

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ рОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

**«Дальневосточный федеральный университет»**

(ДВФУ)

|  |  |  |
| --- | --- | --- |
| **ШКОЛА ЕСТЕСТВЕННЫХ НАУК**  **кафедра информатики, математического и компьютерного моделирования** | | |
|  | |
| Комплексирование небольшого количества прогнозов с помощью градиентного бустинга |
|  |

**КУРСОВАЯ РАБОТА**

по образовательной программе подготовки бакалавров

по направлению 01.03.02 «Прикладная математика и информатика»

|  |  |
| --- | --- |
| Работа защищена  с оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Регистрационный номер \_\_\_\_\_\_\_  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_г. | Студент группы № Б8\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Иванов И.И.  (подпись)  «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_\_\_г.  Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (должность, ученое звание)  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (подпись) (ФИО)  «\_\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_г. |

г. Владивосток

2019

# Введение

Данная курсовая работа посвящена исследованию путей улучшения качества составления прогнозов погоды на метеостанциях Приморского края.

Прогнозом погоды называется описание метеорологических условий, ожидаемых в определенный момент или период времени в определённом районе или части воздушного пространства.

В связи с изменчивость метеорологических величин в пространстве и во времени, а также ввиду несовершенства методов прогнозирование и изменения некоторых величин, конкретное значение любой указанной в прогнозе величины следует рассматривать как наиболее вероятное значение, которое данная величина может иметь в период действия прогноза. Точно так же время возникновения какого-либо явления следует понимать, как наиболее вероятное время.

Конкретные требования к точности прогноза разрабатываются потребителями прогноза совместно с метеорологами с учетом специфики деятельности и возможности прогнозирование. Наибольший практический интерес представляют прогнозы общего назначения, ориентированные на широкий круг, пользователь, главным образом население.

В связи с этим элементы погоды были ограничены следующими показателями: экстремальная температура воздуха на высоте 2 м., максимальная скорость ветра, преобладающее направление ветра, сумма осадков за полусутки и рассматривалось только краткосрочного прогнозирование.

# Описание предметной области

Модель исследования погоды и прогнозирования (WRF) - численная мезомасштабная атмосферная модель, разработанная как для научных исследований, так и для численного прогнозирования погоды. С момента своего первого публичного выпуска в 2000 году WRF стала самой используемой атмосферной моделью в мире. Модель служит для широкого спектра метеорологических применений в масштабах от десятков метров до тысяч километров.

В ней численно решается полная система уравнений гидротермодинамики атмосферы, которая включает в себя уравнения неразрывности, переноса импульса, переноса влаги и внутренней энергии.

Требует для инициализации и задания граничных условий результаты расчетов глобальной модели численного прогноза погоды. В качестве начальных условий должны задаваться трехмерные распределения полей трех компонент скорости, температуры, давления и влажности. В качестве граничных условий должны задаваться значения температуры, влажности и компонент скорости на боковых и на верхней границах области, а также значения потоков тепла, влаги и импульса на нижней границе области, прилегающей к поверхности Земли.

Результаты расчета WRF на прямую не используются синоптиками. Из-за недостатка точности и несовершенного начального состояния численной модели, они подвержены систематическим и случайным ошибкам, которые имеют тенденцию расти со временем. Если оставить результаты расчетов необработанными и по их результатам составить прогноз элементов погоды, то оправдываемость их окажется на уровне 60-70%.

Для обхода данной проблемы существует класс методов статистической постобработки. Одним из самых распространённых является метод статистическая интерпретация гидродинамического прогноза (MOS). MOS был определен Гланом и Лору [4, 5] в 1972 году следующим образом: это объективный метод прогнозирование погоды, в основе которого лежит предположение, что поля воспроизводимые моделями и данные наблюдений связаны линейной регрессией.

4. Klein W.H., Glahn H.R. Forecasting local weather by means of model output statistics // Bull. Am. Meteorol. Soc. 1974. Vol. 55, № 10. P. 1217–1227.

5. Glahn H.R., Lowry D.A. The use of model output statistics (MOS) in objective weather forecasting. // J. Appl. Meteorol. 1972. Vol. 11, № 8. P. 1203–1211.

Не смотря на огромное распространения и доказанную эффективность метода MOS. Он требует не менее двух лет архивных результатов и наблюдении, в течении которых параметры численной модели должна оставаться неизменными или почти такими же. Это требование необходимо для того, чтобы полностью охватить характеристики ошибок модели при широком разнообразии погодных условий для любого конкретного местоположения или региона. Учет экстремальных метеорологических явлений, таких как проливные дожди, снегопады, сильные ветра и т.д. необходим при разработке надежной системы MOS.

К сожалению, метеорологические центры Приморья не обладают требуемым объемом наблюдений, а имеющиеся данные частично испорчены. Поэтому применение технологии MOS не эффективно.

Требуется разработать новые методы прогноза погоды по данным численной модели WRF. Исследуем возможность применения ассемблерных композиций в качестве данного метода.

В конце 80-ых начали появляться работы с исследованием проблемы связи слабой и сильной обучаемости алгоритмов. Слабая обучаемость алгоритма означает, что за полиномиальное время возможно построить алгоритм распознавания, точность которого будет хотя бы немного больше 50%. Под сильной обучаемостью подразумевается, что возможно за полиномиальное время построить алгоритм, который бы мог давать сколь угодно точные результаты. Исследования показали, что сильная обучаемость эквивалентна слабой, поскольку любую слабую модель можно усилить, построив правильную композицию, но потребовалось несколько лет до появления работоспособных алгоритмов.

Robert E. Schapire, The Strength of Weak Learnability - 1990

В 1996 эти идеи сформировались и приобрели законченную форму в виде алгоритма AdaBoost [7].. Несколькими годами позже появилось важное обобщение этого алгоритма – градиентный бустинг [1].

[1] Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. — IMS 1999 Reitz Lecture.

[7] Freund Y., Shapire R. Experiments with a New Boosting Algorithm. —1996.

# Постановка задачи

Цель работы – комплексирование небольшого количества прогнозов с помощью градиентного бустинга для исследования применимости данного метода на метеостанциях Приморского края.

# Градиентный бустинг

## Математическая постановка задачи

Будем решать задачу восстановления функции в контексте обучения с учителем. Пусть дана набор пар признаков  и целевых переменных

Необходимо восстановить зависимость вида

.

Ограничим пространство поиска каким-нибудь конкретным параметризованным семейством функций

# Построение градиентного бустинга

Будем восстанавливать функцию приближением   в виде композиции

Будем строить такую композицию путем жадного наращивания, каждый раз добавляя в сумму слагаемое, являющееся наиболее оптимальным алгоритмом из возможных. Таким образом задача сводиться к поиску пары наиболее оптимальных параметров длины

Будем минимизируется функционал ошибки любым итеративными алгоритом

Сделаем это, реализуя один шаг метода градиентного спуска. Найдем градиент функционала ошибки

В силу метода градиентного спуска, наиболее выгодно добавить новое слагаемое следующим образом

где подбирается линейным поиском по вещественным числам :

Однако представляет из себя лишь вектор оптимальных значений для каждого объекта , а не базовый алгоритм из семейства , определенный . Поэтому нам необходимо найти наиболее похожий на . Сделаем это, опять минимизируя функционал ошибки:

Что соответствует базовому алгоритму обучения. Будем искать линейный поиском

# Псевдокод GBM

 На вход алгоритма поступает:

* набор данных ;
* число итераций ;
* выбор функции потерь  с выписанным градиентом;
* выбор семейства функций базовых алгоритмов , с процедурой их обучения;
* дополнительные гиперпараметры , например, глубина дерева у деревьев решений;

В качестве инициализации начального приближения  используется просто константное значение . Его, а также оптимальный коэффициент  находят бинарным поиском относительно исходной функции потерь. Итак, GBM алгоритм:

1. Инициализировать GBM константным значением
2. Для каждой итерации  повторять:
   1. Посчитать псевдо-остатки
   2. Построить новый базовый алгоритм  как регрессию на псевдо-остатках
   3. Найти оптимальный коэффициент  при  относительно исходной функции потерь
   4. Сохранить
   5. Обновить текущее приближение
3. Скомпоновать итоговую GBM модель

# Градиентный бустинг над решающими деревьями

Сейчас же рассмотрим в качестве базового семейства алгоритмов регрессионные решаюшие деревья из вершин. [9]

[9] Breiman L., Friedman J., Olshen R., Stone C. Classiﬁcation and Regression Trees. — Wadsworth, 1983.

Каждое решающее дерево имеет листовых вершин, соответствующие непересекающимся областям , на которые разбивается пространство объектов . Каждой листовой вершине соответствует некоторое значение регрессии , которое будет ответом в случае попадания анализируемого объекта в соответствующую область. Можно записать этот факт следующей формулой:

где – индикатор события . Видно, что в этой сумме ровно одно слагаемое будет ненулевым. Тогда добавление слагаемого в градиентном бустинге будет происходить следующим образом:

Таким образом мы просто прибавляем к имеющему алгоритму некоторое другое решающее дерево. Попробуем найти оптимальные значения для уже построенных :

И, поскольку области не пересекаются, можем переписать эту формулу в следующем виде:

Таким образом вместо того, чтобы выполнять линейный поиск коэффициента перед новым слагаемым, как в классическом градиентном бустинге, мы полностью перенастраиваем параметры дерева с фиксированными . Это позволяет строить более качественную композицию. Такая разновидность бустинга называется Tree-boost.