Повышение точности распознавания молний на основе данных доплеровского радара с использованием методов машинного обучения

Введение

Молния представляет собой серьезную природную опасность, которая представляет значительный риск для жизни людей и инфраструктуры. Поэтому точное прогнозирование грозовых явлений имеет первостепенное значение для снижения этих рисков. В этом исследовании мы изучили возможность использования алгоритмов машинного обучения для повышения точности прогнозирования молний на основе данных доплеровского радара. Мы собрали и предварительно обработали набор радиолокационных и грозопеленгационных данных и применили к этим данным различные методы машинного обучения. Наши результаты показали, что алгоритмы машинного обучения могут существенно повысить точность классификаций молний. Было обнаружено, что методы глубокого обучения особенно хорошо работают в этом контексте. Эти выводы имеют важное значение для повышения безопасности и надежности систем прогнозирования погоды и обеспечивают основу для дальнейших исследований в этой области.

Данные

Данные, использованные в этом исследовании, состояли из радиолокационных и грозопеленгационных измерений, собранных с доплеровского метеорологического радара РГГМУ в течение двух лет. Данные радара включали доплеровскую отражательную способность, которая дает информацию о структуре и движении облачности.

Данные грозопеленгации, были получены с сети Blitzortung. Blitzortung — это сеть управляемых добровольцами станций обнаружения молний, ​​которые используют метод времени прибытия (TOA) для обнаружения и локализации ударов молнии. Данные, собранные станциями Blitzortung, включают информацию о местоположении, времени и типе ударов молнии, а также об амплитуде и полярности сигналов. Сеть Blitzortung состоит из нескольких сотен станций обнаружения, расположенных по всему миру, и данные с этих станций собираются, обрабатываются и публикуются практически в режиме реального времени через веб-сайт Blitzortung. Данные обычно представляются в виде карт, показывающих место и время ударов молнии, а также графиков, показывающих количество ударов с течением времени.

Одной из основных особенностей данных Blitzortung является их высокое временное разрешение, что позволяет обнаруживать удары молнии почти в реальном времени. Эти данные также обеспечивают глобальный охват и могут использоваться для изучения пространственного и временного распределения грозовой активности. Кроме того, данные Blitzortung можно использовать в сочетании с другими метеорологическими данными для изучения взаимосвязи между молниями и погодными системами.

Чтобы подготовить данные для анализа, мы выровняли радиолокационные и грозопеленгационные данные по времени и местоположению, чтобы обеспечить синхронизацию двух наборов данных. Мы также включили в наш анализ дополнительные данные о высоте изотерм 0°C и -22°C, а также высоту тропопаузы. Затем мы пометили ячейки в данных радара в зависимости от наличия или отсутствия ударов молнии, создав набор данных бинарной классификации. Наконец, мы удалили все ложные точки данных, которые были идентифицированы как потенциальные источники ошибок.

В целом, набор данных, использованный в этом исследовании, содержал в общей сложности 6690 радиолокационных измерений, охватывающих 118815 различных грозовых явлений. Данные были разделены на обучающую и тестовую выборки в соотношении 80/20, при этом обучающая выборка использовалась для подбора моделей машинного обучения, а тестовая — для оценки их производительности.

Методы

В этом исследовании мы оценили производительность нескольких различных алгоритмов машинного обучения для прогнозирования молний на основе данных доплеровского радара. Эти алгоритмы включали логистическую регрессию, классификатор SGD, нейронную сеть и случайный лес. Мы обучили и протестировали эти модели, используя тщательно подобранный набор данных радиолокационных измерений, и сравнили их производительность с точки зрения точности, воспроизводимости и полноты.

Логистическая регрессия — это алгоритм линейной классификации, который обычно используется для прогнозирования бинарных результатов (например, да/нет, 0/1). Он работает путем подгонки линейной модели к входным данным и использования логистической функции для сопоставления прогнозируемого результата с вероятностью от 0 до 1. Логистическую регрессию легко и быстро реализовать, но она может столкнуться с нелинейными отношениями и высокой точностью. размерные данные.

Классификатор SGD — это линейный классификатор, который использует алгоритм оптимизации стохастического градиентного спуска (SGD) для изучения параметров модели. Он часто используется для крупномасштабных задач классификации из-за его эффективности и масштабируемости. Однако он чувствителен к скорости обучения и может быть подвержен переобучению, если данные зашумлены или имеют высокую степень дисперсии.

Нейронная сеть — это модель машинного обучения, состоящая из нескольких взаимосвязанных слоев искусственных нейронов или «узлов». Его можно научить распознавать закономерности и делать прогнозы на основе входных данных. Функция потерь “Focal loss” — это вариант функции кросс-энтропийных потерь, предназначенный для решения проблемы дисбаланса классов в задачах обнаружения объектов. Это уменьшает потери для хорошо классифицированных примеров и увеличивает потери для плохо классифицированных примеров, помогая модели уделять больше внимания сложным примерам. Нейронные сети являются мощными, но требуют значительных объемов данных и вычислительных ресурсов для обучения и могут быть склонны к переобучению, если не будут должным образом упорядочены.

Случайный лес — это ансамблевый метод обучения, состоящий из нескольких деревьев решений, обученных на разных подмножествах данных. Он делает прогнозы, объединяя прогнозы всех деревьев решений, что помогает уменьшить переоснащение и улучшить обобщение. Случайный лес — это универсальный и надежный алгоритм, который может обрабатывать многомерные данные и нелинейные отношения, но его обучение может быть медленным, и для достижения хорошей производительности может потребоваться обширная настройка гиперпараметров. Он также чувствителен к наличию выбросов или зашумленных данных и может быть менее интерпретируемым, чем более простые модели, из-за большого количества используемых деревьев решений.

Скользящее окно — это метод, который включает в себя разделение набора данных на перекрывающиеся сегменты или «окна» фиксированного размера и формы. Затем окна систематически перемещаются по набору данных, что позволяет анализировать и обрабатывать данные в каждом окне. Скользящее окно часто используется для извлечения признаков или шаблонов из временных рядов или последовательных данных или для анализа данных в разных масштабах или контекстах.

В контексте машинного обучения скользящее окно можно использовать для оценки производительности различных вариантов модели путем обучения и тестирования модели на разных подмножествах данных. Например, в исследовании прогнозирования молний на основе данных доплеровского радара можно использовать скользящее окно для анализа данных различной формы (например, 3x3x11, 5х5х11, 7х7х11), где первые две цифры говорят о горизонтальном размере окна, а 11 – количество уровней отражаемости по вертикали. Это позволит исследователям выявить взаимосвязи и тенденции в данных, которые не очевидны только из необработанных данных, и определить, какой размер и форма окна наиболее эффективны для классификции грозовых явлений.

Использование скользящего окна для оценки производительности моделей машинного обучения имеет несколько преимуществ. Одним из преимуществ является то, что он позволяет исследователю анализировать данные в разных масштабах и контекстах, что может обеспечить более тонкое понимание данных и содержащихся в них взаимосвязей. Еще одно преимущество заключается в том, что он может помочь определить шаблоны или функции, которые важны для создания точных прогнозов, которые могут помочь в разработке будущих моделей машинного обучения.