2023 г. оно составляло 129. Это приводит как к более активному рассеянию различных ионосферных типов, так и к ухудшению точности прогноза распространения радиоволн моделью IRI в возмущенных условиях (ошибка траекторных расчетов обычно увеличивается с ростом дальности). Другой особенностью этих данных является, по-видимому, менее точная калибровка радара по углу места (видно, например, по изменению распределения метеоров на рис. 11, *б*).

Следует отметить, что доля сигналов, рассеянных земной поверхностью, по сравнению с 2021 г. сократилась до 24 %, доля сигналов, рассеянных ионосферой, увеличилась до 51 %, доля неинтерпретируемых сигналов возросла в два раза до 10 %, а доля сигналов, которые по-разному определяются различными сетями, выросла в 1.5 раза до 15 %. Таким образом, за первую половину 2023 г. алгоритм позволил автоматически проинтерпретировать 75 % всех сигналов, что на 10 % меньше, чем в 2021 г.

На рис. 10 видны достоинства и ограничения предложенного метода. К достоинствам можно отнести качественное согласие суточно-дальностных распределений сигналов различных классов с данными 2021 г., что говорит о возможности применения метода на новых данных.

К недостаткам нужно отнести наблюдаемое ухудшение точности определения качества классификации

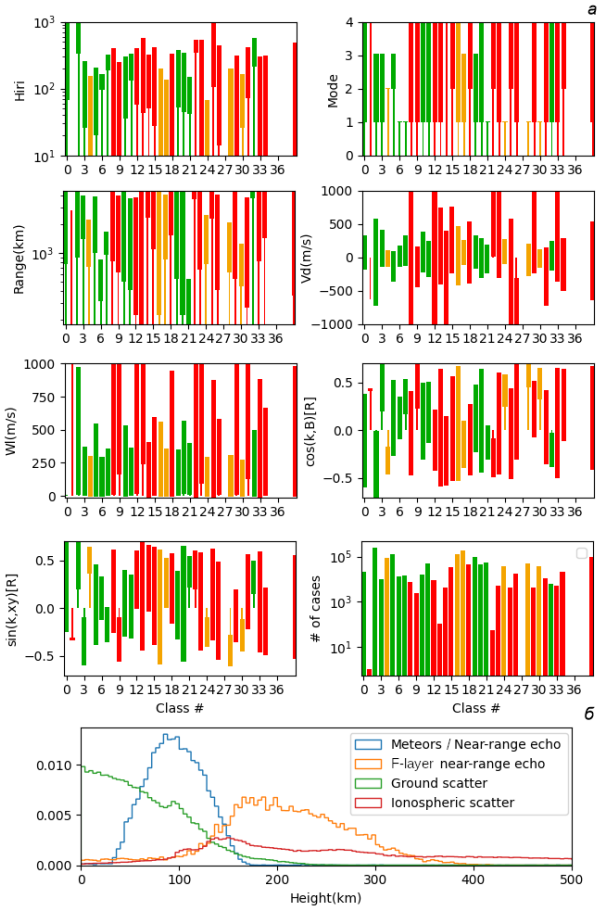

Рис. 11. Результаты классификации данных радара MAGW за январь–май 2023 г.: 95%-й интервал изменения основных параметров в каждом классе, цвета и типы параметров аналогичны рис. 5 (*а*); распределение сигналов различных типов по высотам, типы сигналов аналогичны рис. 8 (*б*)

на дальностях выше 2000 км. Наиболее показательным является класс 20 (F-layer near-range echo) на рис. 6, 7, 10. Из сравнения рисунков видно, что метод в 2023 г. чаще ошибается на дальностях выше 2000 км, где ожидаемо накапливается существенная ошибка расчетов траектории. Похожий эффект наблюдается и в некоторых других классах: ионосферном рассеянии (классы 10, 11) и метеорном/near-range эхе (класс 21). Косвенным признаком понижения качества расчетов являются также изменения в модовом составе сигналов (см. рис. 5, 11): практически все классы стали включать более высокие моды, чем в 2021 г., что говорит о сложностях в траекторных расчетах, и может быть связано в том числе с увеличением уровня фоновой возмущенности ионосферы в 2023 г. по сравнению с 2021 г. Высокие скорости в классах рассеяния от земли также говорят об увеличении возмущенности в ионосфере и о сопровождающей крупномасштабной волновой активности в фоновой ионосфере.

К недостаткам модели можно отнести невозможность разделения очень близких классов, например E-layer near-range echo и метеорного эха, объединенного этой моделью в один класс (класс 21). Этот недостаток модели вызван двумя ее особенностями — локальностью (она не учитывает временного поведения неоднородностей на больших временах жизни, поскольку использует эквивалентное стандартное спектральное разрешение 7-импульсной последовательности, ограничивающее времена жизни ~50 мс), и точностью определения высоты (параметры этих неоднородностей не могут быть разделены с необходимой точностью из-за недостаточной точности определения угла места).

Сезонно-суточные особенности наблюдения различных классов

Сезонно-суточные особенности наблюдения различных классов сигналов в течение 2021 г. показаны на рис. 12.

На панелях *а*–*г* даны суточная зависимость встречаемости различных классов (приведенная к локальному солнечному времени в расчетной точке рассеяния) и сезонные зависимости наблюдения различных классов сигналов. Видно, что большая часть сигналов на радаре EKB наблюдается днем, на радаре MAGW суточный эффект в сигналах менее выражен. Сезонная зависимость сильнее выражена на радаре MAGW и слабее — на EKB. Подобный эффект может быть связан с тем, что радар MAGW расположен полярнее EKB. Поэтому освещенность на радаре MAGW имеет более выраженную сезонную динамику, а ионосферная динамика контролируется магнитосферой в большей степени, чем на EKB.

На панелях *д*, *е* показана статистика рассеяния на земной поверхности 2-го и 1-го скачков (классы 17 и 24). Как и ожидалось, это рассеяние наблюдается в основном в дневное время, когда электронная концентрация в ионосфере достаточно высока для отражения радиосигнала от ионосферы, а в летнее время сигналы экранируются near-range echo слоев E и F.

На панели *ж* показано F-layer near-range echo (класс 20), наиболее интенсивное в летнее время, которое является одной из причин экранировки сигналов, рассеянных от земной поверхности. Видно, что на радаре MAGW этот тип рассеяния может наблюдаться также в области солнечного терминатора.

На панели *з* показан смешанный класс сигналов — meteor echo/near-range echo (класс 21), который сложно разделить по характеристикам рассеяния — близким высотам и низким скоростям [Ponomarenko et al., 2016]. Видно, что наиболее часто сигналы этого класса наблюдается в ночное время (соответствует наблюдению метеоров) и в летнее время (соответствует наблюдениям near-range echo).

На панели *и* показано возможное рассеяние на луче Педерсена (класс 10) [Ponomarenko et al., 2011]. Видно, что наиболее часто этот тип рассеяния наблюдается зимой в дневное время.

На панели *к* показан возможный кандидат на квазидвухпозиционное рассеяние на магнитоориентированных неоднородностях (класс 7), возможность которого предсказывалась в работах [Kravtsov, Namazov, 1980; Berngardt et al., 2016] и связана с тем, что траектории сигнала в прямом и обратном направлениях могут не совпадать, поэтому условие ортогональности рассеяния к магнитному полю Земли для траектории принятой рассеянной радиоволны не соблюдается [Berngardt et al., 2016].

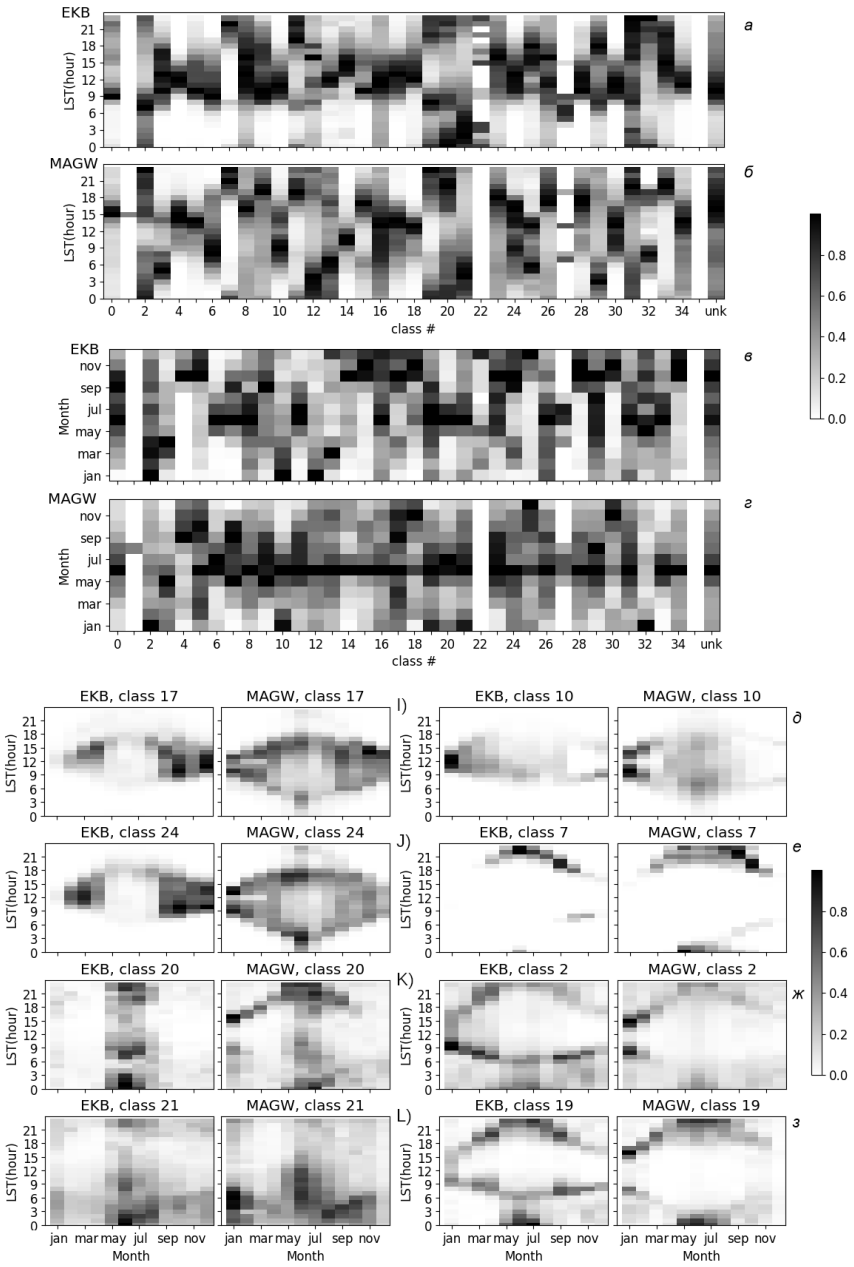

Рис. 12. Суточная зависимость встречаемости различных классов, приведенная к локальному солнечному времени в расчетной точке рассеяния (*а*, *б*); сезонная зависимость встречаемости различных классов (*в*, *г*); суточно-сезонные зависимости наблюдения различных классов: *д*, *е* — рассеяния на земной поверхности второго и первого скачка (классы 17 и 24); *ж* — F-layer near-range echo (класс 20); *з* — meteor echo/near-range echo (класс 21); *и* — рассеяние на луче Педерсона (класс 10); *к* — квази-двухпозиционное рассеяние (класс 7); *л* — ракурсное рассеяние в F-слое 0.5 скачка (класс 2); *м* — рассеяние в E/F-слое 0.5 скачка со слабой ракурсной зависимостью (класс 19)

На панелях *л*, *м* показаны два примера рассеяния в ионосфере: ракурсное рассеяние в F-слое скачка 0.5 (класс 2), и рассеяние в E/F-слое скачка 0.5 со слабой ракурсной зависимостью (класс 19). Видно, что рассеяние этих типов в основном наблюдается в неосвещенное время суток, что качественно соответствует эмпирическим закономерностям.

На панелях *д*–*м* видно, что многие неоднородности интенсифицируются вблизи солнечного терминатора, что связано с его высокой пространственно-временной динамикой. Из рис. 12 также видно, что рассеянные сигналы многих типов можно подразделять на преимущественно дневные (в основном рассеяние от земной поверхности) и преимущественно ночные (в основном рассеяние на ионосферных неоднородностях).

ЗАКЛЮЧЕНИЕ

В работе сделана попытка решения задачи автоматической классификации данных радаров когерентного рассеяния с помощью минимизации влияния субъективного человеческого мнения на подготовку и интерпретацию данных, а также анализа полученного решения.

В рамках самосогласованного управляемого данными подхода был разработан метод автоматиче-

ского построения такого классификатора. В результате применения этого метода была построена, обучена и исследована компактная математическая модель (8)–(9), позволяющая автоматически классифицировать данные радаров EKB и MAGW по радиофизическим особенностям их распространения и рассеяния. Количество свободных параметров модели составляет 2240, количество обнаруженных классов сигналов — 35. Работа является обобщением, улучшением и математически более строгим развитием подхода, изложенного в предыдущих работах [Berngardt et al., 2022; Бернгардт, 2022].

Получены следующие результаты.

Разработана эмпирическая модель для аугментации результатов зондирования 16-импульсной последовательностью до точностей 7-импульсной последовательности (3)–(4). Такое предыскажение необходимо для тренировки единой классифицирующей модели, не зависящей от типа используемого зондирующего сигнала.

Создана и обучена нейронная сеть (классификатор), использующая абсолютные функции активации [Vallés-Pérez et al., 2023] и выходной слой линейной нормализации (см. рис. 1, *в*). С использованием метода поиска минимального числа нейронов в полносвязном слое нейронной сети [Berngardt, 2024] найдено минимальное число нейронов в этой сети: 49 и 35 нейронов на первом и втором слое соответственно. Число нейронов на втором слое соответствует числу скрытых классов сигналов в данных. Обученная нейронная сеть обеспечивает высокое качество повторения результатов использованной кластеризации (0.92 по метрике AUC-PR), существенно превышающее метрику качества 0.68 предыдущей модели [Berngardt et al., 2022; Бернгардт, 2022]. Полученная сеть намного меньше, чем предыдущий вариант [Berngardt et al., 2022; Бернгардт, 2022] и имеет 2240 параметров.

Сравнением результатов обучения сети для двух кластеризаций (Gaussian mixture+BIC критерий для оценки числа кластеров (GMBIC) и GMsDB) [Berngardt, 2023] было показано, что количество разделимых классов в данных радаров составляет от 23 до 35 (см. рис. 4), для удобства дальнейшей интерпретации было выбрано 35.

В зависимости от набора данных для тренировки сети (кросс-валидация по трем фолдам) форма обнаруженных классов может незначительно меняться: скорректированный индекс Рэнда между прогнозами моделей равен 0.936–0.967 для радара EKB, и 0.897–0.937 для радара MAGW. На основе трех вариантов обученной модели была построена и проанализирована ансамблевая модель классификатора по принципу голосования. Случаи, когда все три модели не предсказывают идентичного класса для данных, объединялись в отдельный класс, интерпретируемый как неуверенный результат прогноза. Доля таких данных сравнительно мала (10 %), что говорит о высокой преемственности результатов каждой из сетей и возможности их раздельного использования.

Были проанализированы выявленные 35 наблюдаемых классов сигналов, для каждого из этих классов проведена его предварительная интерпретация. Показано что 19 классов могут быть интерпретированы с физической точки зрения (13 типов рассеяния от ионосферы и шесть типов рассеяния от земной поверхности), остальные включают в себя высокие скорости и спектральные ширины (~1000 м/с). Найдены параметры, наиболее сильно влияющие на определение каждого класса, а также показано, что самыми важными параметрами являются высота рассеяния и угол места распространения сигнала в точке рассеяния. Спектральная ширина сигнала и мода его распространения относятся к наименее важным параметрам (см. рис. 9).

Приведены суточно-дальностные распределения этих сигналов на радарах EKB и MAGW (см. рис. 6, 7), по которым можно видеть, что многие интерпретированные классы обладают ожидаемым суточным ходом.

Результаты анализа данных радара MAGW 2023 г., не используемых при обучении моделей, показали успешную работоспособность модели в условиях более возмущенной ионосферы и выявили слабые стороны этой модели: ожидаемую зависимость от качества расчетов траектории распространения радиоволны, приводящую к падению качества классификации на дальностях выше 2000 км (см. рис. 10, 11).

Предложенный метод позволил автоматически классифицировать примерно 85 % всех данных, принимаемых в главном лепестке диаграммы направленности, на радарах EKB и MAGW в спокойном 2021 г. и примерно 75 % данных радара MAGW за первую половину возмущенного 2023 г. Параметры обученной модели трех вариантов доступны на [https://github.com/berng/WrappedClassifier/tree/master/v.3.0]. Результаты обработки сигналов предложенным алгоритмом в реальном масштабе времени доступны на [http://sdrus.iszf.irk.ru/node/107].

Исследование выполнено за счет гранта Российского научного фонда № 24-22-00436, [https://rscf.ru/project/24-22-00436/].

СПИСОК ЛИТЕРАТУРЫ

Бернгардт О.И. Первый сравнительный анализ метеорного эхо и спорадического рассеяния, идентифицированных самообучившейся нейронной сетью по данным радаров EKB и MAGW ИСЗФ СО РАН. *Солнечно-земная физика*. 2022, т. 8, № 4, с. 66–76. DOI: 10.12737/szf-84202206 / Berngardt O.I. The first comparative analysis of meteor echo and sporadic scattering identified by a self-learned neural network in EKB and MAGW ISTP SB RAS radar data. *Solar-Terrestrial Physics*. 2022, vol. 8, no. 4, pp. 63–72. DOI: 10.12737/stp-84202206.

Бернгардт О.И., Куркин В.И., Кушнарев Д.С. и др. Декаметровые радары ИСЗФ СО РАН. *Солнечно-земная физика*. 2020, т. 6, № 2, с. 79–92. DOI: 10.12737/szf-62202006 / Berngardt O., Kurkin V., Kushnarev D., et al. ISTP SB RAS decameter radars. *Solar-Terrestrial Physics*. 2020, vol. 6, no. 2, pp. 63–73. DOI: 10.12737/stp-62202006.

Arnold V. On the function of three variables. *American Mathematical Society Translations*. 1963, pp. 51–54.

Barthes L., Andrè R., Cerisier J.-C., Villain J.-P. Separation of multiple echoes using a high-resolution spectral analysis for SuperDARN HF radars. *Radio Sci*. 1998, vol. 33, no. 4, pp. 1005–1017. DOI: 10.1029/98RS00714.

Berngardt O.I. Superclustering by finding statistically significant separable groups of optimal Gaussian clusters. 2023. DOI: 10.48550/arXiv.2309.02623.

Berngardt, O.I. Minimum number of neurons in fully connected layers of a given neural network (the first approximation). 2024. DOI: 10.48550/arXiv.2405.14147.

Berngardt O.I., Kutelev K.A., Potekhin A.P. SuperDARN scalar radar equations. *Radio Sci*. 2016, vol. 51, no. 10, pp. 1703–1724. DOI: 10.1002/2016RS006081.

Berngardt O.I., Grkovich K.V., Fedorov R.R. Synthesis of Symmetric Sounding Sequences for Ekaterinburg Coherent Decameter Radar. *Radiophysics and Quantum Electronics*. 2020, vol. 62, no. 11, pp. 721–733. DOI: 10.1007/s11141-020-10018-y.

Berngardt O.I., Fedorov R.R., Ponomarenko P., Grkovich K.V. Interferometric calibration and the first elevation observations at EKB ISTP SB RAS radar at 10–12 MHz. *Polar Science*. 2021, vol. 28, p. 100628. DOI: 10.1016/j.polar.2020.100628.

Berngardt O.I., Kusonsky O.A., Poddelsky A.I., Oinats A.V. Self-trained artificial neural network for physical classification of ionospheric radar data. *Adv. Space Res*. 2022, vol. 70, no. 10, pp. 2905–2919. DOI: 10.1016/j.asr.2022.07.054.

Bilitza D., Pezzopane M., Truhlik V., et al. The International Reference Ionosphere Model: A review and description of an ionospheric benchmark. *Rev. Geophys*. 2022, vol. 60, no. 4, e2022RG000792. DOI: 10.1029/2022RG000792.

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, vol. 44, no. 5. DOI: 10.1029/2009RS004141.

Breiman Leo. Random forests. *Machine Learning*. 2001, vol. 45, no. 1, pp. 5–32. DOI: 10.1023/a:1010933404324.

Burrell A.G., Milan S.E., Perry G.W., Automatically determining the origin direction and propagation mode of high-frequency radar backscatter. *Radio Sci*. 2015. vol. 50, no. 12, pp. 1225–1245. DOI: 10.1002/2015RS005808.

Chisham G., Freeman M.P. A reassessment of SuperDARN meteor echoes from the upper mesosphere and lower thermosphere. *J. Atmos. Solar-Terr. Phys*. 2013, vol. 102, pp. 207–221. DOI: 10.1016/j.jastp.2013.05.018.

Chisham G., Lester M., Milan S.E., et al. A decade of the Super Dual Auroral Radar Network (SuperDARN): scientific achievements, new techniques and future directions. *Surveys in Geophysics*. 2007, vol. 28, pp. 33–109. DOI: 10.1007/s10712-007-9017-8.

Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. B: *Proc. Second International Conference on Knowledge Discovery and Data Mining*. KDD96. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

Goodfellow I., Bengio Y., Courville A. Deep learning. *Adaptive Computation and Machine Learning*. MIT Press, 2016.

Greenwald R.A., Baker K.B., Dudeney J.R., et al. Darn/Superdarn: A global view of the dynamics of high-latitude convection. *Space Sci. Rev*. 1995, vol. 71, pp. 761–796. DOI: 10.1007/BF00751350.

Greenwald R.A., Oksavik K., Barnes R., et al. First radar measurements of ionospheric electric fields at sub-second temporal resolution. *Geophys. Res. Lett*. 2008, vol. 35, no. 3. DOI: 10.1029/2007GL032164.

Guo Chuan, Geoff Pleiss, Yu Sun, Weinberger K.Q. On calibration of modern neural networks. 2017. DOI: 10.48550/arXiv.1706.04599.

Huang X., Kroening D., Ruan W., et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Sci. Rev*. 2020, vol. 37, p. 100270. DOI: 10.1016/j.cosrev.2020.100270.

Hubert L., Arabie P. Comparing partitions. *J. Classification*. 1985, vol. 2, no. 1, pp. 193–218. DOI: 10.1007/bf01908075.

Ioffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. DOI: 10.48550/arXiv.1502.03167.

Kolmogoroff A. Grundbegriffe der Wahrscheinlichkeitsrechnung. *Springer Berlin Heidelberg*. 1933. DOI: 10.1007/978-3-642-49888-6.

Kolmogorov A.N. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*. 1957, pp. 953–956.

Kong Xing, Liu E., Shi S., Chen F. The implementation of deep clustering for SuperDARN backscatter echoes. *Adv. Space Res*. 2024, vol. 74, no. 1, pp. 243–254. DOI: 10.1016/j.asr.2024.03.039.

Kravtsov Y.A., Namazov S.A. Characteristics of scattering of radio waves from magnetically oriented inhomogeneities of the ionosphere near critical frequency. *Radiotekhnika i elektronika* [J. Communications Technology and Electronics]. 1980, pp. 459–466. [In Russian].

Kunduri B.S.R., Baker J.B.H., Ruohoniemi J.M., et al. An examination of SuperDARN backscatter modes using machine learning guided by ray-tracing. *Space Weather*. 2022, vol. 20, no. 9, e2022SW003130. DOI: 10.1029/2022SW003130.

Lester M., Chapman P.J., Cowley S.W.H., et al. Stereo CUTLASS — A new capability for the SuperDARN HF radars. *Ann. Geophys*. 2004, vol. 22, no. 2, pp. 459–473. DOI: 10.5194/angeo-22-459-2004.

Milan S.E., Jones T.B., Robinson T.R., et al. Interferometric evidence for the observation of ground backscatter originating behind the CUTLASS coherent HF radars. *Ann. Geophys*. 1997, vol. 15, no. 1, pp. 29–39. DOI: 10.1007/s00585-997-0029-y.

Nishitani N., Ruohoniemi J.M., Lester M., et al. Review of the accomplishments of mid-latitude Super Dual Auroral Radar Network (SuperDARN) HF radars. *Progress in Earth and Planetary Science*. 2019, vol. 6, no. 1. DOI: 10.1186/s40645-019-0270-5.

Ponomarenko P.V. Blessing Iserhienrhien и Jean-Pierre St.-Maurice. Morphology and possible origins of near-range oblique HF backscatter at high and midlatitudes. *Radio Sci*. 2016, vol. 51, no. 6, pp. 718–730. DOI: 10.1002/2016RS006088.

Ponomarenko P., McWilliams K.A. Climatology of HF Propagation Characteristics at Very High Latitudes From SuperDARN Observations. *Radio Sci*. 2023, vol. 58, no. 5, e2023RS007657. DOI: 10.1029/2023RS007657.

Ponomarenko P.V., Waters C.L., Menk F.W. Factors determining spectral width of HF echoes from high latitudes. *Ann. Geophys*. 2007, vol. 25, no. 3, pp. 675–687. DOI: 10.5194/ angeo-25-675-2007.

Ponomarenko P.V., Koustov A.V., St.-Maurice J.-P., Wiid J. Monitoring the F-region peak electron density using HF backscatter interferometry. *Geophys. Res. Lett*. 2011, vol. 38, no. 21. DOI: 10.1029/2011GL049675.

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, vol. 46, no. 4. DOI: 10.1029/2011RS004676.

Ribeiro A.J., Ruohoniemi J.M., Ponomarenko P.V., et al. A comparison of SuperDARN ACF fitting methods. *Radio Sci*. 2013, vol. 48, no. 3, pp. 274–282. DOI: 10.1002/rds.20031.

Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Computational and Applied Mathematics*. 1987, vol. 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

Rumelhart D.E., Hinton G.E., Williams R.J. *Learning Internal Representations by Error Propagation. B: Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations.* The MIT Press. 1986. DOI: 10.7551/mitpress/5236.003.0012.

Saxena Amit, Mukesh Prasad, Akshansh Gupta, et al. A review of clustering techniques and developments. *Neuro-*

computing. 2017, vol. 267, pp. 664–681. DOI: 10.1016/j.neucom.2017.06.053.

Schwarz Gideon. Estimating the dimension of a model. *The Annals of Statistics*. 1978, vol. 6, no. 2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Shorten C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *J. Big Data.* 2019, vol. 6, no. 1. DOI: 10.1186/s40537-019-0197-0.

Smiti A., Elouedi Z. Fuzzy density based clustering method: Soft DBSCAN-GM. *B: 2016 IEEE 8th International Conference on Intelligent Systems*, 2016, pp. 443–448. DOI: 10.1109/IS.2016.7737459.

Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. *Second*. The MIT Press. 2018.

Vallés-Pérez I., Soria-Olivas E., Martıénez-Sober M., et al. Empirical study of the modulus as activation function in computer vision applications. *Engineering Applications of Artificial Intelligence*. 2023, vol. 20, p. 105863. DOI: 10.1016/j.engappai.2023.105863.

URL: https://github.com/berng/WrappedClassifier/tree/master/v.3.0 (дата обращения 15 апреля 2025 г.).

URL: http://sdrus.iszf.irk.ru/node/107 (дата обращения 15 апреля 2025 г.).

URL: https://rscf.ru/project/24-22-00436/ (дата обращения 15 апреля 2025 г.).

Как цитировать эту статью:

Бернгардт О.И. Управляемый данными подход к классификации данных среднеширотных радаров когерентного рассеяния. *Солнечно-земная физика*. 2025, т. 11, № 2, с. 22–44. DOI: 10.12737/szf-112202503.

Приложение 1

АЛГОРИТМ ОЦЕНКИ ЗНАЧИМОСТИ ПАРАМЕТРОВ

Для определения набора наиболее важных параметров при классификации воспользуемся следующей (жадной) модификацией перестановочного алгоритма:

1. Выбираем скрытый класс C, который хотим проанализировать.

2. Для всего набора данных x_{ij} прогнозируем скрытые классы для всех доступных данных.

3. Отбираем часть из набора данных x_{ijC}, прогнозный класс которого соответствует выбранному классу (x_{ijC}: $Prediction(x_{ijC})=C$), далее работаем только с ним;

4. Создаем пустое множество важных компонент $F=\varnothing$;

5. Если ранее были определены какие-то компоненты как важные, перемешиваем (permutate) каждый соответствующий столбец независимо, неважные компоненты добавляем в набор данных без перемешивания:

$$x'_{i,\,jC}=\begin{cases} i\notin F: x_{i\,jC} \\ i\in F: permutate\left(x_{i\,jC}\right) \end{cases}.$$

6. Вычисляем значение качества классификации по набору данных $x'_{i,\,jC}$, соответствующее такому перемешиванию, как долю правильных прогнозов этого класса в наборе данных x_{ijC} шага 3 (по отношению к исходному полному набору данных x_{ij} шага 2 мы фактически вычисляем метрику Precision по отношению к выбранному классу C):

$$Q_F=\frac{Dim\left(Prediction\left(x'_{i\,jC}\right)=C\right)}{Dim\left(x_{i\,jC}\right)}.$$

7. Выбираем l-й столбец в наборе данных $x'_{i\,jC}$ и вычисляем по нему качество при перемешивании

$$x^{(l)}_{i,\,jC}=\begin{cases} i\neq l: x'_{i\,jC} \\ i=l: permutate\left(x'_{i\,jC}\right) \end{cases},$$

$$Q_{F,\,l}=\frac{Dim\left(Prediction\left(x^{(l)}_{i\,jC}\right)=C\right)}{Dim\left(x_{i\,jC}\right)}.$$

8. Вычисляем уменьшение качества:

$$\Delta Q_l=Q_F-Q_{F,\,l}.$$

9. Выбираем столбец l_{opt}, на котором уменьшение качества ΔQ_l при пермутации наибольшее,

$$l_{opt}=argmax\left(\Delta Q_l\right).$$

10. Если этот столбец совпадает с одним из выбранных ранее столбцов ($l_{opt}\in F$) или если уменьшения качества нет (найденное наименьшее качество не хуже качества, полученного на шаге 6, $\left(\Delta Q_{l_{opt}}\le 0\right)$ — завершаем поиск.

11. В остальных случаях добавляем найденный столбец в список важных параметров $\left(F=F\vee l_{opt}\right)$ и возвращаемся к шагу 5.

Полученное множество значений F является множеством параметров, существенно влияющих на результат прогноза класса C.

Алгоритм является алгоритмом поиска с жадным добавлением компонент и имеет критерий останова, позволяющий учитывать только значимые компоненты, перемешивание по которым ухудшает качество классификации.

Для большей статистической значимости метода качество Q_F, $Q_{F,l}$ на шагах 6, 7 определяется не как качество Q, определенное по всем элементам наборов данных $x'_{i,\,jC}$, $x^{(l)}_{i,\,jC}$, а как 95%-й интервал значений этого качества [min(Q), max(Q)] с использованием множества случайных выборок (с повторениями) из этих наборов данных той же длины (bootstrap-метод). Уменьшение качества $\triangle Q_{l_{opt}}$ в этом случае (шаг 8) оценивается как разница между верхней гранью доверительного интервала на проверяемом параметре (шаг 7) и нижней гранью доверительного интервала на наборе данных предыдущего шага (шаг 6): $\triangle Q_l=\min\left(Q_F\right)-\max\left(Q_{F,l}\right)$. Такой алгоритм позволяет выявить более важные параметры с приблизительным уровнем значимости 0.05, а статистически менее значимые параметры отвергнуть.

Приложение 2

ПОЯСНЕНИЕ НЕКОТОРЫХ ТЕРМИНОВ

- *argmax*(*f*(*x*)) — функция, возвращающая значение аргумента *x*, при котором функция *f*(*x*) максимальна, для векторов — номер координаты, значение которой максимально;
- $\overrightarrow{SoftMax}(\vec{x})$ — распространенная в машинном обучении векторнозначная функция, возвращающая вектор с неотрицательными компонентами согласно формуле

$$SoftMax_i(\vec{x}) = \frac{\exp(x_i)}{\sum_j \exp(x_j)};$$

- $\overrightarrow{BatchNormalization}(\vec{x})$ — суперпозиция двух линейных преобразований $\vec{x}$: первое возвращает на обучающем наборе данных величину с нулевым средним и единичной дисперсией, а второе — с фитируемыми в процессе обучения коэффициентами, улучшающими качество обучения модели;
- ширина слоя нейронной сети — количество нейронов в этом слое;
- фолд — часть набора данных при обучении методом кросс-валидации, обычно при этом весь набор данных делится на равные части (фолды);
- области рассеяния на скачках распространения 0.5, 1, 1.5 и т. д. поясняются на рис. 13. Границами разделения являются область отражения от ионосферы и область рассеяния от земной поверхности — места смены знака вертикальной компоненты волнового вектора радиоволны.

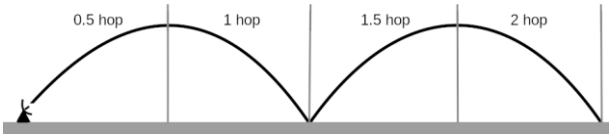

Рис. 13. К объяснению скачковых областей рассеяния