УДК 519.9

АДАПТИВНОЕ ПРОГНОЗИРОВАНИЕ МНОГОМЕРНОГО ВРЕМЕННОГО РЯДА

© Неделько С.В.

Институт математики СО РАН
Лаб. Анализа данных
пр-т Коптюга, 4, г. Новосибирск, 630090, Россия

e-mail: *nedelko@math.nsc.ru*

Abstract. The method of heterogeneous multidimensional time series adaptive prediction is considered. This method is based on adaptive forming a discrete states set and uses some kind of informativity criterion. The method of adaptation the deciding function to changing of probabilistic properties of the stochastic process is also offered[4].

Введение

Прогнозирование многомерных разнотипных временных рядов является хорошо известной задачей анализа данных. Несмотря на немалое количество работ, посвященных данной проблеме, к настоящему времени остается ряд нерешенных вопросов. Прежде всего, особенностью задачи является наличие нескольких целевых переменных, что не отражено в большинстве алгоритмов прогнозирования, когда решающая функция строится для каждой переменной отдельно, без учета их зависимостей [2]. Такой подход в случае сильной зависимости целевых переменных проигрывает в качестве прогноза [3].

Обычно прогнозирование временного ряда подразумевает построение решающей функции, которая по заданной предыстории ряда дает прогнозируемый набор значений переменных ряда для следующего момента времени. Размерность пространства, в котором ведется поиск решающей функции, увеличивается с ростом глубины предыстории, что требует использования различных эвристик либо упрощения класса решающих функций. Кроме того, построение решения в пространстве большой размерности сказывается на быстродействии работы алгоритма.

В настоящей работе рассматривается алгоритм прогнозирования многомерного разнотипного временного ряда, основанный на адаптивном формировании пространства состояний [4] ряда в классе логических решающих функций [2]. Оптимальное разбиение при этом ищется непосредственно в пространстве переменных, описывающих временной ряд, без предварительного разделения на прогнозирующие переменные (пространство значений X) и целевые переменные (пространство значений Y), что существенно снижает трудоемкость алгоритма. Такое разбиение исходного пространства переменных учитывает их зависимость и решает проблему роста размерности с увеличением глубины предыстории.

Метод оценивания условного распределения в заданном классе кусочно-постоянных распределений основан на критерии информативности и не требует

[4]Работа выполнена при поддержке РФФИ, проект № 07-01-00331-а

каких-либо метрических свойств пространства Z. Классические же непараметрические методы оценивания условного распределения предполагают те или иные метрические свойства пространства Z, что в разнотипном случае не вполне оправдано.

Предложен метод адаптации решающей функции к изменению вероятностных свойств процесса, основанный на критерии минимума энтропии.

Задача обнаружения изменения вероятностных свойств случайного процесса (задача о разладке) известна давно и до настоящего времени активно исследуется [1]. Особенность данной работы заключается в использовании достаточно универсальной модели временного ряда.

1. Постановка задачи

Пусть дан n-мерный разнотипный временной ряд $v = \left\{z^t \,|\, t = \overline{1,N}\right\}$, $z^t = (z_1^t, \ldots, z_n^t)$, $z_j^t \in Z_j$. Множество Z_j является множеством допустимых значений j-й переменной ряда. В силу разнотипности ряда в наборе переменных могут присутствовать одновременно непрерывные и дискретные переменные, а также переменные с упорядоченным и неупорядоченным множеством значений. Пространство значений ряда обозначим $Z = \prod\limits_{j=1}^{n} Z_j$. Требуется дать прогноз значений временного ряда в моменты времени $t > N$ на основе анализа имеющихся эмпирических данных, то есть реализации v.

Рассматриваем статистическую постановку задачи, когда v является реализацией некоторого случайного процесса $z(t)$ с дискретным временем. Предполагаем, что процесс задан переходной (условной) вероятностной мерой $P\left[Z|z(t-1), z(t-2), \ldots, z(t-d)\right]$, определяемой предысторией длины d. Квадратные скобки здесь означают, что имеется в виду не мера множества Z, а мера, заданная на некоторой σ–алгебре его подмножеств.

2. Критерии качества

Критерии качества вероятностной модели временного ряда основаны на понятии информативности распределений [6], то есть степени отличия от априорного распределения [7].

Зафиксируем некоторое разбиение $\lambda = \left\{E^\omega \subseteq Z | \omega = \overline{1,k}\right\}$, $\prod\limits_{\omega=1}^{k} E^\omega = Z$, $\omega \neq \overline{\omega} \Rightarrow E^\omega \cap E^{\overline{\omega}} = \varnothing$, пространства Z. Это позволяет исходному многомерному ряду v сопоставить одномерную символьную последовательность $w = \left\{\omega^t | z^t \in E^{\omega^t}, t = \overline{1,N}\right\}$. Тогда случайному процессу $z(t)$ будет соответствовать процесс $\omega(t)$, переходные вероятности для которого обозначим

$$p_{\omega_0|\omega_1,\omega_2,\ldots,\omega_d} = P(\omega(t) = \omega_0 / \omega(t-1) = \omega_1, \ldots, \omega(t-d) = \omega_d).$$

Аналогично вводится совместная вероятность

$$p_{\omega_0 \ldots \omega_d} = P\left(\bigwedge_{\tau=0}^{d} (\omega(t-\tau) = \omega_\tau)\right) = P\left(\bigwedge_{\tau=0}^{d} (Z(t-\tau) \in E^{\omega_\tau})\right),$$

то есть вероятность появления заданной предыстории длины d.

Критерий информативности определим как

$$K(\lambda) = \sum_{\omega_0=1}^{k} \dots \sum_{\omega_d=1}^{k} \left| p_{\omega_0} - p_{\omega_0|\omega_1,\omega_2,\dots,\omega_d} \right| \cdot p_{\omega_1\omega_2\dots\omega_d}.$$

Он представляет собой средний модуль разности между вероятностями перехода и безусловными вероятностями нахождения в состояниях.

После тождественных преобразований имеем:

$$K(\lambda) = \sum_{\omega_0=1}^{k} \dots \sum_{\omega_d=1}^{k} \left| p_{\omega_0\omega_1\dots\omega_d} - p_{\omega_0} p_{\omega_1\dots\omega_d} \right|.$$

Здесь $p_{\omega_1\dots\omega_d} = \sum\limits_{\omega_0=1}^{k} p_{\omega_0\omega_1\dots\omega_d}$, $p_{\omega_0} = \sum\limits_{\omega_1=1}^{k} \dots \sum\limits_{\omega_d=1}^{k} p_{\omega_0\omega_1\dots\omega_d}$.

При длине предыстории $d = 1$ критерий принимает вид

$$K(\lambda) = \sum_{\omega_0=1}^{k} \sum_{\omega_1=1}^{k} \left| p_{\omega_0\omega_1} - p_{\omega_0 p_{\omega_1}} \right|.$$

Также будем использовать критерий, основанный на энтропии:

$$K_e(\lambda) = \sum_{\omega_0=1}^{k} \dots \sum_{\omega_d=1}^{k} p_{\omega_0|\omega_1,\omega_2,\dots,\omega_d} \ln\left(p_{\omega_0|\omega_1,\omega_2,\dots,\omega_d}\right) \cdot p_{\omega_1\omega_2\dots\omega_d} - \sum_{\omega_0=1}^{k} p_{\omega_0} \ln(p_{\omega_0}).$$

Данный критерий будет использован для обнаружения неоднородности процесса (изменения вероятностных свойств).

Для оценки введенных критериев по выборке достаточно заменить $p_{\omega_0\dots\omega_d}$ на $N_{\omega_0\dots\omega_d}/N$ – частоту реализации предыстории на обучающей последовательности v.

3. Алгоритм прогнозирования и оценка значимости закономерностей

Приведем краткое описание алгоритма адаптивного прогнозирования многомерного разнотипного временного ряда на основе выбора пространства состояний случайного процесса. Применяется направленный поиск (алгоритм LRP) для нахождения разбиения λ пространства переменных временного ряда. При этом приближенное к оптимальному разбиение строится максимизацией критерия информативности K. Алгоритм находит решение в виде дерева или непересекающихся многомерных интервалов. Полученные области из разбиения соответствуют состояниям случайного процесса, описывающего исходный многомерный временной ряд. Аппроксимация переходной вероятности случайного процесса дает возможность делать прогноз значений ряда.

Оценивание статистической значимости полученного алгоритмом решения основано на методе статистического моделирования. Предлагается исследование поведения алгоритма на случайных перестановках по времени значений временного ряда. Строится эмпирическая функция распределения для критерия информативности

матрицы переходных вероятностей и ее аппроксимация соответствующим распределением, например, методом моментов. Значение критерия информативности, полученное на реальной последовательности событий, может лежать в области вероятных значений критерия на перестановках, что объясняет полученные закономерности случайными флуктуациями. Метод опробован при нахождении закономерностей в сейсмических данных [5]. Закономерности, заключенные в относительных переходных вероятностях для состояний процесса, оказались статистически значимыми.

4. Обнаружение изменения свойств процесса

Построение логико-вероятностной модели временного ряда на основе выбора пространства состояний процесса можно также применять для обнаружения изменения вероятностных свойств. Это важно в случаях, когда исходный временной ряд составлен из реализаций нескольких случайных процессов. Требуется найти моменты времени, когда происходит изменение модели ряда.

Зафиксируем некоторый момент времени t_0, начиная с которого будем последовательно проверять соответствие фрагмента ряда выбранной модели. Временной ряд разбивается выбранным моментом на две части. Первая часть используется для построения модели, то есть для построения пространства состояний случайного процесса и оценивания матрицы переходных вероятностей между состояниями.

В начале второй части ряда, не задействованной в построении модели, выделяем некоторый фрагмент длины r, соответствующий моментам времени $t_0 + 1$, ..., $t_0 + r$, и вычисляем для него частоты переходов между состояниями. Далее проверяем согласованность частот с построенной моделью.

В случае близости частот переходов переходным вероятностям считаем, что до момента времени $t_0 + r$ ряд определяется одной и той же моделью. Иначе принимаем, что в момент t_0 происходит изменение вероятностных свойств случайного процесса.

Если смены процесса не произошло, то проводим аналогичное исследование для следующего фрагмента ряда длины r. По его результатам также решаем, происходит смена модели в момент времени $t_0 + r$ или нет. Так можем продолжить до конца исходного ряда.

Если же в момент t_0 обнаружено изменение случайного процесса, определяющего ряд, то поступаем со второй частью ряда тем же образом, что и с исходным рядом. А именно, разбиваем остаток ряда на две части, первую используем для аппроксимации переходной вероятности, а вторую последовательно проверяем на соответствие второй построенной модели. Действуя таким образом, находим моменты времени (а их может быть несколько), в которые происходит изменение вероятностных свойств случайного процесса, а также сами оцененные матрицы переходных вероятностей.

При использовании предложенного метода могут возникнуть следующие трудности. Предполагается, что фрагмент ряда до момента t_0 подчиняется одной вероятностной модели, то есть в начале ряда не происходит быстрого изменения модели. Это необходимо, чтобы длина усеченных данных была достаточной для обучения. В то же время не следует выбирать t_0 слишком большим, рискуя пропустить возможный момент разладки между рядом и моделью. Нахождение оптимального значения параметра r также представляет собой нетривиальную задачу.

5. Выявление моментов разладки на основе критерия минимума энтропии

Определение моментов изменения вероятностных свойств процесса можно интерпретировать как расстановку на всем временном интервале границ интервалов постоянства модели. При этом важен выбор критерия расстановки.

Рассмотрим сначала простой иллюстративный пример.

Пусть дан дискретный случайный процесс $z(t) \in \{1, 2\}$ с непрерывным временем $t \in [0, 1]$. При этом вероятности нахождения в каждом из двух возможных состояний не зависят от предыстории и постоянны в каждом из двух временных интервалов: $P\left(z(t) = 1\right) \equiv p_1$, при $t \in [0, \alpha)$, и $P\left(z(t) = 1\right) \equiv p_2$, при $t \in [\alpha, 1]$, $0 < \alpha < 1$.

Таким образом, изменение свойств процесса происходит в точке α. Если изменение процесса в точке α не предполагать, то вероятность нахождения процесса в первом состоянии оценится как среднее на всем интервале $[0, 1]$, то есть $p = \alpha p_1 + (1-\alpha)p_2$.

В качестве критерия, позволяющего определить момент изменения свойств, может быть использована энтропия. При отсутствии границы энтропия есть $H = -p \ln p$. При правильно поставленной границе α энтропия уменьшится и составит $H = -(\alpha p_1 \ln p_1 + (1 - \alpha) p_2 \ln p_2)$. При этом минимум энтропии достигается при правильной постановке границы, при любых параметрах α, p_1, p_2.

Для предложенного алгоритма прогнозирования выполнение подобного свойства становится неочевидным, поскольку используемый в алгоритме энтропийный критерий устроен сложнее, кроме того, на разных интервалах ряда алгоритм строит разные множества состояний процесса. Однако проведенное статистическое моделирование показывает, что данный критерий позволяет находить момент разладки.

При этом критерий, основанный на отличии от априорных вероятностей, не находит момент разладки, если для одного из процессов переходные вероятности близки к априорным.

6. Пример работы алгоритма

Пусть временной ряд $(Z_1(t), Z_2(t))$ задается случайным процессом, у которого в некоторый момент изменяются свойства. Фактически имеем два случайных процесса, а временной ряд есть соединение их реализаций. Первый процесс выделяет в пространстве Z три области: $[0,\ 1] \times (0,65,\ 1]$, $(0,5,\ 1] \times [0,\ 0,65]$ и $[0,\ 0,5] \times [0,\ 0,65]$. Для второго процесса выделены $(0,65,\ 1] \times [0,\ 1]$, $[0,\ 0,65] \times [0,\ 0,5]$ и $[0,\ 0,65] \times (0,5,\ 1]$. Области состояний процессов наглядно изображены на рис. 1. Стрелками показаны наиболее вероятные переходы.

Переходы между областями для каждого процесса определяются матрицами переходных вероятностей. В примере обе матрицы имели вид

$$p_{j|i} = \begin{pmatrix} \delta & 1-2\delta & \delta \\ \delta & \delta & 1-2\delta \\ 1-2\delta & \delta & \delta \end{pmatrix},$$

где i соответствует строке, а j – столбцу.

В пределах каждой области условные распределения равномерны.

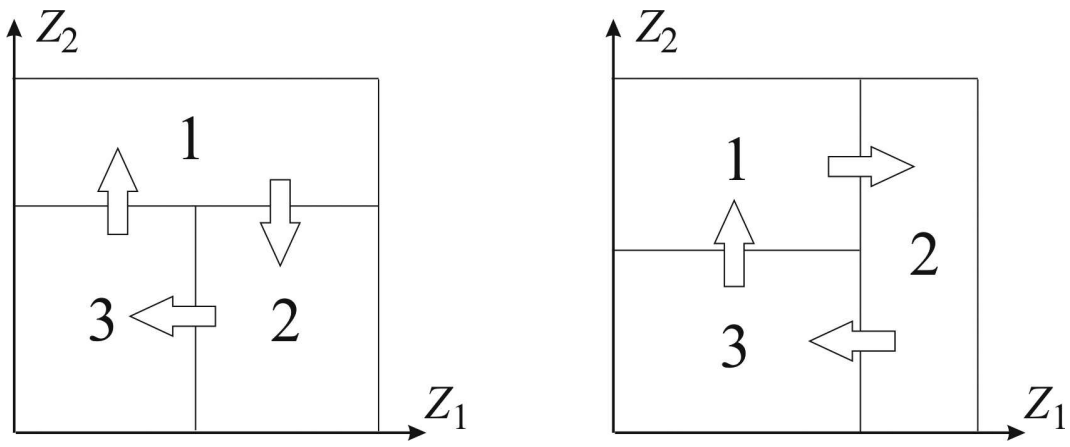

Рис. 1. Модели процесса для различных временных интервалов.

Параметр δ задается для каждого процесса отдельно (δ_1 и δ_2) и позволяет варьировать степень выраженности закономерностей (при $\delta = \frac{1}{3}$ все переходы равновероятны и закономерности отсутствуют).

Заметим, что модели процессов на обоих временных интервалах намеренно выбраны похожими, чтобы проиллюстрировать возможности метода по их разделению.

В случае реализации одного случайного процесса, при $N = 200$, алгоритм поиска логических закономерностей с любым из критериев (K или K_e) практически точно восстанавливал разбиение, если $\delta < 0,2$.

В случае изменяющихся свойств процесса использование критерия K_e позволило определять момент разладки с точностью до нескольких отсчетов (при $N = 500$), если хотя бы один из δ_1 или δ_2 не превосходил $0,15$.

Заключение

В работе предложен метод прогнозирования многомерного разнотипного временного ряда с изменяющимися свойствами, основанный на выделении состояний процесса в классе логических решающих функций. В роли критерия качества модели используются различные варианты меры информативности матрицы переходных вероятностей. Исследование путем статистического моделирования показывает способность метода адекватно оценивать вероятностную модель временного ряда, а также обнаруживать момент изменения вероятностных свойств (разладки).

Список литературы

1. *Боровков А.А.* Асимптотические оптимальные решения в задаче о разладке. // Теория вероятностей и ее применения, 1998, Т.43, № 4, С. 625–654.
2. *Лбов Г.С., Старцева Н.Г.* Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999. 211 с.
3. *Миренкова С.В.* Метод прогозирования многомерного разнотипного временного ряда в классе логических решающих функций. // Искусственный интеллект. Изд-во НАН Украины, 2002, № 2. С. 197–201.

4. *Неделько С.В.* Критерий информативности матрицы переходов и прогнозирование разнотипного временного ряда. // Искусственный интеллект. Изд-во НАН Украины, 2004, № 2. С. 145–149.
5. *Неделько С.В., Ступина Т.А.* Построение логико-вероятностных моделей временного ряда при анализе сейсмических данных. // Научный вестник НГТУ. Новосибирск, 2007, № 4(29). С. 33–42.
6. *Lbov G.S., Nedelko V.M.* A maximum informativity criterion for the forecasting several variables of different types. // Proceedings of the 6-th International Conference "Computer Data Analysis and Modeling". Minsk, 2001. P. 43–48. 1999. 211 c.
7. *Ростовцев П.С.* Алгоритм построения типологий для больших массивов социально-экономической информации. // Модели агрегирования социально-экономической информации. Сборник научных трудов. Изд-во ИЭ и ОПП СО АН СССР, 1978.

Статья поступила в редакцию 01.05.2008