ПРОГНОЗИРОВАНИЕ РАЗНОТИПНОГО ВРЕМЕННОГО РЯДА МЕТОДОМ АДАПТИВНОГО ФОРМИРОВАНИЯ ПРОСТРАНСТВА СОСТОЯНИЙ В КЛАССЕ ЛОГИЧЕСКИХ РЕШАЮЩИХ ФУНКЦИЙ[1]

Светлана Неделько

Abstract: The method of heterogeneous multidimensional time series probabilistic model reconstruction based on adaptive forming a discrete state set is offered. To estimate a deciding function quality some kind of informativity criterion for conditional distribution is used. The algorithm based on the proposed method was implemented and tested on model and applied tasks.

Keywords: multidimensional heterogeneous time series, pattern recognition, classification, statistical robustness, deciding functions, complexity.

ACM Classification Keywords: G.3 Probability and statistics: time series analysis, Markov processes, multivariate statistics, nonparametric statistics; G.1.6. Numerical analysis: Optimization.

Введение

Для многих методов прогнозирования многомерного разнотипного временного ряда имеет место проблема резкого увеличения размерности пространства, в котором ведется поиск решающей функции, при увеличении глубины предыстории, от которой зависит прогноз. В этой ситуации приходится либо упрощать класс решающих функций и уменьшать глубину предыстории, либо использовать различные эвристики и усложнять алгоритмы построения решения.

Кроме того, одна из особенностей задачи прогнозирования многомерного временного ряда состоит в том, что прогнозируется одновременно несколько переменных. В большинстве алгоритмов эта особенность не принимается во внимание, и решающие функции строятся для каждой переменной отдельно, без учета зависимости целевых переменных.

В настоящей работе рассматривается алгоритм прогнозирования многомерного разнотипного временного ряда, основанный на адаптивном формировании пространства состояний в классе логических решающих функций. Поиск оптимального с точки зрения критерия информативности разбиения исходного пространства переменных учитывает зависимость целевых переменных и снимает проблему роста размерности пространства, хотя и огрубляет прогноз.

Постановка задачи

Пусть дан n-мерный разнотипный временной ряд $v = \left\{ z^t \;\middle|\; t = \overline{1,N} \right\},\ z^t = \left(z_1^t, \ldots, z_n^t \right),\ z_j^t \in Z_j$. Здесь Z_j – множество допустимых значений j-й переменной ряда. В наборе переменных могут присутствовать

[1] Работа выполнена при поддержке РФФИ, грант 04-01-00858-а.

одновременно непрерывные и дискретные, а также переменные с упорядоченным и неупорядоченным множеством значений. Пространство значений ряда обозначим $Z = \sum_{j=1}^{n} Z_j$.

Задача состоит в том, чтобы на основе анализа имеющихся эмпирических данных v прогнозировать значения временного ряда в моменты времени $t > N$.

Будем рассматривать статистическую постановку задачи, когда v является реализацией некоторого случайного процесса $z(t)$ с дискретным временем. При этом предположим, что процесс задается переходной (условной) вероятностной мерой $P[Z/z(t-1), z(t-2), ..., z(t-d)]$, определяемой предысторией длины d. Квадратные скобки здесь и далее в аналогичных ситуациях означают, что имеется в виду не мера множества Z, а мера, заданная на некоторой σ-алгебре его подмножеств.

Обычно прогнозирование временного ряда подразумевает построение решающей функции $f : Z^d \to Z$, которая по заданной предыстории ряда для моментов $t-1,\ t-2, \ldots,\ t-d$ дает прогнозируемый набор значений переменных ряда для момента t. Мы будем строить аппроксимацию самой переходной вероятности. Этот подход имеет потенциально большую гибкость при оценивании модели, определяющей временной ряд. Знание переходной вероятности позволяет, в том числе, и строить оптимальный прогноз значений.

При этом, классические непараметрические методы оценивания условного распределения требуют относительно большого объема эмпирических данных и предполагают те или иные метрические свойства пространства Z, что в разнотипном случае не вполне оправдано.

В настоящей работе используется основанный на критерии информативности метод оценивания условного распределения в заданном классе кусочно-постоянных распределений.

Данный метод не требует каких-либо метрических свойств пространства Z и позволяет гибко подстраивать сложность модели под объем выборки.

Критерий качества вероятностной модели

Описываемый критерий основан на понятии информативности распределений. Под информативностью здесь понимается степень отличия от априорного распределения. Степень отличия может характеризоваться расстоянием в некоторой выбранной метрике (однако требование выполнения всех свойств метрики, вообще говоря, необязательно).

Зафиксируем некоторое разбиение $\lambda = \left\{ E^{\omega} \subseteq Z \;\middle|\; \omega = \overline{1,k} \right\},\ \bigcup_{\omega=1}^{k} E^{\omega} = Z ,\ \omega \neq \varpi \Rightarrow E^{\omega} \cap E^{\varpi} = \varnothing ,$

пространства Z.

Теперь исходному многомерному ряду v можно сопоставить одномерную символьную последовательность $w = \left\{ \omega^t \;\middle|\; z^t \in E^{\omega^t},\ t = \overline{1,N} \right\}$.

Случайному процессу $z(t)$ будет соответствовать процесс $\omega(t)$, переходные вероятности для которого обозначим

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

Также будем использовать совместную вероятность $p_{\omega_0 \ldots \omega_d} = \mathrm{P}\left(\bigwedge_{\tau=0}^{d} \left(\omega(t-\tau) = \omega_\tau\right) \right) = \mathrm{P}\left(\bigwedge_{\tau=0}^{d} \left(Z(t-\tau) \in E^{\omega_\tau}\right) \right)$ — вероятность заданной предыстории длины d.

Критерий информативности определим как

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

Данный критерий есть средний модуль разности между вероятностями перехода и безусловными вероятностями нахождения в состояниях.

После тождественных преобразований имеем:

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

Для оценки данного критерия по выборке достаточно заменить $p_{\omega_0 \ldots \omega_d}$ на $N_{\omega_0 \ldots \omega_d} / N$ – частоту реализации предыстории на обучающей последовательности w.

Описание алгоритма

Во многих распространенных алгоритмах прогнозирования временного ряда производится сведение реализации временного ряда к выборке в виде таблицы данных так называемым методом «гусеницы» или «змейки» [Данилов, Жиглявский, 1997]. Для этого вводится переобозначение переменных: прогнозируемые переменные обозначаются $y_j(t) = z_j(t),\ j = 1, \ldots, n$, а значения ряда на предыстории обозначаются $x_j(t) = z_j(t-1)$, $x_{j+n}(t) = z_j(t-2)$, ..., $x_{j+n(d-1)}(t) = z_j(t-d),\ j = \overline{1,n}$. После этого могут использоваться алгоритмы построения решающих функций на основе таблиц данных. Для случая разнотипного пространства независимых переменных известны соответствующие методы распознавания образов и регрессионного анализа в классе логических решающих функций [Лбов, Старцева, 1999].

Однако подобное сведение существенно (в d раз) увеличивает размерность пространства, в котором строится решающая функция.

В предлагаемом алгоритме разбиение строится непосредственно в пространстве Z, что позволяет избежать увеличения размерности и строить решения при относительно коротких реализациях

Для нахождения приближенного к оптимальному в соответствии с критерием K разбиения λ применим алгоритм направленного поиска (LRP), строящий решение в виде дерева [Лбов, Старцева, 1999] или в виде непересекающихся многомерных интервалов. Под интервалом понимается произвольное множество соседних значений для переменной с упорядоченными значениями и произвольное подмножество значений, если переменная с неупорядоченными значениями. Многомерный интервал представляет собой декартово произведение интервалов по переменным.

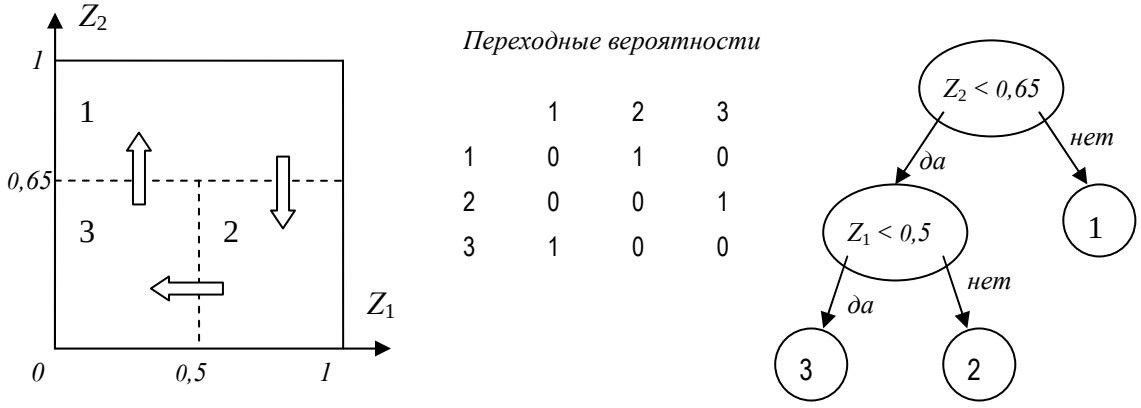

Рис. 1. Для тестового примера изображены соответствующие состояниям области в пространстве значений переменных, переходные вероятности и дерево решений, построенное предложенным алгоритмом.

Тестовый пример

Для иллюстрации работоспособности изложенного алгоритма, решим модельную задачу.

Пусть двумерный временной ряд задается случайным процессом с тремя состояниями. Области пространства, соответствующие состояниям, и вероятности переходов приведены на рис. 1. В каждой из областей при условии попадания в нее распределение равномерно.

Реализованный алгоритм построения разбиения на основе критерия информативности продемонстрировал правильное нахождение закономерностей, заложенных во временной ряд. Построенное им дерево решений изображено в правой части рисунка.

Решение прикладной задачи

Предложенный алгоритм был испытан на задаче анализа временного ряда, представленного метеорологическими данными.

Ряд включал за период с 1915 по 2000 гг. среднегодовые значения следующих величин:

Z_1 – температура воздуха;

Z_2 – объем осадков;

Z_3 – объем водостока.

Траектория данного ряда в проекции на плоскость Z_2, Z_3 приведена на рис. 2.

Для выявления закономерностей данного ряда с помощью описанного алгоритма было построено дерево решений с 6-ю конечными

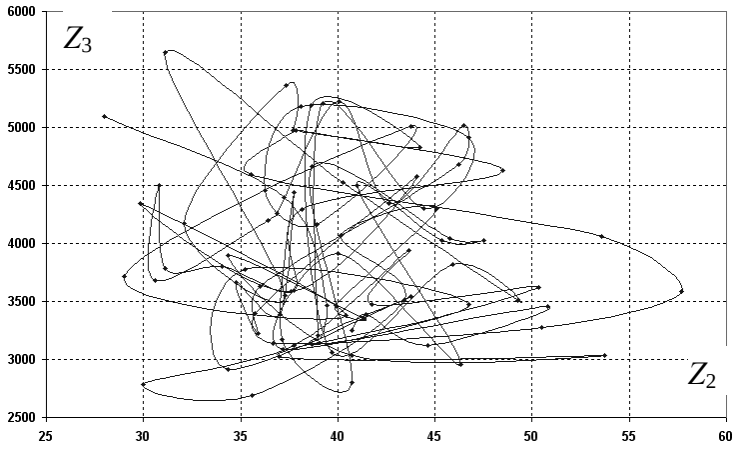

Рис. 2. Траектория временного ряда в пространстве переменных Z_2 и Z_3.

вершинами, которые определяют состояния случайного процесса. Глубина предыстории d = 1 .

Ниже полученное решение представлено в виде списка областей.

1: Z_3 **< 3960 &** Z_2 **< 37.8 &** Z_3 **< 3163**

2: Z_3 **< 3960 &** Z_2 **< 37.8 &** Z_3 **≥ 3163 &** Z_1 **< -0.46**

3: Z_3 **< 3960 &** Z_2 **< 37.8 &** Z_3 **≥ 3163 &** Z_1 **≥ -0.46**

4: Z_3 **< 3960 &** Z_2 **≥ 37.8**

5: Z_3 **≥ 3960 &** Z_2 **< 37.8**

6: Z_3 **≥ 3960 &** Z_2 **≥ 37.8**

безусловн..		**1**	**2**	**3**	**4**	**5**	**6**
0.08	**1**	0.29	0	0	0.71	0	0
0.12	**2**	0.1	0.4	0	0	0.4	0.1
0.06	**3**	0.2	0.2	0	0.6	0	0
0.3	**4**	0.12	0	0.08	0.54	0.04	0.19
0.15	**5**	0	0.23	0.08	0.08	0.38	0.23
0.29	**6**	0	0.08	0.08	0.12	0.08	0.64

В таблице приведены оценки безусловных вероятностей нахождения в соответствующих состояниях (первый столбец), а также переходных вероятностей.

Значение критерия для полученного решения K = 0,8.

Как можно заметить, переходные вероятности существенно отличаются друг от друга и от безусловных вероятностей. Это свидетельствует о том, что предложенный алгоритм может находить закономерности в многомерных временных рядах.

Заключение

Методы одновременного прогнозирования нескольких целевых переменных при решении задачи прогнозирования многомерного разнотипного временного ряда, по сравнению с методами построения решающей функции по каждой целевой переменной в отдельности, позволяют точнее учитывать взаимозависимости переменных. В изложенном в работе методе в качестве решающей функции выступает матрица переходов для состояний случайного процесса, соответствующих адаптивно подобранному разбиению пространства переменных. При этом разбиение строится непосредственно в пространстве переменных, описывающих временной ряд, что существенно снижает трудоемкость алгоритма. Критерием качества решения в таком методе является информативность матрицы переходов.

Применение предложенного метода к анализу метеорологических данных продемонстрировало эффективность метода и возможность получения закономерностей, представляющих интерес для дальнейшего содержательного анализа.

Литература

[Лбов, Старцева, 1999] Г.С. Лбов, Н.Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999, 211 с.

[Ростовцев, 1978] П.С. Ростовцев. Алгоритм построения типологий для больших массивов социально-экономической информации. // Модели агрегирования социально-экономической информации. Сборник научных трудов, изд. ИЭ и ОПП СО АН СССР, 1978.

[Lbov, Nedel'ko, 2001] G.S. Lbov, V.M. Nedel'ko. A Maximum informativity criterion for the forecasting several variables of different types. // Computer data analysis and modeling. Robustness and computer intensive methods. Minsk, 2001, vol 2, 43–48.

[Неделько, 2004] С. В. Неделько. Критерий информативности матрицы переходов и прогнозирование многомерного разнотипного временного ряда. // Искусственный интеллект, № 2, 2004, с. 145–149.

[Данилов, Жиглявский, 1997] Д.Л. Данилов, А.А. Жиглявский. Главные компоненты временных рядов: метод «Гусеница». Санкт-Петербургский Государственный Университет, 1997.

Authors' Information

Svetlana Valeryevna Nedel'ko – Institute of Mathematics SB RAS, Laboratory of Data Analysis, 630090, pr. Koptyuga, 4, Novosibirsk, Russia, e-mail: nedelko@math.nsc.ru