УДК 004.852, 551.59
DOI: 10.12737/szf-84202206

Поступила в редакцию 04.05.2022
Принята к публикации 19.09.2022

ПЕРВЫЙ СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТЕОРНОГО ЭХА И СПОРАДИЧЕСКОГО РАССЕЯНИЯ, ИДЕНТИФИЦИРОВАННЫХ САМООБУЧИВШЕЙСЯ НЕЙРОННОЙ СЕТЬЮ ПО ДАННЫМ РАДАРОВ EKB И MAGW ИСЗФ СО РАН

THE FIRST COMPARATIVE ANALYSIS OF METEOR ECHO AND SPORADIC SCATTERING IDENTIFIED BY A SELF-LEARNING NEURAL NETWORK IN EKB AND MAGW ISTP SB RAS RADAR DATA

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. В работе описана текущая версия алгоритма автоматической классификации сигналов (v.1.1), принимаемых радарами декаметрового когерентного рассеяния ИСЗФ СО РАН. Алгоритм представляет собой самообучающуюся нейронную сеть, определяющую тип рассеянных сигналов по результатам физического моделирования распространения радиоволн с использованием радарных данных и международных ссылочных моделей ионосферы и магнитного поля Земли. Используя данные радаров MAGW и EKB ИСЗФ СО РАН за 2021 г., алгоритм самостоятельно обучается группировать рассеянные сигналы на изначально неизвестные классы. Такое деление основано на физически интерпретируемых параметрах распространения радиоволн и измеренных радаром данных, при этом из 20 возможных скрытых классов выделяются 15 часто наблюдаемых, из которых 14 могут быть интерпретированы с физической точки зрения. Для демонстрации работы алгоритма представлен первый статистический анализ наблюдений сигналов, отнесенных алгоритмом к двум классам, интерпретируемым нами как рассеяние на метеорных следах и рассеяние с участием спорадического слоя E соответственно. На основе статистического анализа данных радаров EKB и MAGW за 2021–2022 гг. определены дальностно-высотные характеристики сигналов этих классов, показана корреляция между среднечасовыми количествами наблюдений обоих классов, а также их среднечасовыми продольными скоростями. Полученные результаты позволяют интерпретировать сигналы этих классов как метеорное эхо и спорадическое рассеяние соответственно и использовать их для изучения процессов взаимодействия нейтральной атмосферы, изучаемой по данным метеорного рассеяния, и нижней ионосферы, изучаемой по наблюдениям за спорадическим рассеянием. В настоящее время представленный алгоритм классификации работает на радарах ИСЗФ СО РАН в автоматическом режиме.

Ключевые слова: машинное обучение, классификация сигналов, радары когерентного рассеяния, метеорное эхо, спорадическое рассеяние.

Abstract. The paper describes the current version (v.1.1) of the algorithm for automatic classification of signals received by ISTP SB RAS decameter coherent scatter radars. The algorithm is a self-learning neural network that determines the type of scattered signals from the results of physical modeling of radio wave propagation, using radar data and international reference models of the ionosphere and geomagnetic field. According to MAGW and EKB ISTP SB RAS radar data for 2021, the algorithm self-learns to classify scattered signals into initially unknown classes based on physically interpreted parameters of radio wave propagation and data measured by the radar, with 15 frequently observed out of 20 possible hidden classes identified, 14 of which can be interpreted from a physical point of view. To demonstrate the operation of the algorithm, we present the first statistical analysis of observations of signals assigned by the algorithm to classes which we interpret as scattering by meteor trails and scattering with the sporadic E layer respectively. Through a statistical analysis of EKB and MAGW radar data during 2021–2022, we demonstrate the range-altitude characteristics of signals of these types. A correlation is shown between the hourly average numbers of observations of both classes, as well as between the hourly average line-of-sight velocities obtained for both classes. The results obtained make it possible to interpret these classes as a meteor echo and sporadic scattering respectively, and to use radar data to study the interaction between the neutral atmosphere (studied from meteor scattering data) and the lower ionosphere (studied from observations of sporadic scattering). Currently, this classification algorithm works in ISTP SB RAS radars in automatic mode.

Keywords: machine learning, signal classification, coherent scatter radars, meteor echo, sporadic scattering.

ВВЕДЕНИЕ

Проблема классификации многомерных экспериментальных данных сложна, она интенсивно изучается в геофизике [Siwei, Ma, 2021]. Одним из средств диагностики и мониторинга состояния магнитосферы, ионосферы и верхних слоев атмосферы является международная сеть радаров когерентного рассеяния SuperDARN (Super Dual Auroral Radar Network) и подобные им импульсные декаметровые радары когерентного рассеяния. На данный момент в мире насчитывается более 35 таких инструментов [Nishitani et al., 2019]. Большое количество данных, предоставляемых этими радарами, сопровождается трудностями автоматической интерпретации. Каждый радар излучает последовательности зондирующих импульсов и принимает рассеянные сигналы, используемые для изучения рассеивающих неоднородностей в атмосфере и ионосфере. Принимаемые радиолокационные сигналы представляют собой смесь сигналов, сформированных различными физическими механизмами рассеяния [Nishitani et al., 2019]. Поэтому важной проблемой интерпретации данных этих радаров является идентификация рассеянных сигналов различных типов. В настоящее время в этой задаче используются разные методы [Blanchard et al., 2009; Ribeiro et al., 2011; Lavygin et al., 2020], в том числе основанные на статистических методах и методах машинного обучения.

С 2012 г. ИСЗФ СО РАН эксплуатирует радар когерентного рассеяния EKB в Свердловской области, с 2020 г. — радар MAGW в Магаданской области [Бернгардт и др., 2020]. Сектор обзора радара EKB составляет −4°−+48°, радара MAGW — −66°−−16°. Радары работают в диапазоне частот 8–20 МГц, что увеличивает количество различных рассеянных сигналов, принимаемых ими, и расширяет диапазон дальности работы радара до 3500–4500 км за счет скачкового распространения радиоволн. С другой стороны, это усложняет интерпретацию принятых сигналов из-за сложности учета траектории распространения этих радиоволн в ионосфере. Эти радары аналогичны по программно-аппаратному обеспечению радарам SuperDARN. В конце 2020 г. на обоих радарах были начаты регулярные угломестные измерения и проведена их качественная калибровка. Использование угла места принятого сигнала позволяет оценить траекторию его распространения и сформулировать задачу автоматической классификации принятых сигналов с учетом процесса их распространения и рассеяния. Традиционно радары SuperDARN не используют угломестную информацию для автоматической классификации данных.

Обычно такую задачу решают с первоначальной разметки данных на классы: рассеяние от ионосферы, рассеяние от земной поверхности, рассеяние от метеоров и т. д., обычно без учета траектории распространения. В предлагаемом подходе задача решается как построение и обучение схемы, которая позволит алгоритму самостоятельно разделить данные на подходящие классы с учетом процесса распространения этих сигналов, а потом дать возможность исследователю интерпретировать каждый полученный класс с физической точки зрения.

Предложенный подход объединяет процесс обучения (классификацию) и процесс автоматической разметки (кластеризацию) в единую схему поиска заранее неизвестных классов в данных только по физически интерпретируемым параметрам, как измеренным радаром, так и полученным в результате численного моделирования. За счет этого обеспечивается, с одной стороны, автоматизация процесса обучения на огромном массиве доступных данных, а с другой стороны, — физическая интерпретируемость результатов.

Основная идея предложенного двухэтапного метода заключается в использовании на первом этапе кластеризации для всех доступных данных — как экспериментальных, так и численно смоделированных. Эта кластеризация математическая, она обоснована лишь относительно понятным с физической точки зрения требованием, чтобы исследуемые сигналы были разбиты в многомерном пространстве параметров на некие ограниченные области, близкие по форме к многомерным элипсоидам. Это соответствует предположению, что сигналы, имеющие различные физические механизмы формирования, должны хоть каким-то образом отличаться друг от друга во всем многомерном пространстве измеренных и полученных в результате численного моделирования параметров.

После такой кластеризации данных на втором этапе обучается классификатор (искусственная нейронная сеть) так, чтобы он, с одной стороны, использовал только хорошо интерпретируемые с физической точки зрения параметры, отобранные нами из всех доступных, с другой стороны, чтобы получаемые им классы («скрытые классы») были похожи на те, которые выдает кластеризация. С точки зрения машинного обучения подход близок к поиску оптимального векторного представления данных (оптимального эмбеддинга), в качестве координат которого выступают вероятности принадлежности сигнала к каждому из скрытых классов. Метод обучения назван «Обернутый классификатор с наивным учителем».

Традиционно исследователям проще всего классифицировать данные радаров визуально, анализируя суточный ход сигнала, учитывая непрерывность областей наблюдения сигналов и их динамику в координатах время — дальность. Например, метеорное эхо обычно наблюдается, как фрагментарные во времени сигналы на малых (до 450 км) дальностях. Рассеяние от земной поверхности в этих координатах наблюдается в виде подковообразной области, где большие дальности (концы подковы) соответствуют восходу и закату, а близкие (середина подковы) — полудню [Nishitani et al., 2019]. Рассеяние на магнитоориентированных неоднородностях часто наблюдается после заката в виде широкой по дальности и времени области. Поэтому в качестве алгоритма кластеризации была выбрана кластеризация на некие ограниченные области, причем кластеризация каждых суток измерения на каждом радаре

(экспериментов) проводилась независимо от остальных. Особенность алгоритма состоит в том, чтобы обучаемый классификатор научился разделять классы на всем наборе доступных экспериментов (радаров, дней) таким образом, чтобы генерируемая им классификация была близка к разбиению на кластеры каждого из таких экспериментов по отдельности с точностью до перестановки номеров классов.

МАТЕМАТИЧЕСКОЕ ОПИСАНИЕ ЗАДАЧИ И ОБУЧЕНИЕ КЛАССИФИКАТОРА

Предлагаемый алгоритм является развитием модели [Berngardt et al., 2022] и является его более простой и лучше интерпретируемой с физической и математической точки зрения версией.

Исходный набор данных, используемых для обучения алгоритма, включает в себя как измеренные, так и смоделированные параметры. Он представляет собой набор 15-мерных векторов $\vec{X}_{e,m}$, где индекс e нумерует эксперименты (дни и радары), а m — порядковый номер вектора наблюдений в рамках одного эксперимента. Координаты этого вектора соответствуют различным параметрам.

Часть параметров непосредственно измеряется радаром: время, азимут, радиолокационная дальность, частота зондирующего сигнала, угол места принятого сигнала, доплеровское смещение частоты и ширина спектра принимаемого сигнала.

Другая часть получается в результате численного моделирования распространения радиоволн, основанного на измеренных параметрах сигнала и постановке эксперимента — времени, географических координат радара, радиолокационной дальности, угла места, азимута зондирования и зондирующей частоты. В результате моделирования вычисляются восемь параметров, характеризующих траекторию распространения сигнала до точки рассеяния: высота рассеяния, эффективная высота рассеяния, количество скачков распространения, угол между траекторией распространения и магнитным полем Земли в точке рассеяния и углы между траекторией и горизонтом в точке рассеяния и на каждой четверти радиолокационного расстояния до точки рассеяния. Траектория распространения рассчитывается методом геометрической оптики (трассировка лучей) [Ginzburg, 1970] в приближении незамагниченной ионосферы. Выходные параметры модели распространения радиосигнала, используемые для обучения нейронной сети, показаны на рис. 1.

Для описания траектории выбраны следующие восемь параметров: синус угла между траекторией луча и горизонтальным направлением в точке на 1/4, 2/4, 3/4 и 4/4 радиолокационной дальности до точки рассеяния (групповой задержки) — параметры $\sin(k, xy)[R/4]$, $\sin(k, xy)[R/2]$, $\sin(k, xy)[R3/4]$, $\sin(k, xy)[R]$ соответственно; косинус угла между траекторией луча и магнитным полем Земли в точке на радиолокационной дальности сигнала — параметр $\cos(k, B)$; Mode — количество отражений от земной поверхности +1; высота рассеяния (высота точки траектории на зарегистрированной радиолокационной дальности) с учетом рефракции в ионосфере (h_{iri}) и без ее учета (h_{eff}). Дополнительные два параметра, достаточно часто используемые при разделении сигналов разных типов — зарегистрированные доплеровское смещение частоты и спектральное уширение сигнала (м/с). Параметры траектории выбраны таким образом, чтобы они были близки к нулю в следующих случаях: 1) рассеяние на магнитоориентированных неоднородностях, $\cos(k, B)$ близок к нулю; 2) рассеяние на земной поверхности (groundscatter 1-го скачка), $\sin(k, xy)[R/2]$ близок к нулю, поскольку вблизи этой точки происходит отражение от ионосферы и волновой вектор радиоволны в ней почти горизонтален; 3) двукратное рассеяние на земной поверхности (groundscatter 2-го скачка), $\sin(k, xy)[R/4]$, $\sin(k, xy)[R3/4]$ близки к нулю, поскольку вблизи этих точек происходит отражение от ионосферы и волновой вектор радиоволны в них почти горизонтален. Знак параметра $\sin(k, xy)[R]$ позволяет определить направление волнового вектора радиоволны: к Земле (отрицательное) или от Земли (положительное), что также упрощает последующую интерпретацию сигналов. Параметр Mode позволяет также разделить сигналы односкачкового и двухскачкового распространения, что важно при интерпретации. Высоты, скорости и спектральные ширины позволяют более корректно интерпретировать рассеянный сигнал и часто используются в различных методах классификации данных радаров SuperDARN и аналогичных [Ribeiro et al., 2013]. Описанные выше десять параметров полностью повторяют параметры, используемые в [Berngardt et al., 2022]. Таким образом, модель классификатора учитывает региональные особенности формирования рассеянных сигналов различных классов только через модели ионосферы и магнитного поля Земли (IRI и IGRF), а наблюдение сигналов того или иного класса проявляется только в относительной частоте появления таких сигналов, что является одним из физических предположений этой модели. Как будет показано ниже, различия в частоте появления сигналов различных классов на разных радарах действительно присутствуют. Очевидно, это предположение достаточно грубое, но в первом приближении, как показано в [Berngardt et al., 2022] и в этой работе, позволяет получать интерпретируемые результаты.

Ионосферная рефракция рассчитывается по международной эталонной модели ионосферы IRI [Bilitza et al., 2014] с параметрами, рекомендованными разработчиками. Магнитное поле Земли рассчитывается на основе международной эталонной модели магнитного поля IGRF [Thébault, 2015]. В качестве входных параметров расчета траектории радиоволны использовались угол места принимаемого сигнала (предполагалось, что он совпадает с углом места излучения); азимут луча радара; рабочая частота радара; географическое положение радара; дата; время; радиолокационная дальность (групповая задержка сигнала) от радара до точки рассеяния. Необходимая гладкость ионосферы в расчетах обеспечивается ее аппроксимацией локальными B-сплайнами 2-го порядка. Ионосфера предполагается двумерно-неоднородной (в плоскости распространения сигнала

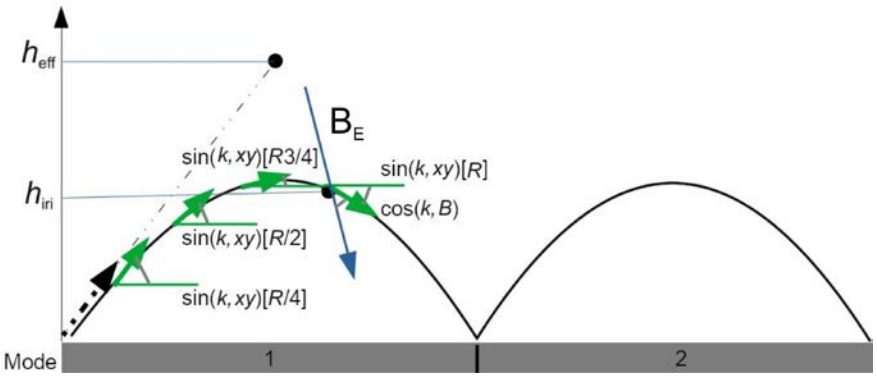

Рис. 1. Схема расчета физических параметров, определенных в результате численного моделирования. Черная сплошная линия — траектория распространения радиосигнала, рассчитанная с помощью трассировки лучей. Синяя стрелка — направление магнитного поля Земли, зеленые горизонтальные линии отмечают направление горизонта, зеленые сплошные стрелки — направление распространения радиосигнала

по дальности и по высоте). Более детально особенности моделирования рассмотрены в [Berngardt et al., 2022].

Первым этапом обучения модели является разбиение всех доступных 15-мерных данных $\vec{X}_{e,m}$ на кластеры — компактные области, по форме напоминающие элипсоиды в 15-мерном пространстве. Для этого распределение значений данных $P_e\left(\vec{X}_{e,m}\right)$ аппроксимируется линейной комбинацией двадцати 15-мерных распределений Гаусса $p_{y_{e,n}}$ с уникальными для каждого эксперимента параметрами

$$P_e\left(\vec{X}_{e,m}\right) \approx \sum_{y_{e,n}=1}^{N} A_{y_{e,n}} p_{y_{e,n}}\left(\vec{X}_{e,m}\middle|\vec{\Theta}_{y_{e,n}}\right), \quad \sum_{y_{e,n}=1}^{N} A_{y_{e,n}} = 1. \tag{1}$$

Здесь $N=20$, веса $A_{y_{e,n}}$ и параметры $\vec{\Theta}_{y_{e,n}}$ 15-мерных распределений Гаусса рассчитываются путем подгонки экспериментальных данных $\vec{X}_{e,m}$ методом максимального правдоподобия [Dempster et al., 1977]. Решением этой задачи является поиск неизвестных параметров $A_{y_{e,n}}$, $\vec{\Theta}_{y_{e,n}}$, определяющих разделение набора данных $\vec{X}_{e,m}$, полученных в каждом из экспериментов e, на кластеры с номерами $y_{e,n}$.

Номера кластеров $y_{e,n}$, которым принадлежат соответствующие точки $\vec{X}_{e,m}$, — это числа от 1 до N, которые получаются автоматически как результат работы алгоритма кластеризации.

На втором этапе по построенной разметке данных $y_{e,n}$ обучается итоговый классификатор данных. Задача построения оптимального классификатора состоит в том, чтобы найти функцию $\vec{g}()$, которая получает на вход вектор $\vec{x}_{e,m}$, являющийся проекцией вектора $\vec{X}_{e,m}$ в 10-мерное подпространство отобранных нами физических параметров, удобных для последующей интерпретации.

Таким образом, классификация, которую мы хотим получить, должна, с одной стороны, достаточно хорошо повторять кластеризацию, полученную на предыдущем этапе, с другой стороны, использовать для этого только те параметры, которые мы впоследствии сможем уверенно интерпретировать с физической точки зрения. Это решение находится как приближенное решение задачи

$$\vec{f}_e\left(\vec{g}\left(\overrightarrow{PF}\left(\vec{x}_{e,m}\right)\right)\right) \approx \overrightarrow{OHE}\left(y_{e,n}\right). \tag{2}$$

Здесь функция-обертка $\vec{f}_e\left(\vec{z}\right)$ уникальна для каждого эксперимента e и имеет вид

$$\vec{f}_e\left(\vec{g}\right) = \overrightarrow{soft\max}\left(\sum_{l=1}^{K} C_l\left(\sum_{k=1}^{K} b_{k,l,e} g_k\right)\right), \quad b_{k,l,e}, C_{l,e} \geq 0, \quad soft\max_j\left(\vec{z}\right) = \frac{e^{z_j}}{\sum_i e^{z_j}}. \tag{3}$$

Все неизвестные параметры (C_l, $b_{k,l,e}$ и параметры функции $\vec{g}()$) находятся путем подбора экспериментальных данных как решения (2). Размерности векторов $\vec{f}_e$, $\vec{g}$ — это количество кластеров ($N=20$) и скрытых классов ($K=20$), принятых нами равными. Суммирование в формуле ведется по скрытым классам k для обеспечения корректной перестановки номеров классов в рамках каждого отдельного эксперимента e (более детально этот процесс рассмотрен в [Berngardt et al., 2022]. Эта перестановка уникальна для каждого эксперимента e, а суммирование по скрытым классам l ведется для возможного объединения похожих скрытых классов и фактически старается сделать прогноз вероятности класса, предсказываемого наивным учителем по линейной комбинации вероятностей скрытых классов, выдаваемых обернутым классификатором.

Решение задачи проводилось одной из широко используемых модернизаций метода градиентного спуска — импульсным методом ADAM [Goodfellow et al., 2016]. В качестве условия оптимальности ис-

пользовался классический подход к решению задач классификации — минимизация взвешанной перекрестной энтропии *WCE* [Goodfellow et al., 2016]

$$WCE = -\sum_{j,k} W_k Y_{right,\,j,k} \log\left(Y_{left,\,j,k}\right), \quad (4)$$

где W_k — балансировочные веса, обратно пропорциональные числу данных в классе k, индекс j нумерует объекты в обучающем датасете: пары (номер эксперимента, номер объекта в эксперименте), $Y_{right,\,j,\,k}$, $Y_{left,\,j,\,k}$ — правая и левая части уравнения (2) соответственно.

Имеются два критерия оптимальности полученного классификатора. Первый, математический, состоит в том, чтобы классификация каждого эксперимента, производимая обернутым классификатором с точностью до перестановки номеров классов наилучшим образом соответствовала разбиению данного эксперимента, производимому наивным учителем. Выполнение этого критерия является результатом обучения нейронной сети, и отклонения от такой оптимальной классификации используются при обучении как функция, которую мы хотим минимизировать, обучив нейронную сеть (так называемая функция потерь, выражение (4)). Качество результата оценивается численно метрикой качества (так называемая внутренняя оценка качества AURPC), которая приведена в работе ниже.

Второй критерий, физический, состоит в том, чтобы получаемые в результате классификации данные каждого отдельного класса можно было интерпретировать с физической точки зрения как сигналы, полученные в результате конкретного механизма рассеяния. Выполнение этого критерия проверяется статистическим анализом данных экспертом после их классификации обученным классификатором (так называемая внешняя оценка качества). Оптимизация этого критерия заключается в подборе структуры (архитектуры) нейронной сети и входных параметров до достижения необходимого качества интерпретации. Пример такого экспертного анализа в аналогичной задаче приведен в [Berngardt et al., 2022].

В качестве наивного учителя выбран вероятностный метод Gaussian Mixture [Vander Plas, 2016], представляющий данные как случайные, подчиняющиеся распределению вероятностей, равному линейной комбинации многомерных нормальных распределений. Их параметры (вес, среднее и матрица ковариации) определяются автоматически в процессе анализа данных, а количество этих нормальных распределений фиксировано и задается исследователем. Выбор количества классов обусловлен следующим. Модель Gaussian Mixture характеризуется тем, что разделяет данные именно на то число классов, которое задано N. Схема построения обернутого классификатора только ограничивает число скрытых классов K сверху. В случае, если это возможно, реальное количество скрытых классов с ненулевым количеством элементов после обучения может стать меньше, т. е. алгоритм способен к объединению близких классов, если это не ухудшает точность аппроксимации. Таким образом, в качестве N и K удобно выбрать достаточно большое число, заранее превышающее ожидаемое количество различных типов сигналов. Традиционно в радарных исследованиях различают порядка десятка таких качественно различимых по дальности или спектральным характеристикам сигнала типов: рассеяние от земной поверхности первого и второго скачков, метеорное рассеяние, рассеяние от магниториентированных неоднородностей E-слоя (двухпотоковых и градиентно-дрейфовых), рассеяние от магнитоориентированных неоднородностей F-слоя, рассеяние от спорадических слоев, мезосферное эхо и т. д. Поэтому количество классов N, K было выбрано вдвое превышающим ожидаемое количество различных классов, чтобы обернутый классификатор смог соответствующим образом уменьшить количество обнаруженных классов, если необходимо. С этим связана и простота используемого кластеризатора (Gaussian Mixture), имеющая целью загрубить и упростить кластеризацию в надежде, что статистический характер ее ошибок в каждом эксперименте позволит сформировать более устойчивые к таким ошибкам скрытые классы. Как показал дальнейший анализ, актуальное количество интерпретируемых скрытых классов со значимым количеством наблюдений действительно составляет от 15 до 18.

Функция $\overrightarrow{OHE}\left(y_{e,n}\right)$ имеет только одну ненулевую координату, соответствующую значению номера кластера $y_{e,\,n}$,

$$OHE_m\left(y_{e,\,n}\right) = \delta_{y_{e,\,n},m}. \quad (5)$$

Функция классификатора $\vec{g}()$ аппроксимируется трехслойной полносвязной нейронной сетью примерно с 30 тысячами свободных параметров, 133 нейронами и функциями активации *ReLU* на каждом скрытом слое, и функцией активации *SoftMax* на выходном слое. Благодаря этому функция нормализована так, что

$$\sum_{i=1}^{K} g_i = 1, \qquad g_i \geq 0. \quad (6)$$

Это позволяет интерпретировать выход функции $\vec{g}()$ как вероятности того, что $\vec{x}_{e,\,m}$ принадлежит к скрытым классам $\{1..K\}$, определяемым координатами g_i. Для улучшения качества обучения слои отделены друг от друга слоями batch-нормализации. Архитектура сети приведена на рис. 2, *б*.

Как показало тестирование, трехслойной сети $\vec{g}()$ достаточно для качественного решения задачи, более и менее глубокие сети качества не улучшают.

Для улучшения качества работы классификатора использовалось увеличение размерности входных данных методом спрямляющего пространства. Эффективность его использования в аналогичной задаче показана в [Berngardt et al., 2022]. Предполагается, что в этом варианте его использование также будет эффективным. Оно было выбрано в виде полиномиального преобразования $\overrightarrow{PF}\left(\vec{x}\right)$

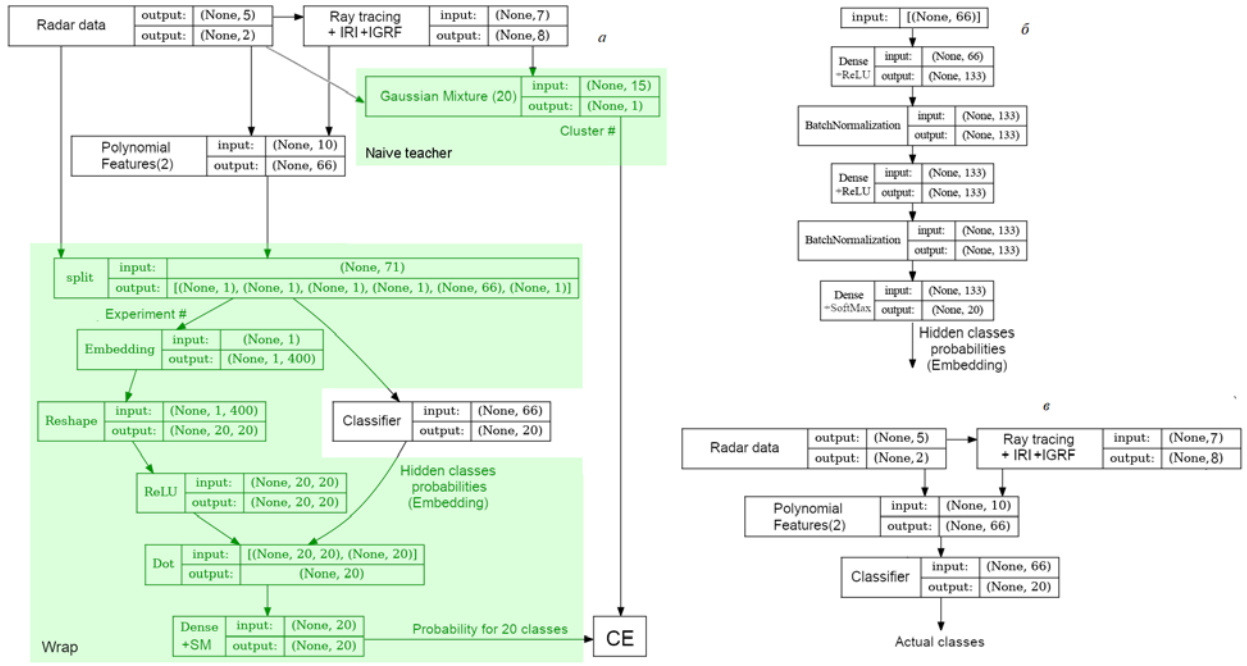

Рис. 2. Архитектура обернутого классификатора с наивным учителем и размерность векторов: *а* — архитектура сети в режиме обучения; *б* — архитектура классификатора; *в* — архитектура сети в режиме обработки. Метки None соответствуют числу записей в наборе данных

$$\overrightarrow{PF}\left(\vec{x}\right)=\left(1,\, x_0,\, x_1,\, ...,\, x_{N-1},\, x_0^2,\, x_0x_1,\right.$$
$$\left. x_0x_2,\, ...,\, x_{N-2}x_{N-1},\, x_{N-1}^2\right). \quad (7)$$

Это преобразование увеличивает исходную размерность N входного вектора $\vec{x}_{e,\,m}$ до размерности $N+N(N+1)/2+1$ и упрощает тем самым решение задачи классификации данных. Использованию этого преобразования способствовало то, что взвешенная сумма квадратов доплеровского сдвига и ширины спектра сигнала уже достаточно широко используется в качестве хорошего критерия идентификации сигналов рассеяния от земной поверхности [Blanchard et al., 2009]. Таким образом, учет квадратов входных признаков и перекрестных произведений входных признаков может повышать эффективность классификации. Характерная особенность использования спрямляющего пространства *PF* на входе нейронной сети заключается в том, что нейронная сеть в процессе обучения самостоятельно отсеет несущественные для оптимального решения координаты, оставив только существенные. Увеличение размерности пространства позволит использовать более простые сети для классификации, однако за счет увеличения количества свободных коэффициентов нейронной сети и увеличения размеров обучающего датасета, необходимого для ее обучения. Эффективность использования полиномиального спрямляющего пространства в этой задаче была продемонстрирована в [Berngardt et al., 2022] при рассмотрении предыдущего варианта обернутого классификатора. Неиспользование *PF* в кластеризаторе имело целью загрубить кластеризатор и внести в него больше ошибок в каждом отдельном эксперименте, чтобы скрытые классы классификатора сделать более устойчивыми к ошибкам за счет статистического характера обучения.

Большое количество неизвестных значений $C_{l,\,e}$, $b_{k,\,l,\,e}$ в (3) позволяет оптимально аппроксимировать произвольную перестановку $y_{e,n}$ по координатам g_i функции $\vec{g}()$ для каждого эксперимента e отдельно, так что $\vec{g}$ не зависит от e и, как и ожидалось, определяется исключительно формой кластеров сигналов в N-мерном пространстве их параметров.

Основным отличием данной модели от предыдущего варианта, описанного в [Berngardt et al., 2022], является форма модели $\vec{f}_e\left(\vec{g}\right)$ и требование $C_{l,\,e}\geq 0$ в (3). Это позволяет более уверенно интерпретировать функцию $\vec{g}$ как вероятности скрытых классов, из которых можно получить вероятности кластеров $\vec{f}_e$ простым линейным преобразованием с неотрицательными коэффициентами. Кроме того, такая модификация сделала возможным существенное упрощение модели функции $\vec{g}$, получение более качественных результатов разделения и проведение более качественного ее обучения за счет меньшего количества свободных коэффициентов в модели (30 тыс. вместо 80 тыс. свободных параметров).

Таким образом, предложенная схема классификации состоит из двух последовательных нейронных сетей (рис. 2, *а*), одна из которых (классификатор) создает оптимальное представление данных в виде 20-мерного вектора, интерпретируемое нами позже как вероятности скрытых классов. Вторая сеть (обертка,Wrap) преобразует вероятности скрытых классов в номера кластеров набора данных, размеченного наивным учителем. Сеть-классификатор

и сеть-обертку совместно обучают для наилучшего соответствия их выхода разметке этого набора данных кластеризатором (наивным учителем).

Кластеризатор независимо кластеризует данные от эксперимента к эксперименту (ото дня ко дню, от радара к радару): одни и те же виды рассеяния в разных экспериментах могут иметь разные номера кластеров. Поэтому при интерпретации результатов кластеризации мы должны перенумеровывать их. Мы не можем полностью доверять этому учителю и быть уверенными в правильности его кластеризации. Поэтому мы называем такого учителя наивным. Основываясь на результатах его кластеризации, мы строим и обучаем оптимальный классификатор, который проводит оптимальную классификацию на основе физики распространения радиоволн и обладает более широкими возможностями обобщения и интерпретации, чем наивный учитель.

Количество нейронов в каждом слое классификатора (133) было выбрано в соответствии с теоремой Колмогорова — Арнольда [Arnold, 1963] для наиболее оптимального представления входных данных. Классификатор имеет около 30 000 обученных параметров, а обертка — около 153 000, что существенно меньше, чем в предыдущем варианте классификатора [Berngardt et al., 2022].

Для ускорения обучения сети каждый этап проводился последовательно с сохранением полученных наборов данных в хранилище. Для интерпретации новых точек данных требуется только обученный классификатор и модель распространения радиоволн. Сеть-обертка и кластеризатор при прогнозе не используются (рис. 2, *в*).

Модель классификатора получает на вход десять входных параметров, только два из которых (доплеровский сдвиг и ширина спектра) были непосредственно измерены радаром, а остальные восемь получены с помощью численного моделирования на основе измеренных параметров принимаемого сигнала. Эти параметры показаны на рис. 1, более детально они обсуждались в [Berngardt et al., 2022].

Эта архитектура (классификатор + обертка) позволяет автоматически перенумеровать номера кластеров для каждого эксперимента независимо и повышает точность восстановления функции-классификатора $\vec{g}$. Во время обучения в качестве функции потерь используется взвешенная кросс-энтропия, где веса представляют собой обратное количество встреченного кластера в наборе обучающих данных. Это позволяет автоматически сбалансировать набор данных и тем самым повысить качество фитирования.

AURPC (AUC-RP) используется как метрика качества предсказания, так как работает достаточно корректно в случае возможного дисбаланса классов. Для обнаружения переобучения мы используем метод ранней остановки после более двадцати неудачных эпох обучения на основе значения метрики AURPC в наборе валидационных данных. Метод градиентного спуска используется с оптимизатором ADAM с размером пакета 32. Для обучения набор данных двух радаров за январь–сентябрь 2021 г. (~3 миллиона записей) был разбит на обучающий, валидационный и тестовый наборы данных в соотношении 64:16:20 %. Процесс обучения соответствует описанному в [Berngardt et al., 2022].

На рис. 3 показано разделение экспериментальных данных на классы на примере данных за апрель 2022 г. Номера классов нейронная сеть присвоила самостоятельно в процессе обучения, мы же в дальнейшем будем интерпретировать только классы с номерами 2 и 13.

На рис. 4 показано распределение количества обнаруженных сигналов в различных классах. Видно, что актуальное количество скрытых классов с ненулевым количеством объектов составляет 18, из которых два класса, 9 и 18 — плохо определенные данные (данные, распределенные примерно равномерно по диаграмме дальность — время, в дальнейшем нами обычно не анализируются) и несколько классов, очень редко встречающихся (1, 5 и 7, причем классы 1, 5 встречаются только на радаре EKB). Класс 15 представляет собой неинтерпретируемый с точки зрения распространения — средние высоты рассеяния h_{iri} превышают высоту максимума F2-слоя, что говорит о том, что в данном случае траектория распространения радиосигнала расчитана неверно: либо ионосфера в этих случаях не соответствует модельной (это возможно, учитывая существующие точности модели IRI), либо угол места прихода сигнала рассчитан неверно (это также возможно, учитывая сложности расчета угла места прихода сигнала). Таким образом, общее количество интерпретируемых классов составляет 14, а их суммарная доля в данных различных радаров составляет от 50 до 60 %, что говорит о том, что примерно половина регистрируемых данных радаров может быть автоматически интерпретирована с точки зрения модельного распространения радиосигналов в рамках предложенного метода. Это неплохо согласуется с качественными ожиданиями и результатами работ предыдущей версии алгоритма [Berngardt et al., 2022]. Из рис. 4 видно, что в данных присутствуют особенности, зависящие от радара. В частности классы 1, 5 наблюдаются только на EKB. Отмечается также явный дисбаланс в наблюдении одних и тех же классов на различных радарах (например, 8, 10, 16), который может быть вызван как техническими характеристиками радаров (различный уровень шумов и слегка различные коэффициенты усиления антенн), так и региональными особенностями ионосферы (сектор обзора радара MAGW сильнее наклонен от северного направления, чем у EKB).

ПЕРВЫЕ ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ И ИХ ИНТЕРПРЕТАЦИЯ

Дальнейшее исследование посвящено классу 13, интерпетируемому как метеорное рассеяние и классу 2, интерпретируемому как результат спорадического рассеяния в ионосфере. Детальный анализ остальных классов, определенных нейронной сетью, выходит за рамки данной работы. Для обоснования такой интерпретации классов, выделенных самостоятельно

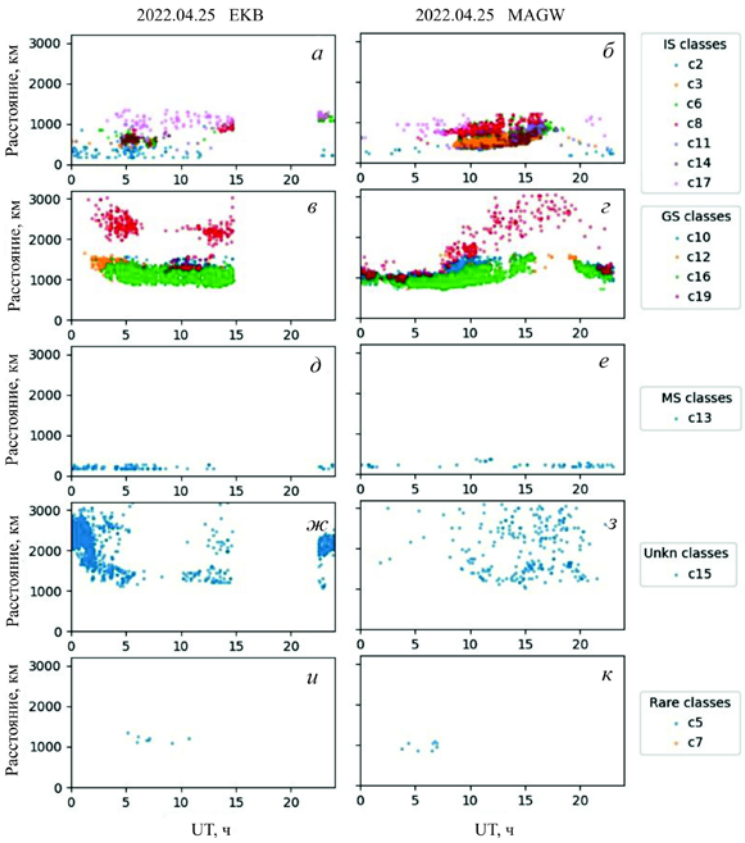

Рис. 3. Пример работы алгоритма на данных радаров EKB (слева) и MAGW ИСЗФ СО РАН (справа) 25 апреля 2022 г. Сверху вниз: классы типа ионосферных рассеяний (*а*, *б*); классы типа рассеяния от земной поверхности (*в*, *г*); классы типа метеорного рассеяния (*д*, *е*); неидентифицированные классы (*ж*, *з*); редко наблюдаемые классы (*и*, *к*)

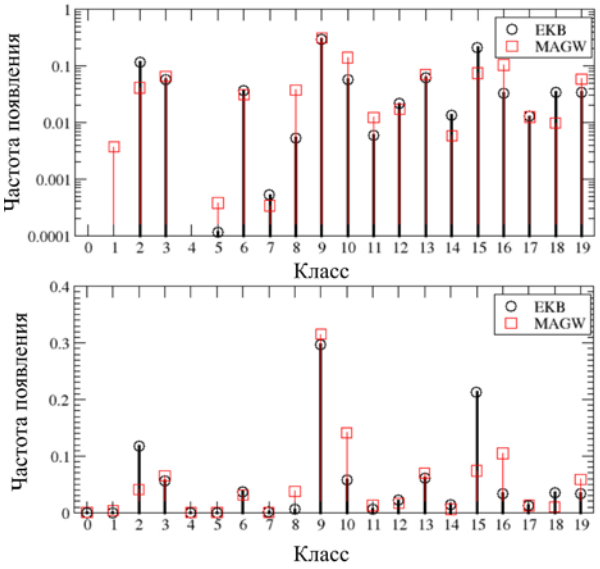

Рис. 4. Распределение частот появления наблюдаемых классов сигналов в данных радаров EKB и MAGW ИСЗФ СО РАН за период обучения в логарифмической (сверху) и линейной (снизу) шкалах

предложенной нейронной сетью, на рис. 5 показаны высотно-дальностные распределения принадлежащих к этим классам данных радаров EKB и MAGW за 2021–2022 гг.

Видно, что распределение высот появления класса 13 на обоих радарах соответствует диапазону высот 50–150 км с максимумом в районе 80–90 км, что хорошо соответствует рассеянию на метеорных следах [Chisham, Freeman, 2013; Федоров, Бернгардт, 2021]. Диапазон дальностей составляет от минимальной дальности радара 180 км до примерно 400 км с сильным спаданием числа появлений с дальностью, то также хорошо соответствует статистике наблюдений метеорного рассеяния, выделяемого из данных другим методом [Федоров, Бернгардт, 2021].

Высотно-дальностное распределение класса 2 отличается от распределения класса 13. Диапазон высот распределения расширился от 0 до 250 км, причем ближние дальности и малые высоты скорее всего соответствуют рассеянию от земной поверхности с высокими углами места (невозможно пока проверить из-за проблемы фазовой неопределенности интерферометрических наблюдений радаров такого типа [Chisham, Freeman, 2013]), что на таких малых дальностях соответствует рассеянию от нижних слоев ионосферы с высотами ниже 200 км. Это можно интерпретировать как рассеяние от спорадического слоя и последующее рассеяние от земной поверхности.

В пользу этого механизма говорит увеличение с дальностью эффективной высоты рассеяния на дальностях ниже 500 км, что может быть связано с недоучетом моделью ионосферы IRI спорадических слоев. Вторая часть рассеянных сигналов сосредоточена вблизи высот 100 км и наблюдается в основном на дальностях 300–500 км. Это говорит

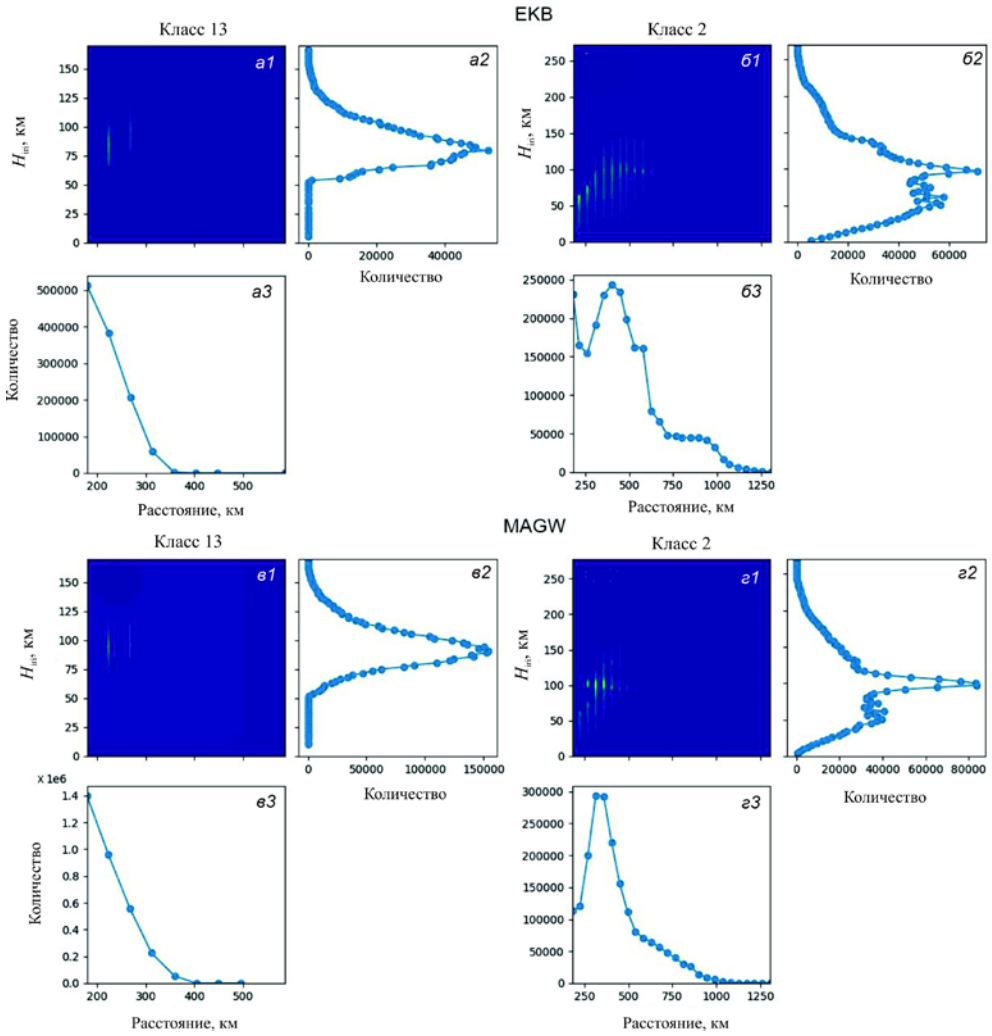

Рис. 5. Сравнение высотно-дальностных распределений классов 13 и 2 за период январь 2021 – март 2022 г.: *а1–г1* — высотно-дальностное распределение частоты наблюдения следов; *а2–г2* — частота появления рассеивателей в зависимости от высоты, рассчитанной по траектории распространения сигнала; *а3–г3* — распределение сигналов как функция радиолокационной дальности

о возможности интерпретации 2-го класса сигналов как смеси непосредственного рассеяния от спорадического слоя и рассеяния на земной поверхности после рассеяния на спорадическом слое.

В пользу интерпретации класса 2 как подходящего для диагностики спорадического слоя говорят и продольные скорости, близкие к скоростям нейтрального ветра, диагностируемые по доплеровскому смещению частоты принятого сигнала. Из рис. 3, *а*, *б*, *д*, *е* видно, что классы 13 и 2 очень фрагментарны (спорадичны) во времени и пространстве (отмечены синими цветами на рис. 3, *а*, *б*, *д*, *е*), поэтому для дальнейшего сравнительного статистического анализа удобно использовать их усредненные по 1 ч параметры.

На рис. 6 приведены примеры поведения продольной скорости (вверху) и количества рассеянных сигналов (внизу), усредненных за 1 ч. Черному цвету соответствует класс 13 (метеорное эхо), красному — класс 2 (спорадическое рассеяние). Видно, что эти два класса неплохо коррелируют по скоростям и периодам появления сигналов. С точки зрения механизмов формирования класс 2 можно интерпретировать как рассеяние от спорадических слоев, механизм образования которых в нижней части ионосферы является одним из противоречивых вопросов в исследовании нижней ионосферы и может быть связан с метеорами [Malhotra et al., 2008] и требует высокого временного и пространственного разрешения для уверенного изучения. Такое высокое временное разрешение (от единиц секунд до минуты) обеспечивается радарами SuperDARN и аналогичными им радарами EKB и MAGW ИСЗФ СО РАН. Статистическое подтверждение такой корреляции по полному набору данных 2021–2022 гг. показано на рис. 7. Коэффициенты линейной (Пирсона) и ранговой (Спирмена) корреляции приведены в таблице. Видно, что коэффициенты корреляции находятся в диапазоне 0.52–0.76, что подтверждает видимую на рис. 7 значимую положительную корреляцию между среднечасовыми значениями параметров сигналов в классах метеорного эха и спорадического эха. Во всех случаях рассчитанный уровень значимости отсутствия корреляции (p-value) менее 10^{-6} (в таблице не приводится), что говорит о значимости корреляции. Незначительное превышение корреляции Спирмена над корре-

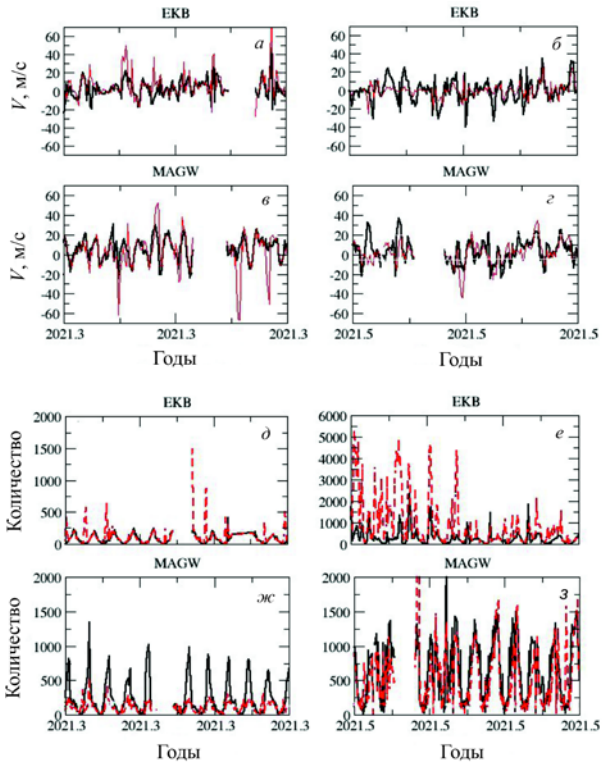

Рис. 6. Пример продольных скоростей (*а–г*), измеренных на радарах EKB и MAGW, и количества обнаруженных сигналов (*д–з*) весной (*а*, *в*, *д*, *ж*) и летом (*б*, *г*, *е*, *з*) 2021 г. для классов 13 (черная линия) и 2 (красная линия)

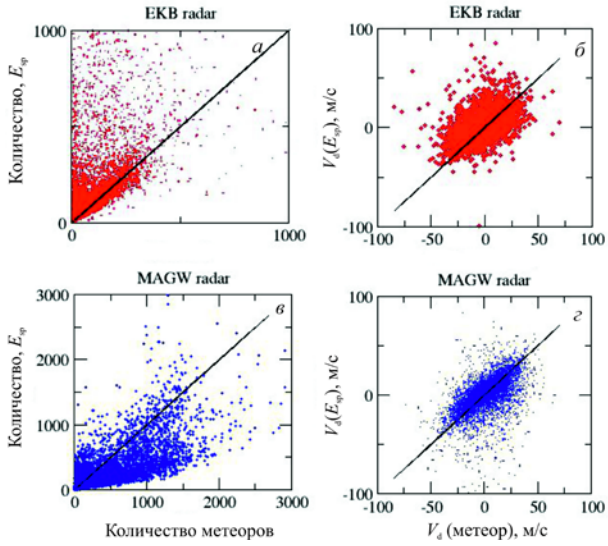

Рис. 7. Взаимосвязь числа рассеянных сигналов (*а*, *в*) и среднечасовых скоростей (*б*, *г*) в классах 13 (метеорное эхо) и 2 (спорадическое рассеяние) по данным радаров EKB (*а*, *б*) и MAGW (*в*, *г*)

ляцией Пирсона (порядка 10 %) говорит о наличии не только сильной линейной связи, но, возможно, и слабой нелинейной связи. Различный наклон на панелях рис. 7, А, С для числа наблюдений спорадического эха на радарах может быть связан как с особенностями геометрии зондирования (сектор обзора радара MAGW сильнее наклонен к западу, чем сектор обзора радара EKB), так и региональными особенностями ионосферы и требует дополнительного анализа. Изложенная в работе методика классификации сигналов, принимаемых радарами и проведенный в работе предварительный анализ дают возможность в будущем использовать сравнительный анализ этих двух классов для диагностики нейтрально-ионосферного взаимодействия в нижней части ионосферы.

Коэффициенты корреляции Пирсона и Спирмена между метеорным эхом и спорадическим эхом по данным радаров EKB и MAGW

Коэффициент корреляции	EKB	MAGW
Пирсона по количеству наблюдений	0.636	0.718
Спирмена по количеству наблюдений	0.716	0.761
Пирсона по доплеровской скорости	0.527	0.564
Спирмена по доплеровской скорости	0.591	0.652

ЗАКЛЮЧЕНИЕ

В работе описана текущая версия алгоритма автоматической классификации сигналов (v.1.1), принимаемых радарами декаметрового когерентного рассеяния ИСЗФ СО РАН. Алгоритм, названный нами обернутым классификатором с наивным учителем, представляет собой самообучающуюся нейронную сеть, определяющую тип рассеянных сигналов по результатам зондирования и численного моделирования распространения радиоволн. Распространение радиосигналов рассчитывается методом геометрической оптики (трассировки лучей) с использованием радарных данных и международных ссылочных моделей ионосферы (IRI) и магнитного поля Земли (IGRF). Модель обучается самостоятельно, используя результаты предварительной классификации данных наивным учителем. Алгоритм наивного учителя представляет собой кластеризацию данных статистической моделью смеси многомерных нормальных распределений. Он разбивает данные на кластеры, которые потом используются для поиска скрытых классов в данных по тем физическим параметрам (полученным по данным радара, а также в результате физического моделирования распространения радиоволн), которые впоследствии можно эффективно интерпретировать. Самостоятельное обучение нейронной сети проводилось на данных радаров MAGW и EKB ИСЗФ СО РАН за 2021 г.

Результирующие скрытые классы, найденные классификатором, могут быть интерпретированы с физической точки зрения путем статистического анализа распределения физически интерпретируемых параметров сигналов, принадлежащих каждому классу. Данный классификатор является модернизацией классификатора, предложенного ранее [Berngardt et al., 2022], и обладает лучшим качеством разбиения на классы, а также более простой архитектурой классификатора с меньшим числом свободных параметров, найденных в результате обучения.

Для демонстрации работы алгоритма классификации представлен первый статистический анализ наблюдений сигналов, отнесенных алгоритмом к классам 13 и 2, интерпретируемых нами как рассеяние на метеорных следах и рассеяние с участием спорадического слоя E соответственно. На основе полного анализа статистических данных на радарах EKB и MAGW за 2021–2022 гг. представлены дальност-

но-высотные характеристики сигналов этих типов. Показана корреляция между среднечасовыми количествами наблюдений обоих классов, а также между среднечасовыми продольными скоростями, получаемыми в обоих классах данных. Предполагается, что результаты позволят использовать радарные данные для изучения процессов взаимодействия нейтральной атмосферы (изучаемой по данным метеорного рассеяния) и нижней ионосферы (изучаемой по наблюдениям за спорадическим слоем E) в пространственно близких точках с высоким временным разрешением. В настоящее время алгоритм классификации данных работает на радарах ИСЗФ СО РАН в автоматическом режиме [http://sdrus.iszf.irk.ru/node/95], код обученной нейронной сети классификатора доступен по адресу [https://github.com/berng/WrappedClassifier].

Радар EKB ИСЗФ СО РАН входит в центр коллективного пользования научной аппаратурой «Ангара» [http://ckp-rf.ru/ckp/3056]. Эксплуатация радаров осуществлялась при финансовой поддержке Министерства науки и высшего образования Российской Федерации (субсидия № 075-ГЗ/С3569/278). Данные радаров EKB и MAGW ИСЗФ СО РАН доступны на сайте [http://sdrus.iszf.irk.ru/ekb/page_example/simple]. Обучение модели выполнено частично на оборудовании Центра коллективного доступа «Биоинформатика» Федерального исследовательского центра «Институт цитологии и генетики СО РАН» (ИЦГ СО РАН). Автор благодарен И.С. Петрушину (Иркутский государственный университет) за полезные обсуждения. Работа выполнена при финансовой поддержке совместного гранта РФФИ-CNRS № 21-55-15012.

СПИСОК ЛИТЕРАТУРЫ

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.

Федоров Р.Р., Бернгардт О.И. Мониторинговые наблюдения метеорного эха на радаре EKB ИСЗФ СО РАН: алгоритмы, валидация, статистика. *Солнечно-земная физика*. 2021. Т. 7, № 1. С. 59–73. DOI: 10.12737/szf-71202107.

Arnold V.I. On functions of three variables. *American Mathematical Society Translations. Ser. 2*. 1963. Vol. 28. P. 51–54. (*Translation of Dokl. Akad. Nauk SSSR*. 1957. Vol. 114, iss. 4. P. 679–681).

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, iss. 10. P. 2905–2919. DOI: 10.1016/j.asr.2022.07.054. (In print).

Bilitza D., Altadill D., Zhang Y., et al. The International Reference Ionosphere 2012 — a model of international collaboration. *J. Space Weather Space Climate*. 2014. Vol. 4, id. A07. 12 p. DOI: 10.1051/swsc/2014004.

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, iss. 5. RS5012. DOI: 10.1029/2009rs004141.

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. P. 207–221. DOI: 10.1016/j.jastp.2013.05.018.

Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society: Ser. B (Methodological)*. 1977. Vol. 39, no. 1. P. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Ginzburg V.L. *The Propagation of Electromagnetic Waves in Plasmas*. Pergamon Press, 1970. 615 p.

Goodfellow I., Bengio Y., Courville A. *Deep Learning (Adaptive Computation and Machine Learning Ser)*. MIT Press, 2016. 800 p.

Lavygin I.A., Berngardt O.I., Lebedev V.P., Grkovich K.V. Identifying ground scatter and ionospheric scatter signals by using their fine structure at Ekaterinburg decametre coherent radar. *IET Radar, Sonar and Navigation*. 2020. Vol. 14, iss. 1. P. 167–176. DOI: 10.1049/iet-rsn.2019.0192.

Malhotra A., Mathews J.D., Urbina J. Effect of meteor ionization on sporadic-E observed at Jicamarca. *Geophys. Res. Lett*. 2008. Vol. 35, iss. 15. DOI: 10.1029/2008GL034661.

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 Sci*. 2019. Vol. 6, iss. 1. P. 27. DOI: 10.1186/s40645-019-0270-5.

Ribeiro A.J., Ruohoniemi J.M., Baker J.B.H., et al. A new approach for identifying ionospheric backscatter in midlatitude SuperDARN HF radar observations. *Radio Sci*. 2011. Vol. 46, iss. 4. RS4011. 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, iss. 3. P. 274–282. DOI: 1002/rds.20031.

Siwei Yu., Ma J. Deep learning for geophysics: current and future trends. *Rev. Geophys*. 2021. Vol. 59, iss. 3. e2021RG000742. DOI: 10.1029/2021rg000742.

Thébault E., Finlay C.C., Beggan C.D., et al. International Geomagnetic Reference Field: the 12th generation. *Earth, Planets and Space*. 2015. Vol. 67, iss. 1. P. 79. DOI: 10.1186/s40623-015-0228-9.

Vander Plas J. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 2016. 548 p.

URL: http://sdrus.iszf.irk.ru/node/95 (дата обращения 12 октября 2022 г.).

URL: https://github.com/berng/WrappedClassifier (дата обращения 12 октября 2022 г.).

URL: http://ckp-rf.ru/ckp/3056 (дата обращения 12 октября 2022 г.).

URL: http://sdrus.iszf.irk.ru/ekb/page_example/simple (дата обращения 12 октября 2022 г.).

Как цитировать эту статью:

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.