САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ - ПРОЦЕССОВ УПРАВЛЕНИЯ

КАФЕДРА КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ И
МНОГОПРОЦЕССОРНЫХ СИСТЕМ

Балакший Андрей Владимирович

Выпускная квалификационная работа бакалавра

Использование методов машинного обучения для прогнозирования опасных конвективных явлений с помощью численной модели конвективного облака

Направление 010400

Прикладная математика и информатика

Заведующий кафедрой,
доктор физ.-мат. наук,
профессор Андрианов С. Н.

Научный руководитель,
кандидат физ.-мат. наук,
доцент Станкова Е. Н.

Рецензент,
ст. науч. сотрудник,
кандидат физ.-мат. наук,
Довгалюк Ю. А.

Санкт-Петербург
2016

Содержание

Введение

На сегодняшний день одним из наиболее актуальных и приоритетных направлений в науке является решение таких практически значимых задач, как предсказание опасных конвективных явлений. Ведь такие явления как град, шквал или гроза оказывают значительное влияние на жизнь людей, не говоря уже об их роли в причинении разрушений в огромных масштабах. Одним из ключевых факторов возникновения таких явлений и формирования погоды в целом являются конвективные облака. В наши дни их изучение производится по трем основным направлениям: исследования в лабораториях, натурные эксперименты и численное моделирование. В силу ряда объективных причин, таких как сложность проведения контрольных экспериментов, их труднодоступности и дороговизны используемых для этого приборов, наиболее эффективным и распространенным методом изучения облаков является их численное моделирование.

В данной работе используется полуторамерная нестационарная модель конвективного облака с подробным описанием микрофизических процессов для расчета параметров облака, которые могут быть использованы в дальнейшем для прогнозирования опасных конвективных явлений. Представлен ранее неиспользуемый подход для автоматической классификации результатов радиозондирования атмосферы (далее данные радиозондирования или зондировки). Прогноз осуществляется по отобранным численным параметрам смоделированного облака с помощью методов машинного обучения. Машинное обучение — математическая дисциплина, позволяющая посредством использования различных разделов теории вероятностей, математической статистики и численных методов, получать знания из имеющихся данных. Она используется для автоматизации решения различных задач в самых разных областях человеческой деятельности. В наши дни в результате повсеместной информатизация накоплены внушительные объёмы данных во всевозможных отраслях таких как производство, наука, бизнес, здравоохранение. В данной работе машинное обучение используется для автоматизации нахождения решения задачи классификации зондировок, в результате чего данная численная модель может быть использована для оперативного прогноза

опасного конвективного явления в различных метеоцентрах. Представлены результаты применения различных методов машинного обучения для классификации зондировок посредством обучения с учителем.

Оказалось, что вплоть до настоящего времени прогноз опасных явлений, связанных с развитием конвекции (гроза, град, шквал) осуществляется с помощью полуэмпирических методов Пескова, Ягудина, Решетова, Лебедевой и др. [1]. Они основаны на расчете комплексных коэффициентов, которые являются функциями некоторых параметров облака, таких как, например, температура и высота верхней границы облака, значение температуры на уровне определенной изобары и др. Такого рода параметры определяются либо с помощью синоптической карты, либо с помощью аэрологической диаграммы, по которой можно определить возможное развитие облака, начиная от уровня конденсации. В данной работе параметры облака рассчитываются с помощью численной модели, используя в качестве начальных и граничных условий данные радиолокационного зондирования атмосферы (вертикальные распределения температуры и влажности).

Постановка задачи

Разработать на основе методов машинного обучения и реализовать в программном коде алгоритм прогнозирования опасных конвективных явлений с помощью полуторамерной нестационарной модели конвективного облака. Осуществить выбор алгоритма определения значимых для прогноза параметров облака. Оценить качество используемых методов машинного обучения.

Для достижения поставленных целей были сформулированы следующие задачи:

1. Выбрать тип используемой далее численной модели конвективного облака и её реализацию.
2. Провести валидацию выбранной модели по данным натурного эксперимента.
3. Провести автоматизацию и настройку модели.
4. Создать обучающую выборку данных, используя информацию о наблюдаемых конвективных явлениях и соответствующие данные радиолокационного зондирования атмосферы.
5. Провести выделение значимых признаков (параметров облака, значимых для реализации прогноза).
6. Получить конкретный вид решающей функции, используя несколько алгоритмов машинного обучения и оценить точность прогноза.
7. Представить конкретный алгоритм прогнозирования опасного конвективного явления.

Обзор литературы

В процессе написания данной выпускной квалификационной работы были использованы различные источники: научная и учебная литература, разные статьи, касающиеся как информационной составляющей данной работы, так и основных понятий физики облака.

”Руководство по прогнозированию метеорологических условий для авиации” под редакцией К.Г. Абрамовича, А.А. Васильева позволило ознакомиться с существующими методами прогноза опасных конвективных явлений. Оно же вдохновило меня поставить перед собой целью разработать новый альтернативный метод предсказания, использующий современные передовые методы науки.

Основной литературой по моделированию атмосферных процессов стали книги и статьи авторов: Ампиловой Н.Б., Веремеева Н.Е., Довгалюк Ю.А., Раба Н.О., Синькевич А.А., Станковой Е.Н. Используя указанные источники я получил всю необходимую информацию о существующих численных моделях конвективного облака и их особенностях. Особенно хочу отметить статью Морозова В.Н., Веремея Н.Е., Довгалюк Ю.А. ”Моделирование процессов электризации в трехмерной численной модели осадкообразующего конвективного облака”, которая помогла мне понять, что используя модели высокой размерности удается получить результаты очень хорошей точности, однако за это приходится платить высокими требованиями к мощности вычислительной техники.

Статьи Станковой Е.Н., Петрова Д.А. касаемо комплексной системы формирования входных данных для численных моделей облаков помогли мне уменьшить время, которое ушло на то, чтобы разобраться в этой крайне полезной системе.

Книга Матвеева Л. ”Курс общей метеорологии. Физика атмосферы” позволила мне поднять свою компетенцию в вопросах основных понятий физики облаков, без которой эта работа не смогла бы быть законченной.

Однако если по вопросам облаков не составило особых проблем найти русскоязычные источники, то с машинным обучением и связанными с ним технологиями литература на английском языке либо существенно выше ка-

чеством, либо вообще не имеет русскоязычных аналогов. Например, для изучения применения методов машинного обучения было проанализировано издание следующих авторов: T. Hastie, R. Tibshirani, J. Friedman ”The elements of Statisticals Learning”. В нем описываются не только все основные методы машинного обучения, но и даются лучшие практики их использования. Аналогов на русском языке данная книга не имеет, также как и перевода.

Таким образом, в указанной литературе я нашел всю необходимую мне теоретическую информацию. При этом хотя литературные издания по практической части в данной области присутствуют, однако там освещены далеко не все необходимые для меня вопросы.

Глава 1. Выбор численной модели конвективного облака

Одним из наиболее эффективных и распространенных способов изучения конвективных облаков является их численное моделирование, позволяющее не прибегать к дорогостоящим натурным экспериментам. Существуют различные численные модели конвективного облака, отличающиеся между собой размерностью и степенью детализации описания микрофизических процессов.

По степени детализации описания микрофизических процессов численные модели делятся на две большие группы: модели с параметризованной и с детальной микрофизикой. В моделях с параметрической микрофизикой рассматривается пространственно-временная эволюция интегральных параметров облачных частиц, таких как удельное содержание капель, ледяных кристаллов, частиц крупы и града. Относительно небольшое число интегральных параметров и уравнений делают такую модели весьма привлекательной с точки зрения эффективности вычислений.

В моделях же с детальной микрофизикой эволюция частиц описывается с помощью функции распределения по массам или размерам. При этом форма спектров всех частиц рассчитывается путем численного решения системы кинетических уравнений. Это позволяет проводить более детальный анализ микрофизических процессов.

По размерности бывают одномерные, полуторамерные, двумерные и трехмерные представления. Понятно, что для научных исследований больше всего подходят двухмерные или трехмерные представления, так как они наиболее полно отражают все происходящие в облаке процессы [2, 3]. Но такие представления весьма сложны и требуют огромных вычислительных ресурсов.

Однако не всегда есть возможность использовать суперкомпьютеры для того, чтобы сделать оперативный прогноз опасного конвективного явления. Поэтому необходимы представления, которые не требуют таких больших вычислительных ресурсов, и при этом могут достаточно точно воспроизводить основные характеристики облака. Чаще всего это достигается путем умень-

шения размерности пространства. Здесь как нельзя лучше подходят полутаромерные модели, способные давать достоверные прогнозы на краткосрочный период времени. В их основе лежит модель Шиино [4], в которой область разбивается на слои по высоте. Значения всех параметров усредняются в каждом слое. В отличии от одномерных представлений, в полутаромерных представлениях учитывается также радиальная составляющая скорости, направленная внутрь цилиндра или во внешнюю среду. Благодаря этому присутствует взаимодействие с внешней средой через боковую поверхность цилиндра.

Однако в классической модели Шиино не происходит стадия диссипации, облако проходит стадию развития и стабилизируется. Такое поведение описано в [5]. Это проблема вызвана отсутствием нисходящего потока. В данной работе используется разработанная Н.О. Раба и Е.Н. Станковой полуторамерная нестационарная модель конвективного облака с подробным описанием микрофизических процессов [6–8], лишенная данного недостатка, так как там присутствует нисходящий поток.

Глава 2. Валидация модели

Для использования численной модели облака при прогнозировании необходимо было провести её валидацию по данным натурного эксперимента. Были использованы данные радиозондирования атмосферы, полученные в Miles City, штат Монтана, 19 июля 1981г [9]. Тогда из-за малого сдвига ветра ситуация благоприятствовала образованию одноячейковых кучевых облаков. Для сопоставления облачных характеристик, полученных при численном моделировании и в эксперименте, была проведена синхронизация сравниваемых характеристик по времени, когда верхняя граница облака достигла высоты, равной 7.2 км. В натурном эксперименте это было время 16.18 по горному дневному времени.

2.1. Настройка диаметров цилиндров

Пусть a и b соответственно радиусы внутреннего и внешнего цилиндров модели.

Случай 1: $a = 3000, b = 10000.$

Данные радиусы были выбраны авторами используемой модели как значения при моделировании по умолчанию. В таблице 1 и на рисунке 1 приведены зависимость значения верхней грани облака от времени при численном моделировании и в эксперименте.

t эксп.	16.18	16.20	16.23	16.25	16.28	16.30	16.32	16.35	16.55
t мод мин	37.2	39.2	42.2	44.2	47.2	49.2	51.2	54.2	64.2
эксп.Км	7.15	7.7	8.3	9.3	10.0	10.4	10.6	10.6	10.6
мод.Км	7.2	7.85	9.1	9.9	10.6	10.7	10.5	10.3	9.9

Таблица 1: Значения высоты верхней грани облака в моделировании и в эксперименте

Из таблицы 1 видно довольно таки схожее поведение облака по модельным и экспериментальным данным, но высота верхней грани по данным моделирования немного превышает высоту в эксперименте, и в

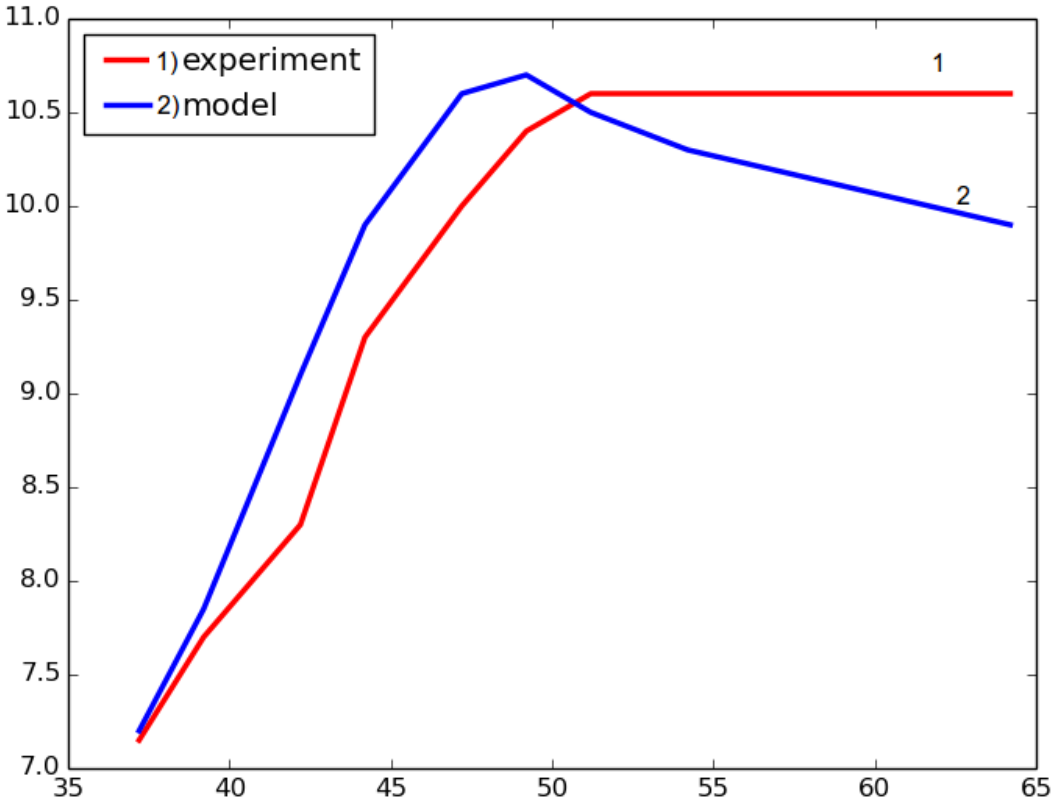

Рис. 1: Зависимость высоты верхней грани от времени. a = 3000м, b = 10000м

целом, по модельным данным верхняя грань всегда была немного выше до точки достижения максимума экспериментальной верхней границы. Так же максимум достигался после 12 минут с начала синхронизации, а в экспериментальных данных через 14 минут.

Случай 2: $a = 3000, b = 9000.$

Необходимо понизить верхнюю грань моделируемого облака для увеличения соответствия с поведением облака в эксперименте. Для этого уменьшим радиус внешнего цилиндра. Уменьшение отношения радиусов приведет к уменьшению высоты верхней грани [10].

Как видим из таблицы 2, мы уменьшили высоту верхней границы, но теперь раньше стала начинаться диссипация, что можно объяснить тем, что когда мы уменьшили радиус внешнего цилиндра, мы увеличили скорость нисходящего потока.

t эксп.	16.18	16.20	16.23	16.25	16.28	16.30	16.32	16.35	16.55
t мод1 мин	38.40	40.40	43.40	45.40	48.40	50.40	52.40	55.40	65.40
z эксп.Км	7.15	7.7	8.3	9.3	10.0	10.4	10.6	10.6	10.6
z мод.Км	7.2	7.9	8.9	9.6	10.3	10.4	10.2	10.1	9.6

Таблица 2: Значения высоты верхней грани облака в моделировании и в эксперименте

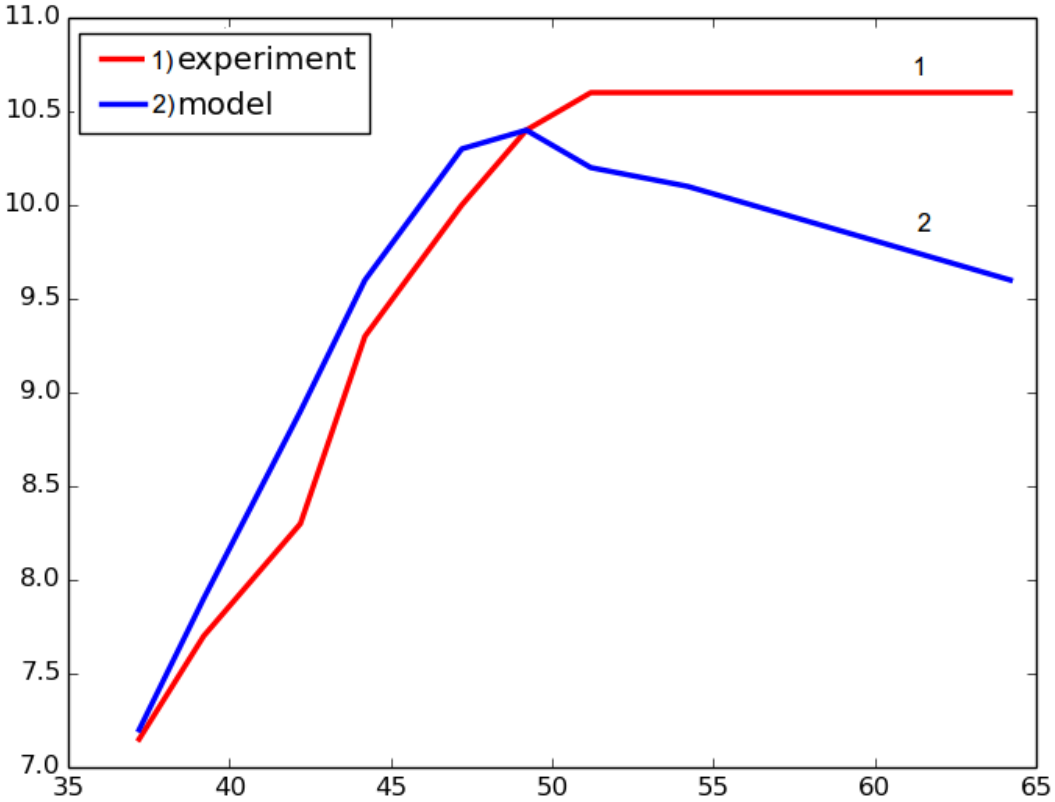

Рис. 2: Зависимость высоты верхней грани от времени. a = 3000м, b = 9000м

Случай 3: $a = 4000, b = 12000.$

Для решения проблемы возникшей в случае 2, попробуем увеличить значения радиусов не изменяя их отношение.

Из таблицы 3 видим, что получили результат схожий с результатами из таблицы 1, когда радиусы внутреннего и внешнего цилиндров были соответственно 3000 м и 10000 м, однако в нашем случае облако быстрее диссипирует.

t эксп.	16.18	16.20	16.23	16.25	16.28	16.30	16.32	16.35	16.55
t мод мин	37.40	39.40	42.40	44.40	47.40	49.40	51.40	54.40	64.40
эксп.Км	7.15	7.7	8.3	9.3	10.0	10.4	10.6	10.6	10.6
мод.Км	7.2	8	9.1	9.9	10.6	10.7	10.5	10.3	9.8

Таблица 3: Значения высоты верхней грани облака в моделировании и в эксперименте

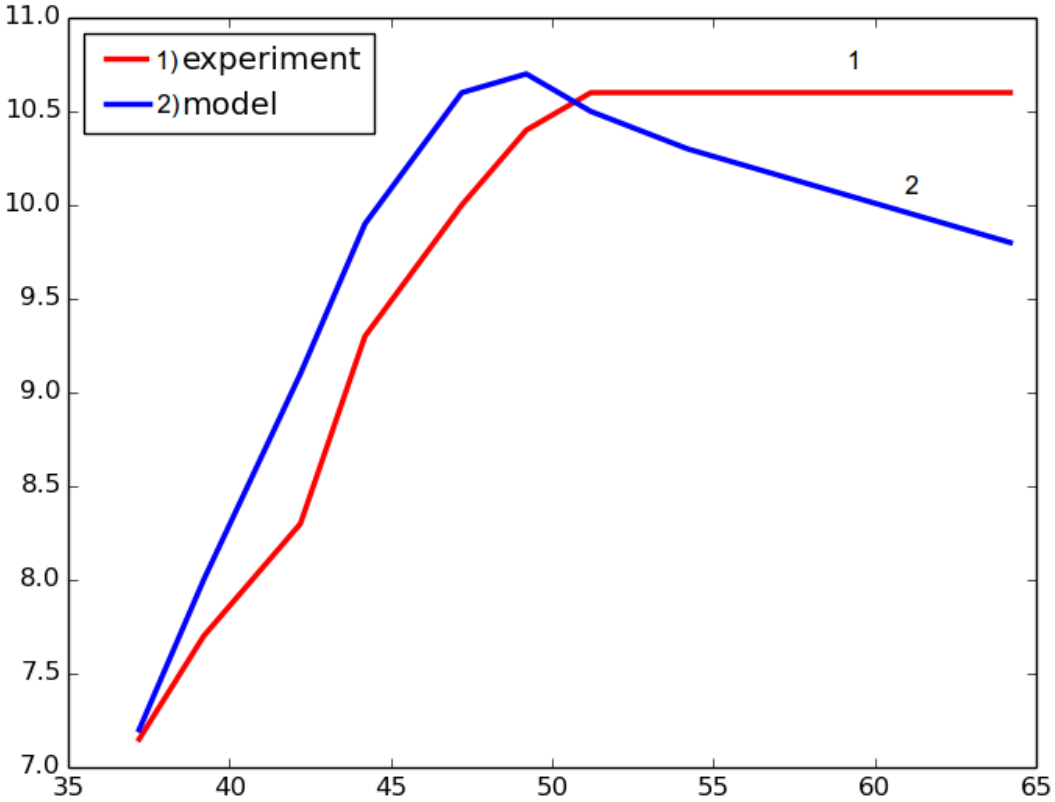

Рис. 3: Зависимость высоты верхней грани от времени. a = 4000м, b = 12000м

Проанализировал таблицы 1-3 кажется разумным использование рекомендуемых авторами модели значений радиусов $a = 3000$ м и $b = 10000$ м. Далее в работе будут использоваться именно эти значения.

2.2. Время жизни облака

По данным эксперимента время жизни облака составляет порядка 30-40мин с 16.00 (учитывая время образования) до 16.36-16.50мин. В нашей модели с

двумя радиусами и с детальной микрофизикой эволюция облака продолжалась в течении, приблизительно, 32мин., т.е. несколько меньше, чем в натурном эксперименте, после чего облако в модели начинает диссипироваться.

2.3. Изменение во времени скорости восходящего потока

t эксп.	16.17	16.20	16.22	16.25	16.29	16.32	16.40	16.42	16.47	16.50
H км	6.0	6.0	6.0	6.0	5.4	5.2	5.3	5.6	6.0	6.1
Wэ.м/с	5	3	10	10	9	6	3	0.3	1	-2
Wм.м/с	11.4	10.8	10.3	8.8	8.0	6.3	2.9	1.1	-2.8	-5.1

Таблица 4: Изменение во времени скорости восходящего потока

Как можно видеть из таблицы 4, скорости по модельным данным достаточно хорошо согласуются с экспериментальными данными (за исключением начала наблюдения). Ход изменения скорости после достижения максимума по модели в целом совпадает с ходом изменения скорости по экспериментальным данным.

2.4. Изменения во времени водности облачных капель на высоте 5.2 – 6.1км

t эксп.	16.17	16.20	16.22	16.25	16.29	16.32	16.40	16.42	16.47	16.50
H км	6.0	6.0	6.0	6.0	5.4	5.2	5.3	5.6	6.0	6.1
LWC э	1.37	1.01	1.75	2.19	1.77	0.71	0.34	0.04	0.06	0.04
LWC м	3.47	3.59	3.72	3.98	2.64	2.41	1.98	0.73	0	0

Таблица 5: Изменения во времени водности облачных капель на высоте 5.2 – 6.1км в 10^{-3} кг/м^3

Из таблицы 5 видно, что водность при моделировании завышена, причем в некоторых местах более чем в 3 раза. Скорее всего так происходит из-за используемого в модели метода решения кинетического уравнения коагуляции

Коветца-Олунда, основным недостатком которого является фиктивное расширение спектра частиц [11]. Так же надо учитывать, что мы усредняем все параметры по сечению облака.

Глава 3. Настройка численной модели конвективного облака

Используемая в данной работе численная модель конвективного облака исходно не была предназначена для одновременной обработки более чем одной зондировки за раз. Так же в ней не было возможности получения каких-либо параметров моделируемого облака в численном виде. Поэтому первоочередной задачей являлась задача автоматизация модели.

3.1. Автоматизация модели

Для обработки всех зондировок и формирования по ним обучающего множества для использования машинного обучения необходимо было автоматизировать модель. В данной работе мы, например, не нуждались в графическом интерфейсе. Вместо всевозможных графиков для различных параметров облака необходимо было сделать возможным получение численных значений любых параметров в любой момент времени на протяжении всего времени моделирования облака. Так же необходимо было сделать возможным обработку сразу группы зондировок, вместо работы с каждой по отдельности, как это было единственно возможным в модели.

Такие параметры моделирования как шаг по высоте, шаг по времени, коэффициент вертикального турбулентного перемешивания и некоторые остальные могут регулироваться, однако для использования численной модели при прогнозировании на практике необходимо зафиксировать их какие-то конкретные значения. Эта необходимость обусловлена тем, что для использования методов машинного обучения важно, чтобы и обучение, и эксплуатация происходили в одинаковых условиях(в противном случае, мы не сможем быть до конца уверенными в истинных причинах правильных или неправильных результатах работы алгоритмов). Смена параметров возможна, но потребует заново сформировать обучающее множество и снова пройти процесс обучения.

Использовались следующие значения параметров при моделировании: шаг по времени $\Delta t = 20c$, шаг по высоте $\Delta h = 150$м , коэффициент бокового

турбулентного перемешивания $\alpha^2 = 0.08$, коэффициент вертикальной турбулентной диффузии $K_v = 100$. Радиусы внутреннего и внешнего столбов были приняты равными 3 км и 10 км соответственно.

Для каждой зондировки облако моделировалось в течение одного часа. Вообще говоря, модель способна моделировать облако в промежутке времени до шести часов подряд. Выбранное время моделирования объясняется снижением продолжительности обработки одной отдельно взятой зондировки и тем, что один час является типичным промежутком времени эволюции облака.

Замечание: было несколько занижено предложенное автором модели значение $\alpha^2 = 0.1$. Это объясняется тем, что, как было показано в главе 2, реализованный в модели метод расчета коагуляции Коветца-Олунда страдает значительным увеличением значения водности.

Глава 4. Использование численной модели конвективного облака

После автоматизации модели для использования методов машинного обучения с учителем требовалась представительная выборка зондировок как содержащих опасное конвективное явление, так и не содержащие.

4.1. Источник данных

Исходные данные для моделирования были получены с помощью использования комплексной информационной системы для формирования входных данных моделей конвективных облаков [12–14]. Всего было получено 615 зондировок: 289 из них были собраны когда никакого явления не наблюдалось и 326 когда было опасное конвективное явление.

4.2. Предобработка данных

Для использования собранных зондировок посредством модели потребовалась их предобработка по причине того, что практически у всех численных моделей конвективного облака существует проблема с развитием облака при наличии задерживающего слоя: облако может не развиться. Проблема практически полностью решилась внесением исправлений в зондировки, а именно, нахождением высоты уровня конденсации с помощью формулы Ипполитова [15] с последующим проведением сухо адиабатического градиента до этого уровня.

Для проверки корректности данного подхода была собрана контрольная выборка, состоящая из 128 зондировок содержащих явление и 104 не содержащих, предоставленных Институтом радарной метеорологии (IRAM). При этом, без внесения в них каких-либо корректировок при моделировании, лишь в 22 из 128 действительно возникало явление, и при этом ни в одной из 104 явление не наблюдалось. После указанной предобработки зондировок выяснилось, что теперь при моделировании в 117 из контрольных 128, содер-

жащих явление, оно действительно возникало, при этом из 104 зондировок, где явление не должно было наблюдаться, оно по прежнему не наблюдалось в 102 и возникло всего в 2 зондировках.

4.3. Использование модели

Для входа модели используются исправленные зондировки, а на выходе получаются данные в формате CSV(Comma-Separated Values) для каждой зондировки в следующем формате: время, высота, название параметра, значение параметра. Выводились следующие параметры: вертикальная составляющая скорости, горизонтальная составляющая скорости, давление, плотность, температура, отклонение температуры от температуры окружающей среды, относительная влажность(над водяной поверхностью), отношение смеси пара, общее отношение смеси аэрозолей, капель, ледяных частиц, крупы, градин, вертикальная мощность облака. Для выбора нижней и верхней границ облака использовалось значение общего отношения смеси водных капель равное 0.0015. Данное значение было подобрано опытным путем в процессе валидации модели, чтобы она воспроизводила облако как можно более схожее к тому, которое было в натурном эксперименте. Пример фрагмента выходного файла приведен в таблице 6.

Время	Высота	Название параметра	Значение параметра
1160	4050	velocity	13.067016
1160	4050	velocityU	9.343779
1160	4050	temperature	272.22715
1160	4050	relativeHumidity	1.0090644
1160	4050	vapor	0.0058138833
1160	4050	pressure	61652.788
1160	4050	density	0.7889759
1160	4050	aerosol	0.134232
1160	4050	drop	0.002514255
1160	4050	iceHailAndGrits	2.2671957E-0006
1160	4200	velocity	11.32452
1160	4200	velocityU	39.735083

Таблица 6: Пример фрагмента выходного файла модели

На процессоре intel core i5 и видеокарте NVidea GeForce 840M одна зондировка моделируется в среднем 2-3 минуты.

Глава 5. Применение методов машинного обучения

Реализации всех методов машинного обучения были использованы из библиотеки scikit-learn [16]. Scikit-learn - библиотека с открытым исходных кодом, написанная на языке программирования Python и содержащая множество различных алгоритмов машинного обучения. Распространяется под лицензией BSD (Berkeley Software Distribution), допускающей помимо прочего также и коммерческое использование данной библиотеки.

Представлен ранее неиспользуемый подход для автоматической классификации зондировок. В данной работе мы имеем дело с задачей классификации зондировок на два класса: содержащие опасное конвективное явление и не содержащие.

Машинное обучение используется для автоматизации нахождения решения задачи определения по зондировке: будет ли опасное конвективное явление или нет, в результате чего данная численная модель может быть использована для оперативного прогноза опасного конвективного явления в различных метеоцентрах. Представлены результаты применения различных методов машинного обучения для классификации зондировок посредством обучения с учителем. Наше обучающее множество состоит из численных параметров, смоделированных численной моделью облака для каждой зондировки, и является размеченным вручную, то есть для каждой зондировки из нашего множества мы знаем, наблюдалось ли какое-либо опасное конвективное явление или нет.

Так же будет показано, что для всех используемых алгоритмов машинного обучения была достигнута точность свыше 95%.

5.1. Краткое определение машинного обучения

Машинное обучение — математическая дисциплина, позволяющая посредством использования различных разделов теории вероятностей, математической статистики и численных методов, получать знания из имеющихся данных. Она используется для автоматизации решения различных задач в са-

мых разных областях человеческой деятельности. Машинное обучение используется в области компьютерного зрения, медицинской диагностики, распознавания речи. При этом его область применений постоянно расширяется. В наши дни в результате повсеместной информатизации накоплены внушительные объёмы данных во всевозможных отраслях, таких как производство, наука, бизнес, здравоохранение. Раньше, когда таких данных не было в наличии, эти задачи либо вообще не ставились, либо решались совершенно иными методами.

5.2. Выделение значимых признаков

В качестве признаков для применения методов машинного обучения в задаче прогнозирования выступили численные параметры смоделированного облака. Было решено для прогнозирования выбирать момент максимального развития облака, то есть момент времени, когда была достигнута максимальная вертикальная мощность облака (разность между высотой верхней и нижней границ облака) так как максимальная мощность облака позволяет наиболее реалистично судить об интенсивности развития конвекции, и, следовательно, о вероятности наступления опасного конвективного явления. Выбор численных параметров для прогнозирования происходил на высотном уровне, где наблюдалось максимальное отношение смеси водных капель. Далее выбранные параметры были подвергнуты нормализации в ввиду того, что большинство градиентных методов, лежащие в основе почти всех алгоритмов машинного обучения, сильно чувствительны к шкалированию данных.

Из всех 13 воспроизводимых моделью типов численных параметров необходимо определить наиболее информативные, то есть те, по которым лучше всего осуществлять прогноз. Для их отбора был использован алгоритм перебора Recursive Feature Elimination [17] с автоматическим выбором признаков (и их количества) посредством техники оценки результата методом скользящего окна для метода опорных векторов с линейным ядром. На рисунке 4 изображен график зависимости точности предсказания от количества задействованных признаков.

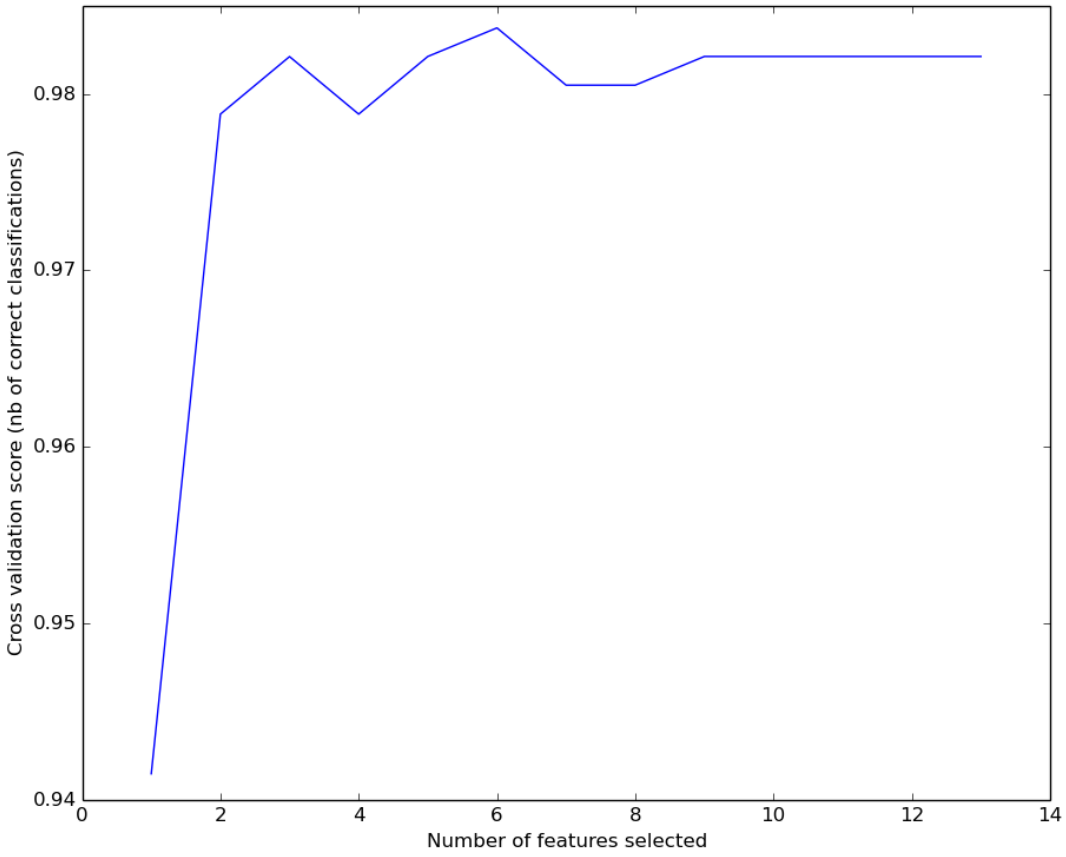

Рис. 4: зависимость точности предсказания от количества задействованных признаков

Оптимальным для нашего случая оказалось использовать следующие шесть параметров модели как признаки для прогнозирования: вертикальная составляющая скорости, отклонение температуры от температуры окружающей среды, относительная влажность(над водяной поверхностью), отношение смеси пара, общее отношение капель, вертикальная мощность облака.

5.3. Использование различных методов

Использовались следующие методы машинного обучения: метод опорных векторов (support vector machine) [18], логистическая регрессия (logit model) [18], Гребневая Регрессия (Ridge Regression) [19]. Для оценки качества работы использовался метод скользящего окна.

Метод опорных векторов относится к семейству линейных классификаторов и является одним из самых популярных методов машинного обучения. Ключевая идея метода заключается в переводе исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Опорными векторами будем называть вектора ближайшие к противоположному классу (то есть лежащие на границе класса). Две параллельные гиперплоскости строятся через опорные вектора. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Использовался метод опорных векторов с линейным ядром. Точность предсказания составила 0.986 (98.6%). Оптимальная гиперплоскость описывается уравнением (1).

$$\begin{aligned} f(\overline{x}) = 2.33572643x_1 + 0.83159455x_2 + \\ +0.59938731x_3 + 0.7031227x_4 + 0.65584503x_5 + \\ +2.77717664x_6 + 1.31997495 = 0 \end{aligned} \tag{1}$$

где $\overline{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ вектор признаков, состоящий из нормализованных значений параметров облака. Здесь x_1 равно значению вертикальной составляющей скорости, x_2 - отклонению температуры от температуры окружающей среды, x_3 - относительной влажности(над водяной поверхностью), x_4 - отношению смеси пара, x_5 - общему отношению капель, x_6 - вертикальной мощности облака. Правило классификации определяется как $G(x) = sign(x)$, то есть в случае, когда $f(x) \geq 0$ — опасное конвективное явление будет наблюдаться, и при $f(x) < 0$ — не будет.

Логистическая регрессия — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой. В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, зна-

чением функции является вероятность того, что данное исходное значение принадлежит к определенному классу.

Точность предсказания составила 0.977 (97.7%). Получившаяся решающая функция описывается уравнением (2).

$$\begin{aligned} z = 3.88523988x_1 + 1.84193614x_2 + \\ +2.8254668x_3 + 1.94093366x_4 + 1.05776969x_5 + \\ +2.90365228x_6 + 3.68298532 \end{aligned} \tag{2}$$

где $\overline{x}$ - такой же вектор признаков как и в (1) и вероятность что будет явление при заданном векторе признаков $Pr(y = 1|\overline{x}) = f(z)$ и $f(z) = \frac{1}{1+e^{-z}}$. Соответственно чем больше $f(z) \in [0, 1]$, тем более вероятно возникновение опасного конвективного явления. Вероятность противоположного события, то есть того, что опасное конвективное явление наблюдаться не будет равно $1 - f(z)$, соответственно если $f(z) \geq 0.5$, то явление будет наблюдаться и при $f(z) < 0.5$ - не будет.

Гребневая Регрессия - вариация Метода наименьших квадратов, решающая некоторые характерные для него проблемы: чаще всего применяется при наличии переизбыточности в данных, когда независимые переменные коррелируют друг с другом (т.е. имеет место мультиколлинеарность). Данная проблема решается с помощью наложения штрафа зависящего от размера коэффициентов, то есть решается задача $\min_w ||Xw-y||_2^2+\alpha||w||_2^2$. Здесь α является комплексным параметром контролирующем величину штрафа: чем больше α, тем более устойчивыми к коллинеарности становятся коэффициенты.

Точность предсказания составила 0.981 (98.1%). Решающая функция описывается уравнением (3).

$$\begin{aligned} f(\overline{x}) = 1.24574578x_1 - 0.19004467x_2 + \\ +1.3471659x_3 - 0.03611449x_4 + 0.65584503x_5 + \\ +0.74499218x_6 - 1.40857864 \end{aligned} \tag{3}$$

где $\overline{x}$ - такой же вектор признаков как и в (1). Правило классификации определяется как $G(x) = sign(x)$, то есть в случае, когда $f(x) \geq 0$ — опасное конвективное явление будет наблюдаться, и при $f(x) < 0$ — не будет.

Выводы

Была выбрана наиболее подходящая для поставленных в данной работе целей полуторамерная нестационарная модель конвективного облака с подробным описанием микрофизических процессов, позволяющая без больших вычислительных затрат моделировать облако с высокой степенью достоверности. Проведена валидация выбранной модели по данным натурного эксперимента. После этого данная модель была автоматизирована.

Комплексная информационная система была использована для сбора метеорологических данных послуживших входными данными для модели.

Был использован алгоритм Recursive Feature Elimination для определения численных параметров облака, лучше всего подходящих для совершения прогноза опасного конвективного явления. В итоге были выбраны шесть оптимальных для прогнозирования параметров облака для использования в качестве признаков для машинного обучения: вертикальная составляющая скорости, отклонение температуры от температуры окружающей среды, относительная влажность(над водяной поверхностью), отношение смеси пара, общее отношение капель, вертикальная мощность облака.

Применены три метода машинного обучения: метод опорных векторов, логистическая регрессия и Гребневая Регрессия. Для каждого из них получены решающие функции. Точности использования этих методов составили 97.7 % для метода опорных векторов, 98.6% для логистической регрессии и 98.1% для Гребней Регрессии.

Объединяя все полученные результаты проделанной работы, можно привести алгоритм применения на практике численной модели для прогнозирования опасного конвективного явления:

1. Полученные оперативные данные радиозондирования атмосферы корректируются проведением сухоадиабатического градиента от высоты найденного с помощью формулы Ипполитова уровня конденсации.

2. Модифицированные зондировки используются как вход для численной модели облака.

3. Моделируется облако и производится выбор необходимых численных параметров в нужный момент времени и на нужной высоте.

4. Выбранные параметры используются как признаки для вычисления значения решающей функции.

5. Делается вывод по результатам прогноза одной или нескольких решающих функций. Например, можно делать вывод о том, что будет наблюдаться опасное конвективное явление, если хотя бы две из трех решающих функции определили что оно будет наблюдаться.

В дальнейшем планируется реализовать двойную классификацию: сначала определять будет явление или нет, а затем уже если будет явление, то какое именно.

Заключение

Все поставленные цели были достигнуты.

Был осуществлен выбор наиболее подходящей для нашей задачи численной модели облака. Проведена её валидация и подбор параметров ”настройки” путем сопоставления основных характеристик облака с соответствующими характеристиками, полученными в результате натурного эксперимента. Расчеты по численной модели была автоматизированы, что позволило осуществлять серии численных экспериментов с различными данными радиозондирования атмосферы, которые использовались в качестве входных параметров. Предложен и обоснован метод модификации исходных данных радиозондирования путем использования формулы Ипполитова для определения уровня конденсации и сухоадиабатического градиента температуры в подоблачном слое. Такая модификация может рассматриваться в качестве универсального способа препроцессинга входных данных для всех видов численных моделей облаков. Она позволяет моделировать развитие облака даже в случае наличия в подоблачном слое слоев температурной инверсии и изотермии (задерживающих слоев).

Была получена обучающая выборка данных радиозондирования атмосферы с помощью комплексной информационной системы, которая позволила в автоматическом режиме отобрать 615 вертикальных профилей температуры и влажности, 326 из которых наблюдались при опасных конвективных явлениях, а 289 - в отсутствии таких явлений.

Реализован прогноз опасного конвективного явления по результатам работы численной модели. При этом удалось достигнуть точности предсказания свыше 95 %.

Список литературы

[1] *Руководство по прогнозированию метеорологических условий для авиации* Под редакцией К.Г. Абрамович, А.А. Васильева. Авторы: Абрамович К.Г., Васильев А.А., Булдовский Г.С., Борисова В.В., Глазунов В.Г., Горлах И.А., Лешкевич Т.В., Ляхов А.А., Рацимор М.Я., Решетов Г.Д., Рубинштейн М.В., Шакина Н.П.. Госкомгидромет, Москва, 1985, 308 стр.

[2] Tao W.-K. *Consistent 2D and 3D Cloud Resolving Model Simulations* // 13th ARM Sci. Team Meeting Proceedings. – 2003. – P. 1–8

[3] Морозов В.Н., Веремей Н.Е., Довгалюк Ю.А. *Моделирование процессов электризации в трехмерной численной модели осадкообразующего конвективного облака* // Труды Главной геофизической обсерватории им. А.И.Воейкова. – СПб., 2009. – Вып. 559. – С. 134–160. – ISSN 0376-1274

[4] Shiino J. A *Numerical Study of Precipitation Development in Cumulus Clouds* // Papers in Meteorology and Geophysics. – 1978. – Vol. 29, N. 4. – P. 157–194.

[5] Довгалюк Ю.А. *Анализ результатов работ по воздействию на облака с целью предотвращения осадков в г. Ленинграде (на примере опыта 7 ноября 1988 г.)* // Метеорология и гидрология. – М., 1998. – No2 – С. 44–53. – ISSN 0130-2906.

[6] N. Raba, E. Stankova and N. Ampilova *One-and- a-half-dimensional Model of Cumulus Cloud with Two Cylinders. Research of Influence of Compensating Descending Flow on Development of Cloud.* // Proceedings of the 5th International Conference “Dynamical Systems and Applications” Ovidius University Annals Series: Civil Engineering Volume 1, Special Issue 11, June 2009, pp.93-101

[7] 7. Raba N.O., Stankova E.N., Ampilova N. *On Investigation of Parallelization Effectiveness with the Help of Multi-core Processors* // Procedia Computer Science. 2010. Vol. 1, Issue 1. P. 2757-2762

[8] 9. N. Raba, E. Stankova *On the Problem of Numerical Modeling of Dangerous Convective Phenomena: Possibilities of Real-Time Forecast with the Help of Multi-core Processors* // Murgante et al. (Eds.): ICCSA 2011, LNCS 6786, pp. 633 – 642, 2011. ISSN 0302-9743

[9] Dye, J.E., Jones, J.J., Winn, W.P., Cerni, T.A., Gardiner, B., Lamb, D., Pitter, R.L., Hallett, J., Saunders, C.P.R. *Early Electrification and Precipitation Development in a Small, Isolated Montana Cumulonimbus* // J. of Geophys. Res., 1986. 91, # D1, 1231-1247

[10] Раба Н.О., Станкова Е.Н. *Исследование влияния компенсирующего нисходящего потока, сопутствующего конвективным течениям, на жизненный цикл облака с помощью полуторомерной модели с двумя цилиндрами* // Труды ГГО. 2009, Вып.559. С. 192-209 (ISSN 0376-1274)

[11] Kovetz A., Olund B *The effect of coalescence and condensation on rain formation in a cloud of finite vertical exten* // J. Atm. Sci. 1969. V. 26. № 9. P. 1060–1065

[12] Dmitry A. Petrov and Elena N. Stankova *Use of Consolidation Technology for Meteorological Data Processing* // B. Murgante et al. (Eds.): ICCSA 2014, Part I, Lecture Notes in Computer Science 8579, pp. 440–451. Springer International Publishing Switzerland (2014) DOI 10.1007 1/978-3-319-09144-0-30.

[13] Станкова Е.Н., Петров Д.А. *Комплексная информационная система, предназначенная для формирования входных данных моделей конвективных облаков* // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. Серия 10. 2015 Выпуск 3. Стр. 83-95

[14] Dmitry A. Petrov and Elena N. Stankova *Integrated Information System for Verification of the Models of Convective Clouds* // O. Gervasi et al. (Eds.): ICCSA 2015, Part IV, LNCS 9158, pp. 321–330, 2015. DOI: 10.1007/978-3-319-21410-8-25

[15] Матвеев Л. *Курс общей метеорологии. Физика атмосферы* Издание второе, переработанное и дополненное. Л.: Гидрометеоиздат, 1984. 751 с.

[16] Scikit-learn. Machine Learning in Python. http://scikit-learn.org/

[17] I.Guyon, J.Weston, S.Barnhill, V.Vapnik *Gene selection for cancer classification using support vector machines* // Machine Learning, vol. 46, nos. 1-3, (2002), pp. 389-422

[18] T. Hastie, R. Tibshirani, J. Friedman *The elements of Statistical Learning* Second edition, Springer, 2009

[19] Arthur E. Hoerl and Robert W. Kennard *Ridge Regression: Biased Estimation for Nonorthogonal Problems* // echnometrics, Vol. 12, No. 1 (Feb., 1970), pp. 55-67