НЕЙРОСЕТЕВОЙ АЛГОРИТМ ПРОГНОЗИРОВАНИЯ СОБЫТИЙ В МНОГОМЕРНЫХ ВРЕМЕННЫХ РЯДАХ И ЕГО ПРИМЕНЕНИЕ ДЛЯ АНАЛИЗА КОСМОФИЗИЧЕСКИХ ДАННЫХ

Ю.С.Шугай, С.А.Доленко, И.Г.Персианцев, Ю.В.Орлов

Научно-исследовательский институт ядерной физики им. Д.В.Скобельцына
МГУ им. М.В.Ломоносова
Россия, 119992, Москва, Воробьевы горы, МГУ, НИИЯФ
тел. +7 (095) 939-46-19
E-mail: jshugai@srd.sinp.msu.ru

Многие практические задачи связаны с поиском взаимосвязи между поведением сложного объекта и сравнительно редкими событиями, инициируемыми этим поведением или коррелирующими с ним. В таких случаях можно предполагать, что возникновению события всякий раз предшествует некоторое явление – некая комбинация значений признаков, описывающих объект, в известном диапазоне задержек по времени. В данной работе авторы продолжают исследование предложенного ими ранее нейросетевого метода анализа таких объектов. Целью применения метода является обнаружение морфологических и динамических признаков, вызывающих событие или предшествующих его возникновению.

Описание алгоритма

Задача поиска корреляционных связей в многомерных временных рядах, как одна из задач анализа пространственно-временных образов, представляется весьма актуальной [3]. В работах [1,5] авторами был предложен метод анализа многомерных временных рядов с целью прогнозирования наступления тех или иных *событий* и поиска их предвестников – *явлений*, представляющих собой некую неизвестную комбинацию значений параметров, описывающих объект. Алгоритм основан на использовании комитета нейронных сетей, обучаемых на различных участках анализируемого временного ряда. Существенной особенностью разработанного алгоритма являются возможность поиска нелинейных связей между событием и явлением.

Рассмотрим подробнее основные принципы исследуемого алгоритма. Анализируемый *интервал поиска* разбивается на перекрывающиеся *сегменты* длиной, равной т.н. *времени инициации*. Взаимное расположение соседних сегментов одинаково и характеризуется *интервалом перекрытия*. Время инициации оценивается из априорных соображений и должно превышать длительность искомого явления. Для каждого сегмента строится отдельная нейросеть, обучающаяся прогнозировать событие на основе признаков в данном сегменте. Во время тренировки интервал поиска сдвигается по оси времени. Когда правая граница интервала поиска оказывается на расстоянии минимально возможной задержки явления от момента наступления события, желаемый выход нейросетей для всех сегментов устанавливается равным 1. Во всех остальных случаях желаемый выход равен нулю. Таким образом, разные нейросети комитета, при совпадении желаемых выходов, получают в качестве входной информации каждая свой участок анализируемого многомерного временного ряда. При этом те из сетей, на чьих участках окажется предвестник (или хотя бы его часть), будут обучаться более эффективно, чем сети, информация на входах которых никак не связана с событием (последние, скорее всего, не будут обучаться вовсе). По окончанию обучения можно сделать вывод о том, что искомое явление (предвестник события) находится на участке той сети, которая по

результатам обучения обеспечивает наиболее точный прогноз события (желательно, на независимых данных). Сдвигая интервал поиска по оси времени и применяя набор нейросетей к соответствующим сегментам анализируемого ряда, мы можем прогнозировать наступление события. При этом задержка между явлением и событием оказывается определённой с точностью не хуже времени инициации. Дальнейшее уточнение значения задержки может быть достигнуто путём изменения интервала перекрытия сегментов, соответствующих разным сетям.

По результатам тестирования на модельных задачах [4] в алгоритм были внесены изменения. В качестве критерия остановки при обучении нейронной сети обычно используется условие достижения минимума средней ошибки на тестовом наборе данных. Такой критерий здесь использовался отдельно для каждой из сетей; при этом под минимумом средней ошибки на тестовом наборе понимается минимум коэффициента множественной детерминации (R-квадрат) для прогнозирования события на тестовом (в противоположность тренировочному) участке временного ряда. При применении такого критерия все нейронные сети обучаются в течение разного количества эпох, в результате чего не используется дополнительное преимущество, состоящее в том, что сети, входные данные для которых содержат явление, обучаются быстрее. Для повышения контрастности общей картины на выходах нейронных сетей системы в качестве критерия остановки обучения использовалось и другое условие - достижение максимума дисперсии (соответствующего максимуму энтропии) коэффициентов, характеризующих степень соответствия ответа каждой из нейронных сетей желаемому выходу. В качестве последних могут использоваться линейные коэффициенты корреляции, коэффициенты множественной детерминации (R-квадрат) и др. При применении этого критерия все нейронные сети обучаются параллельно, т.е. в течение одного и того же количества эпох. Заметим, что в ситуации, когда количество данных (известных событий) невелико и выделение тестового набора невозможно или нежелательно, этот критерий можно использовать и на тренировочном наборе, не опасаясь "переучивания" сетей.

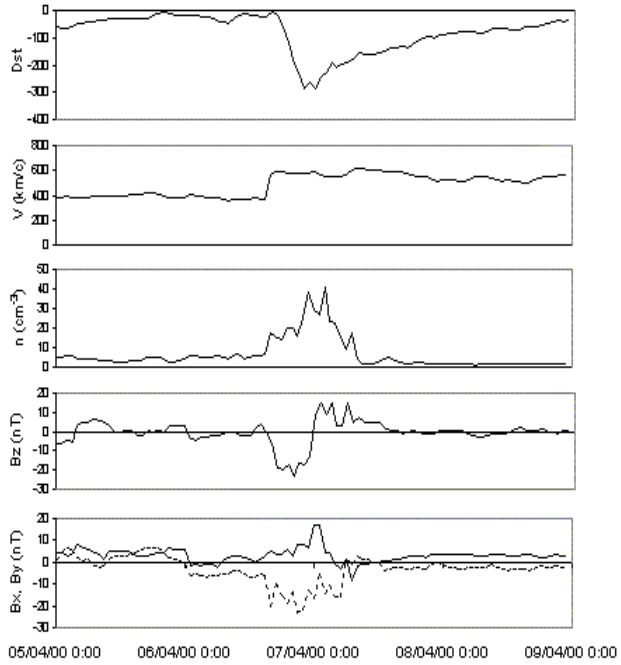
В работе [2] были проанализированы параметры нейронных сетей. Было установлено, что снижение скорости обучения нейронных сетей с 0.05 до 0.001 не приводит к существенному изменению точности прогнозирования. Коэффициент R-квадрат для соответствия ответов нейронных сетей прогнозируемым событиям на экзаменационном наборе остается на уровне 0.8.

На модельных задачах [2] было проверено, как влияют на точность прогнозирования события различные критерии остановки обучения, описанные выше. Сравнивались два критерия остановки обучения нейронных сетей: при достижении максимального значения R-квадрат каждой из сетей на тестовом наборе ("остановка по R-квадрат"); при достижении максимальной дисперсии коэффициентов R-квадрат для разных сетей на тестовом наборе ("остановка по дисперсии"). Полученные результаты показывают, что при остановке тренировки нейронных сетей "по R-квадрат" точность прогнозирования события повышается, а контрастность картины на выходе системы понижается по сравнению с "остановкой по дисперсии".

Космофизические данные

Прогнозировалось событие - начало геомагнитной бури со значениями геомагнитного индекса (см. [6]) Dst < -80 nT. Часовые значения Dst-индекса были взяты с WDC-C2 KYOTO (http://swdcwww.kugi.kyoto-u.ac.jp/dstdir/dst1/final.html). Известно, что развитие геомагнитной бури зависит в первую очередь от двух параметров солнечного ветра: Bz–компоненты межпланетного магнитного поля и скорости солнечного ветра (V). Параметры солнечного ветра регистрировались космическим кораблем ACE (http://www.srl.caltech.edu/ACE/), находящимся в точке гравитационного равновесия между Солнцем и Землей.

Солнечный ветер достигает Земли приблизительно через час после его регистрации спутником ACE.



Параметры геомагнитной активности и солнечного ветра во время геомагнитной бури 7 апреля 2000 года (dst = -288 nT). Сверху вниз: Dst индекс геомагнитной активности; скорость солнечного ветра V; плотность плазмы солнечного ветра n; Bz–компонента межпланетного магнитного поля; другие компоненты межпланетного магнитного поля (By – пунктирная линия, Bx – сплошная).

В качестве входных данных были взяты значения скорости и Bz–компоненты магнитного поля солнечного ветра за 2000-2002 годы. Данные были разбиты на тренировочный набор (2000 год и первая половина 2001), тестовый набор (вторая половина 2001 года) и экзаменационный набор (2002 год). Интервал поиска равнялся 36 часам. Время инициации равнялось 8 часам, интервал перекрытия - 4 часам. Система автоматически определила, что значения параметров солнечного ветра, приводящие к геомагнитной буре, регистрируются спутником за 8 и меньше часов до начала геомагнитной бури с Dst < -80 nT, для большинства исследуемых геомагнитных событий. В то же время результаты, полученные на экзаменационном наборе, показали, что данные недостаточно представительны для надёжного прогнозирования геомагнитных бурь на разных фазах солнечного цикла.

Хотя Bz–компонента магнитного поля является наиболее значимым параметром солнечного ветра, другие параметры также могут влиять на силу и продолжительность геомагнитной бури [6]. Поэтому в новый набор входных данных были включены, кроме скорости и Bz–компоненты магнитного поля, еще и Bx- и By-компоненты магнитного поля и плотность плазмы солнечного ветра. Использовались данные за 2000-2003 годы. Данные были разбиты на тренировочный набор (2000 и 2002 годы), тестовый набор (2001 год) и экзаменационный набор (2003 год). Интервал поиска равнялся 24 часам. Время инициации равнялось 8 часам, интервал перекрытия – 4 часам. Полученные предварительные результаты показали, что система по-прежнему правильно определяет задержку между явлением, являющимся предвестником геомагнитной бури, и самой бурей (событием); при этом уверенность ответов сети (значение коэффициента R-квадрат для соответствия ответов нейронных сетей прогнозируемым событиям) выросло по сравнению с предыдущим экспериментом. Увеличился и процент правильных прогнозов, оставаясь, однако, недостаточным для надёжного прогнозирования событий (было правильно спрогнозировано менее 50% событий). Тем не менее, полученные результаты свидетельствуют о том, что искомое явление (являющееся предвестником события – геомагнитной бури) может быть более адекватно описано расширенным набором входных признаков, использованным во втором эксперименте.

Продолжение исследования возможностей алгоритма в применении к решению задачи о прогнозировании геомагнитных бурь по данным спутника ACE планируется в двух направлениях. Дальнейшее увеличение размерности рассматриваемого временного ряда (например, учёт температуры солнечной плазмы или абсолютной величины межпланетного магнитного поля) может привести к увеличению процента правильных прогнозов; в то же время необходимо учитывать тот факт, что излишнее количество входных переменных

затрудняет обучение сетей и делает его более длительным. Вторая задача будет заключаться в том, чтобы попытаться понять, какие именно из входных переменных необходимы для описания явления, а какие можно исключить без ущерба для решения задачи прогнозирования.

Помимо этого, планируется также дальнейшее усовершенствование самого алгоритма. В частности, планируется повысить точность определения задержки между явлением и событием путем автоматической подстройки интервала перекрытия и времени инициации.

Заключение

В работе продемонстрировано применение предложенного ранее авторами алгоритма анализа временных рядов для решения реальной задачи анализа космофизических данных. Было показано, что алгоритм правильно определяет задержку между явлением и событием, однако надёжность прогнозирования геомагнитных бурь по данным спутника ACE с помощью разрабатываемого алгоритма оказалась недостаточной. Показано также, что увеличение размерности временного ряда путём добавления значимых входных переменных позволяет повысить уверенность ответов сети и несколько увеличить надёжность прогнозирования. В дальнейшем планируется продолжение работ по анализу космофизических данных. Планируется также усовершенствование алгоритма с целью повышения точности определения задержки между явлением и событием.

Работа выполнялась при финансовой поддержке следующих организаций: Российский фонд фундаментальных исследований (РФФИ), проект № 04-01-00506; фонд "Научный потенциал" (Human Capital Foundation), проект № 23-03-70.

Список литературы

1. С.А.Доленко и др. Нейросетевой поиск корреляционных связей во временных рядах. 6-я международная конференция "Распознавание образов и анализ изображений: Новые информационные технологии" (РОАИ-6-2002), Великий Новгород, Россия, 21-26 октября 2002. Труды конференции. Великий Новгород, 2002, т.1, с.198-202.
2. Ю.С.Шугай и др. Нейросетевые алгоритмы прогнозирования событий и поиска предвестников в многомерных временных рядах. Искусственный Интеллект, Донецк, 2004, № 2, с.211-215.
3. J.B.D.Cabrera, K.R.Mehra. Extracting precursor rules from time series, – A Classical Statistical Viewpoint. Proc. 2nd SIAM Int. Conf. on Data Mining (SDM-2002). April 11-13, 2002, Hyatt Regency, Crystal City at Ronald Reagan National Airport, Arlington, VA.
4. S.A.Dolenko et al. Pattern Recognition and Image Analysis, 2003, v.13, No.3, pp.441-446.
5. Yu.V.Orlov et al. Nuclear Instruments and Methods in Physics Research Section A (NIMA A), 2003, v.502, No.2-3, pp.532-534.
6. Sh.Watanabe et al. J. Communications Research Laboratory, 2002, v.49, No.4, pp.69-85.