Построение динамической модели критических температурных аномалий Арктического региона методами машинного обучения по данным длительных космических наблюдений

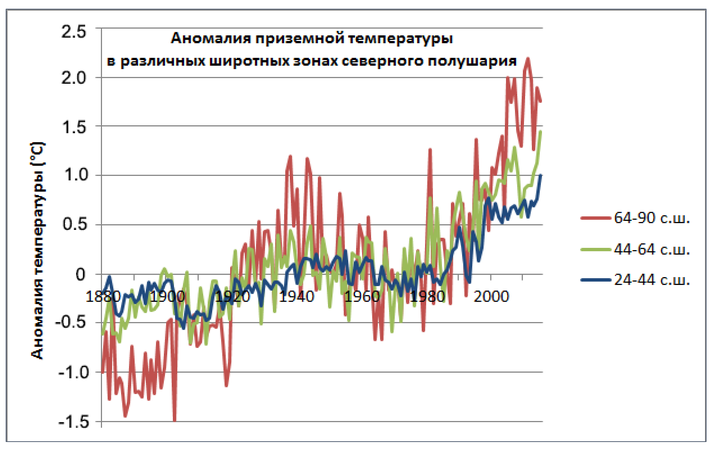
Головко В.А.[1,2)], Федотов И.А. [2)], Синёва А.А. [2)]

1) ФГБУ Научно-исследовательский центр космической гидрометеорологии «Планета»

2) Московский физико-технический институт (государственный университет)

Введение

Глобальный климат Земли определяется радиационным балансом нашей планеты. Последние несколько десятилетий (с 1970-х) космические системы наблюдений фиксируют устойчивый радиационный дисбаланс Земли. Текущая оценка глобального дисбаланса составляет около +0.59 Вт/м^2. Это свидетельствует о том, что Земля получает от Солнца энергии больше, чем её излучает в космическое пространство, т.е. продолжается нагрев нашей планеты [1].

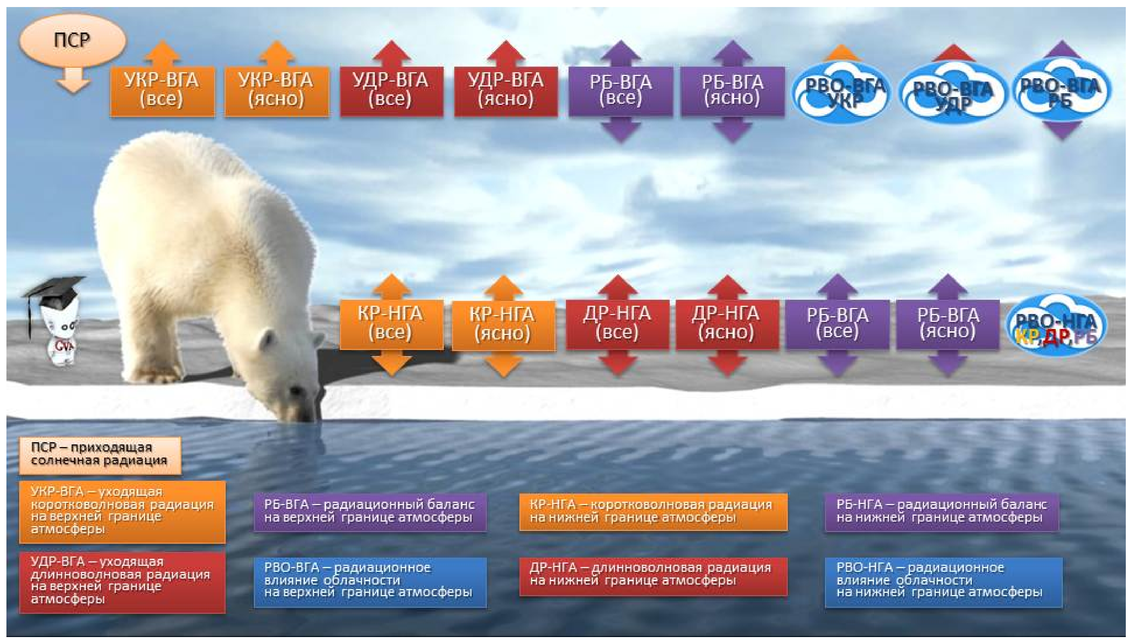


За последние 35 лет (1980-2015 гг.) потепление наблюдалось во всех широтных зонах, однако наиболее значительный рост температуры был зафиксирован в северных широтах. Если глобальная аномалия приземной температуры в 2015 году (признанным самым тёплым за всю историю наблюдений) достигла 0.87°C, то в зоне 64°-90°с.ш. она составила уже 1.76°C.

По мнению большинства экспертов IPCC, если постиндустриальная аномалия приземной температуры возрастет до 2°C (как предсказывает ансамблевый прогноз), то во многих регионах планеты наступит настоящая катастрофа.

Концепция построения модели

Построение математической модели осуществлялось на основе всего набора данных космического мониторинга составляющих радиационного баланса Земли (РБЗ), полученных за последние 38 лет (1978-2016). Однако наиболее полно (с контролируемой точностью) были представлены данные, полученные в интервале с марта 2000 г. по июнь 2016 г. Информация представлена среднемесячными значениями на глобальной сетке 1° x 1° на верхней и нижней границах атмосферы. В качестве независимых переменных рассматривались потоки приходящей и уходящей коротковолновой и длинноволновой радиации, а также данные о радиационном форсинге облачности на эти потоки.



Для учета глобальной и региональной составляющих осцилляций климата, обусловленных особенностями циркуляции, были включены данные индексов Арктической осцилляции (AO), Северо-Атлантической осцилляции (NAO) и Эль-Ниньо/Южное колебание (ENSO). Поскольку большая часть приземной температурной аномалии в Арктике в последнее время связывается с образовавшимся в атмосфере на севере Восточной Сибири планетарным максимумом парниковых газов (в первую очередь метана и CO2), в число независимых переменных были включены значения общего содержания этих газов в атмосфере данного региона, полученные системами космического мониторинга.

В качестве целевой переменной модели рассматривалась приземная температурная аномалия в зоне 64°-90°с.ш.

Построение компьютерной модели

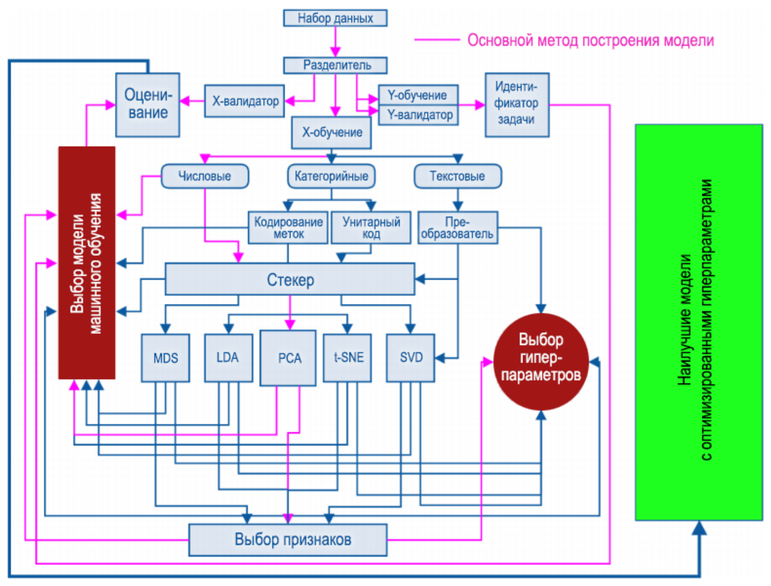
Наиболее трудоёмким является первый этап модельного процесса: валидация, фильтрация и преобразование данных в формат, подходящий для применения алгоритмов машинного обучения. На втором этапе выполняется предобработка и непосредственное обучение моделей.

На первом этапе набор имеющихся данных разделялся методом случайной выборки на две части: обучающей (training set) ~90% и валидационной (validation set) ~10%.

Ввиду больших объёмов исходных данных важными последующими этапами разработки модели были уменьшение размерности и отбор наиболее информативных признаков.

Особенности алгоритмической реализации модели

При решении задачи понижения размерности кроме классических методов, в частности, метода главных компонент (PCA) и многомерного шкалирования (MDS)использовался также один из мощнейших современных методов - t-SNE (t-Distributed Stochastic Neighbor Embedding). Отбор признаков осуществлялся как одним из популярных методов, основанном на "жадном" алгоритме (greedy feature selection), так и современными методами логистической регрессии (logistic regression) и случайного леса (random forest).



Важным этапом разработки являлся выбор алгоритмов машинного обучения и настройка гиперпараметров. В настоящее время доступно большое число библиотек, реализующих самые современные алгоритмы, которые целесообразно было апробировать [2]: случайный лес (random forest), градиентный бустинг (gradient boosting), логистическая регрессия (logistic regression), наивный Байес (naive Bayes), метод опорных векторов (support vector machine).

Разные алгоритмы требуют и различных подходов к выбору оптимальных значений параметров для них. Это наиболее сложная задача, возникающая на этапе настройки гиперпараметров. К сожалению однозначных рекомендаций, на этот счет не существует, и получить приемлемые результаты для каждого алгоритма, можно только постепенно приобретая опыт работы с комбинациями параметров на различных случайных выборках имеющегося набора данных.

Наилучших результатов при построении модели удалось достичь с помощью современных ансамблевых методов (ensemble method). Преимущества ансамблевых методов заключаются в том, что это синтез алгоритмов машинного обучения, которые обучают множество классификаторов, а затем классифицируют новые наблюдения, объединяя прогнозы этих классификаторов на основе взвешенного большинства голосов. В результате использования ансамбля уменьшается смещение (bias), уменьшается дисперсия (variance), минимизируется эффект переобучения. Сейчас в качестве наиболее перспективных новых техник совершенствования модели можно рассматривать такие методы, как бэггинг (bagging) и бустинг (boosting).

Для машинного обучения некоторых блоков модели были использованы быстро развивающиеся сейчас нейросетевые технологии. Они помогают найти в исходных данных уже известные (по прецедентам) паттерны/шаблоны климатических изменений. Но, несмотря на все современные успехи применения нейросетей (в частности, «глубокого обучения» – deep learning), есть и одна общая проблема: полученные с их помощью результаты часто очень трудно интерпретируемы, а это в свою очередь означает, что бывает достаточно трудно определить, когда результат может оказаться ошибочным.

Основная часть исходных данных (за последние 16 лет) представлена в виде качественных непрерывных временных рядов. Это позволяет в рамках динамической модели решать не только задачи классификации и регрессии, но и задачи, присущие анализу исключительно временных рядов. В их число входит выявление трендов, особенностей сезонного поведения и краткосрочный прогноз.

Заключение

Основным достоинством динамической модели является возможность уточнения всех найденных закономерностей по мере пополнения временных рядов. Это некоторая современная версия обучения с подкреплением (reinforcement learning), когда имеет место определенная форма обратной связи для каждого этапа прогнозирования. Данный подход позволяет использовать динамическую модель как само-подстраивающуюся экспертную систему текущих и будущих климатических изменений в Арктике.

Несмотря на понимание мировым научным сообществом важности проявлений последствий глобального потепления климата, для разных стран объективно существуют региональные приоритеты и если говорить о России, то таким приоритетом в рассматриваемой проблеме, несомненно, являются климатические изменения в арктическом регионе.

Литература

1. Головко В.А. Энергетические аспекты изменения климата Земли: взгляд из космоса. // Сб. «Современные проблемы дистанционного зондирования Земли из космоса», – М.: ООО «ДоМира», 2012 , т.9., №5 , с.140-154.

2. Thakur A., Krohn-Grimberghe A. AutoCompete: A Framework for Machine Learning Competitions. // ICML 2015 AutoML Workshop.