Исследование важности входных признаков при прогнозировании геомагнитного индекса алгоритмами машинного обучения

Р.Д. Владимиров
Московский государственный университет имени М.В. Ломоносова
Москва, Россия
vladimirov.rd16@physics.msu.ru

В.Р. Широкий
Московский государственный университет имени М.В. Ломоносова НИИ ядерной физики имени Д.В. Скобельцына
Москва, Россия
shirokiy@srd.sinp.msu.ru

О.Г. Баринов
Московский государственный университет имени М.В. Ломоносова НИИ ядерной физики имени Д.В. Скобельцына
Москва, Россия
obar@sinp.msu.ru

И.Н. Мягкова
Московский государственный университет имени М.В. Ломоносова НИИ ядерной физики имени Д.В. Скобельцына
Москва, Россия
irina@srd.sinp.msu.ru

С.А. Доленко
Московский государственный университет имени М.В. Ломоносова НИИ ядерной физики имени Д.В. Скобельцына
Москва, Россия
dolenko@srd.sinp.msu.ru

Аннотация—Одним из эффективных инструментов для прогнозирования временных рядов является использование методов машинного обучения, в частности, искусственных нейронных сетей. Однако при этом необходимым этапом исследования является понижение размерности входных данных. В данной работе рассматриваются результаты понижения размерности данных на основе ранжирования входных признаков по их существенности при решении задачи прогнозирования геомагнитного индекса Dst. Для оценки относительной существенности признаков используется итеративный подход, связанный с перебором моделей-кандидатов путём отбрасывания признаков по одному.

Ключевые слова— многомерный временной ряд, прогнозирование, отбор существенных признаков, искусственные нейронные сети, магнитосфера Земли, геомагнитный индекс Dst.

1. Введение

Геомагнитные возмущения представляют собой один из существенных факторов космической погоды, который с развитием космической отрасли становится все более важным. Прогнозирование геомагнитных возмущений, представляет интерес, так как сильные возмущения (магнитные бури) могут стать причиной нарушений в работе телеграфных линий и радиосвязи, трубопроводов, линий электропередач и энергосетей [1]. Также они опосредованно оказывают влияние на радиационные условия в космическом пространстве, поскольку после примерно половины магнитных бурь возрастает поток релятивистских электронов внешнего радиационного пояса Земли, что сбоям в электронных микросхемах спутниковой аппаратуры.

Состояние магнитосферы Земли характеризуется геомагнитными индексами. Одним из наиболее используемых является индекс Dst. Магнитосфера Земли и представляет собой динамическую систему, будущее состояние которой зависит не только от её текущего состояния и от текущего воздействия со стороны солнечного ветра, но и от предыстории. Это приводит к тому, что размерность используемых для прогнозирования данных, описывающих текущее состояние магнитосферы и предысторию на необходимую глубину, оказывается достаточно высока.

Одним из эффективных инструментов для прогнозирования временных рядов является использование методов машинного обучения (МО), в частности, искусственных нейронных сетей [2, 3]. Однако при этом высокая размерность входных данных может приводить к нежелательным последствиям, таким, как высокая вычислительная стоимость обучения и переучивание. Также отбор существенных входных признаков (ВП) может позволить сделать некоторые выводы о взаимосвязях различных физических величин, значения которых используются в качестве ВП.

Целью настоящей работы являлось применение адаптивного метода ранжирования ВП по их существенности на основе их отбрасывания по одному, а также оценка полученных результатов.

2. Входные данные и описание алгоритма

В качестве входных признаков использовались исторические значения характеристик солнечного ветра (скорости SW_spd и плотности H_den) и межпланетного магнитного поля (компонент By и Bz в системе GSM и модуля вектора поля |B|), измеренные в точке Лагранжа L1 в система Солнце-Земля, и значения самого прогнозируемого индекса Dst, а также временные характеристики, описывающие фазу вращения Земли вокруг Солнца и вокруг своей оси. Каждый параметр, кроме временных характеристик, описывался 24 среднечасовыми значениями: текущим и 23 историческими значениями за последние сутки (всего 6 признаков * 24 часа + 4 = 148 ВП). Использовались данные за период с октября 1997 по конец 2016 года (тренировочный набор, из которого 20% случайно

выделялись в валидационный набор) и с начала 2017 по январь 2020 г. (тестовый набор).

Для ранжирования ВП по существенности использовался итеративный подход на основе метода МО, связанный с перебором моделей-кандидатов путём удаления ВП по одному. В данной работе для построения моделей использовалась линейная регрессия (ЛР). Общая схема работы алгоритма выглядит так:

1) Обучается модель на данных с полным набором из n ВП, и производится оценка качества модели на тестовом наборе данных.
2) Каждый ВП по очереди удаляется из набора.
3) На каждом из n уменьшенных наборов производится обучение модели, и производится оценка изменения точности по сравнению с моделью, полученной на шаге 1.
4) Тот ВП, удаление которого приводило к максимальному падению (или минимальному увеличению) ошибки прогноза по сравнению с остальными моделями, полученными на шаге 3, окончательно удаляется из набора. Ошибка прогноза полученной модели фиксируется.
5) Алгоритм повторяется с шага 1, используя полученный на шаге 4 набор в качестве полного.
6) Считается, что признаки ранжированы по важности в порядке, обратном порядку удаления: самый важный признак остается последним.

На основании зависимости среднеквадратичного отклонения прогноза модели шага 4 на тестовом наборе от количества ВП в данной модели определяется оптимальный набор признаков на основе линейной регрессии, соответствующий минимуму построенной зависимости.

Преимуществом используемого метода ЛР является его низкая вычислительная стоимость. Однако его недостатком является линейность, в то время как аппроксимируемое отображение является нелинейным. Для компенсации этого недостатка дополнительно осуществлялось построение нейросетевых моделей на отобранных наборах ВП с количеством ВП от 3 до 25. Все используемые нейронные сети представляли собой многослойные персептроны с 1 скрытым слоем и 32 нейронами в нем. Для предотвращения переобучения использовался метод ранней остановки – обучение прекращалось спустя 500 эпох без улучшения результата на валидационном наборе. Для уменьшения влияния фактора, связанного с влиянием начальной инициализации весов на обучение, для каждого рассматриваемого случая обучалось по 5 НС, а статистические показатели их применения усреднялись.

3. Результаты

На Рис.1 показана зависимость среднеквадратичной ошибки прогнозирования на тестовом наборе данных для горизонта прогноза 1 час от номера набора (количества отброшенных ВП). Видно, что более 90% ВП могут быть исключены из рассмотрения без изменения качества прогноза. Аналогичная картина наблюдается для всех горизонтов от 1 до 6 часов.

Анализ наборов ВП, оставшихся в качестве наиболее существенных после осуществления процедуры отбора, показывает, что отобранные признаки согласуются с физическими представлениями (Рис.2).

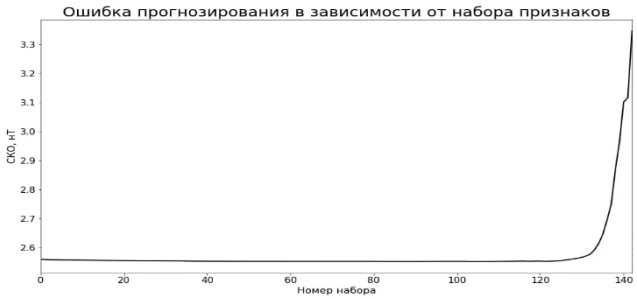

Рис. 1. Среднеквадратичная ошибка прогнозирования на тестовом наборе данных для горизонта прогноза 1 час в зависимости от номера набора (количества отброшенных ВП)

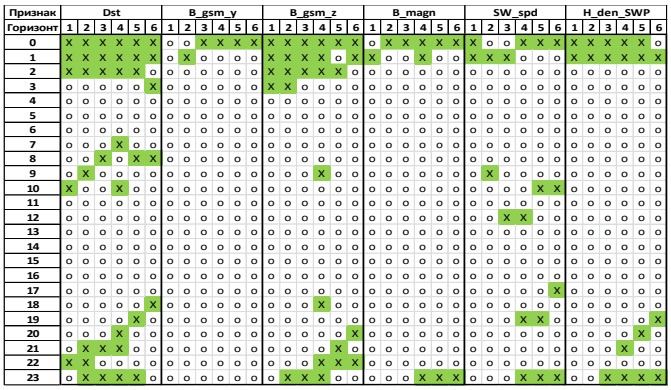

Рис. 2. Отобранные существенные признаки для горизонта прогноза от 1 до 6 часов. По вертикали – глубина задержки в часах

4. Выводы

- В результате применения ранжирования существенных входных признаков были отобраны оптимальные наборы ВП для горизонтов прогноза от 1 до 6 часов. Результаты согласуются с физическими представлениями о группах наиболее существенных ВП.

- Также удалось сократить размерность входных данных в среднем в 7.6 раз без потери качества прогнозирования. Добиться существенного улучшения качества прогнозирования путем отбора существенных ВП с помощью использованного алгоритма не удалось.

- Работы следует продолжить с использованием для построения моделей при отборе признаков нелинейных алгоритмов – градиентного бустинга или многослойных персептронов.

- Отобранные комплекты существенных ВП могут быть зафиксированы и использоваться в дальнейшем в работах по усовершенствованию алгоритмов и методик прогнозирования.

Литература

[1] Лазутин, Л.Л. Мировые и полярные магнитные бури /Л.Л. Лазутин. – М.: МГУ, 2012. – 214 с.

[2] Revallo, M. A neural network Dst index model driven by input time histories of the solar wind-magnetosphere interaction / M. Revallo, F. Valach, P. Hejda, J. Bochníček // J. Atmos. Sol. Terr. Phys. – 2014. – Vol. 110-111. – P. 9-14. DOI: 10.1016/j.jastp.2014.01.011.

[3] Ефиторов, А.О. Прогнозирование Dst-индекса, основанное на методах машинного обучения / А.О. Ефиторов, И.Н. Мягкова, В.Р. Широкий, С.А. Доленко // Косм. иссл. – 2018. – Т. 56, № 6. – С. 353-364. DOI: 10.1134/S0010952518060035.