ДОКЛАДЫ АКАДЕМИИ НАУК, 1994, том 339, № 6, с. 750 - 753

ИНФОРМАТИКА

УДК 519.24

МЕТОД АНАЛИЗА МНОГОМЕРНЫХ РАЗНОТИПНЫХ ВРЕМЕННЫХ РЯДОВ В КЛАССЕ ЛОГИЧЕСКИХ РЕШАЮЩИХ ФУНКЦИЙ

© 1994 г. Г. С. Лбов

Представлено академиком М.М. Лаврентьевым 19.11.93 г.

Поступило 26.11.93 г.

Пусть некоторый объект (процесс) описывается набором случайных характеристик X, изменяющихся во времени. Необходимо на основе анализа информации, представляющей измерения этих характеристик в ряде последовательных моментов времени (предыстория), предсказать значения набора характеристик Y ($Y \subseteq X$) в некоторый будущий момент времени. В качестве примера можно рассматривать задачу прогнозирования погоды по предыдущим наблюдениям.

Отличительной особенностью рассматриваемых задач прогнозирования, связанных с анализом многомерных временных рядов, является, во-первых, разнотипность измеряемых характеристик: набор переменных одновременно может содержать бинарные, номинальные и количественные переменные. В этом случае многомерный ряд представляет собой набор бинарных, символьных и числовых случайных последовательностей. Это обстоятельство не позволяет применять известные методы анализа многомерных временны́х рядов, использующие в том или ином виде понятие близости (метрики) в пространстве переменных. Во-вторых, при создании экспертных систем в области прогнозирования ставится задача обнаружения закономерностей ("знаний"), отражающих внутренние причинно-следственные связи между различными характеристиками изучаемого сложного динамического объекта. В-третьих, поскольку соответствующие распределения многомерного разнотипного случайного процесса неизвестны, решающие функции прогнозирования строятся на основе статистического анализа временны́х рядов (по обучающей выборке). В условиях малой выборки и высокой размерности пространства переменных возникает проблема получения статистически устойчивых решений для прогнозирования состояний динамических объектов.

Институт математики Сибирского отделения Российской Академии наук, Новосибирск

Как показывают теоретические и экспериментальные исследования, наиболее подходящим классом решающих функций в указанных выше условиях является класс логических решающих функций от разнотипных переменных [1, 2]. Это класс используется в данной работе для решения задачи прогнозирования.

1. Постановка задачи. Рассматривается n-мерный разнотипный случайный процесс $G = \{X_1(t), \ldots, X_j(t), \ldots, X_n(t)\}$. Пусть выбран набор прогнозируемых характеристик $X' = \{X_{j_1}, \ldots, X_{j_l}, \ldots, X_{j_m}\}$, $X' \subseteq X$; обозначим этот набор $Y = \{Y_1, \ldots, Y_l, \ldots, Y_m\}$, где $Y_1 = X_{j_l}$. Зафиксируем некоторые последовательные моменты времени $t_1, \ldots, t_d, \ldots, t_R, t_{R+1}$. обозначим: x_j^d – значение случайной переменной X_j в момент времени t_d, $x_j^d \in D_j$, а y_l^d – значение прогнозируемой переменной Y_l, $y_l^d \in D_{y_l}$.

Задача состоит в том, чтобы на основе таблицы данных $b = \{x_j^d\}, j = 1, \ldots, n; d = 1, \ldots, R$, предсказать набор значений $y = (y_1, \ldots, y_l, \ldots, y_m)$ в некоторый будущий момент t_{R+1}.

Таблицу данных b назовем предысторией длины R. Множество возможных предысторий обозначим B, а множество возможных наборов $\{y\}$ обозначим D_y, $b \in B$, $y \in D_y$, $D_y = \prod_{l=1}^{m} D_{y_l}$. Под решающей функцией прогнозирования понимается отображение множества B на множество D_y, т.е. $F: B \to D_y$. При построении решающей функции F используется следующая гипотеза: предполагается, что условное распределение $p(y|b)$ не зависит от сдвига во времени, т.е. указанное распределение для моментов времени $t_1, \ldots, t_R, t_{R+1}$ совпадает с распределением для моментов времени $t_1 + \Delta T, \ldots, t_R + \Delta T, t_{R+1} + \Delta T$. Если известно совместное распределение $p(y, b)$, то можно найти оптимальную решающую функцию прогнозирования F_0. Так как указанное распреде-

ление неизвестно, то решающую функцию будем строить на основе анализа многомерного временнóго ряда.

Пусть для случайного процесса G характеристики $X_1, \ldots, X_j, \ldots, X_n$ измерены в последовательные моменты времени. Обозначим это множество моментов $T = \{t_1, \ldots, t_\nu, \ldots, t_N\}$. Для простоты изложения в дальнейшем будем предполагать, что измерения проведены через равные интервалы времени $\Delta t = t_\nu - t_{\nu-1}$. Таким образом, эмпирическая информация представлена n-мерным разнотипным временнýм рядом $q = \{x_j^\nu\}, j = 1, \ldots, n; \nu = 1, \ldots, N$. Набор значений $x^{\mu-d} = (x_1^{\mu-d}, \ldots, x_j^{\mu-d}, \ldots, x_n^{\mu-d})$ назовем предысторией с номером d, соотнесенной с моментом времени $t_\mu, \mu = R+1, \ldots, N$. Для указанного момента времени t_μ под предысторией длины R, $1 \le R \ll N$, будем понимать таблицу $b^\mu = \{x^{\mu-d}\}$, $d = 1, \ldots, R$. Заметим, что одномерная символьная последовательность при $R = 1$ является реализацией простого марковского процесса с матрицей переходных вероятностей $\{p(y|x)\}, x \in A, y \in A$; A – алфавит символов.

Решающую функцию $\bar{F}$, построенную на основе анализа набора предысторий длины R, будем называть выборочной решающей функцией прогнозирования. Функцию $\bar{F}$ будем строить из класса логических решающих функций от разнотипных переменных.

2. Класс логических решающих функций. Пусть для описания изучаемого объекта выбран набор разнотипных характеристик $X_1, \ldots, X_j, \ldots, X_n$. Обозначим: D_j – множество возможных значений переменной X_j. Для бинарной переменной $D_j = \{0, 1\}$, для номинальной переменной множество $D_j = \{h_1, \ldots, h_{l_j}\}$ представляет собой набор имен (символов), для количественной переменной множество D_j – множество дискретных числовых значений либо некоторый интервал на вещественной оси. Пространство переменных представим как декартово произведение $D = \prod_{j=1}^{n} D_j$. Набор значений переменных обозначим $x = (x_1, \ldots, x_n)$, $x \in D$. Ставится задача: по набору x предсказать значение некоторой характеристики объекта Y. Эту характеристику назовем целевой переменной с множеством значений D_y. Целевая переменная Y может быть бинарной, номинальной или количественной переменной. Под решающей функцией $\tilde{y} = f(x)$ понимается отображение f: $D \to D_y$. Логическая решающая функция представляется в виде пары $\langle \alpha, r(\alpha) \rangle$, где α – разбиение множества D на M подмножеств $\alpha = \{E^1, \ldots, E^i, \ldots, E^M\}$, $1 \le M < \infty$, $r(\alpha)$ – набор решений $\{\tilde{y}^1, \ldots, \tilde{y}^t, \ldots, \tilde{y}^M\}$, $\tilde{y}^t \in D_y$. Класс логических решающих функций Φ_M задается как декартово произведение множества разбиений Ψ_M на множество решений R_M. Разбиение $\alpha \in \Psi_M$, если каждое подмножество из α представляется в виде $E^t = \prod_{j=1}^{n} E_j, E_j \subseteq D_j, E_j \subseteq W_j$, где для бинарных и номинальных переменных W_j – множество всевозможных подмножеств множества D_j, для количественных дискретных переменных W_j – множество подмножеств, включающих соседние элементы; для вещественных переменных W_j – множество подынтервалов интервала $D_j = [\alpha_j, \beta_j]$.

Пусть для некоторых номеров переменных $j \in \{j_1, \ldots, j_{m_t}\}$ $E_j \subset D_j$, для остальных переменных $E_j = D_j$, $m_t \le n$. Таким образом, отличительной особенностью предлагаемого класса разбиений Ψ_M является то, что для описания подмножества E^t используется свой набор из m_t переменных $X^t = \{X_{j_1}, \ldots, X_{jm_t}\}$, $X^t \subseteq X$. Введем предикат вида $J(a, E_j) =$ "$X_j(a) \in E_j$", $a \in \Gamma, E_j \in W_j$, где Γ – множество испытаний (наблюдений) случайной величины X_j. При фиксированных a и E_j предикат имеет два значения: истина, ложь. Очевидно, что конъюнкция $s(a, E^t) = J(a, E_1) \wedge \ldots \wedge J(a, E_n)$ эквивалентна конъюнкции $J(a, E_{j_1}) \wedge \ldots \wedge J(a, E_{jm_t})$.

Если известно совместное распределение $p(y, x)$, можно определить оптимальную решающую функцию f_0 из класса Φ_M. Поскольку в прикладных задачах распределение $p(y, x)$ неизвестно, решающая функция определяется на основе анализа эмпирической таблицы (обучающей выборки) $v = \{x_j^i, y^i\}, i = 1, \ldots, N; j = 1, \ldots, n$, где N – число наблюдений. Такую функцию будем называть выборочной решающей функцией и обозначать $\bar{f}$. Задача построения решающей функции $\bar{f}$ сводится либо к задаче распознавания образов, если переменная Y – бинарная или номинальная переменная, либо к задаче регрессионного анализа, если Y – количественная переменная. В первом случае минимизируется оценка вероятности ошибки распознавания, во втором случае – средний квадрат отклонений.

В работах [2, 3] предложены алгоритмы распознавания образов и регрессионного анализа для построения функции $\tilde{y} = f(x)$ в классе Φ_M. Функция $\bar{f}$ в том и другом случае представляется списком конъюнкций $\{s(a, E^t)\}, t = 1, \ldots, M$, обладающих определенными информативными свойствами. Такие конъюнкции будем называть

логическими закономерностями ("знаниями").

3. Метод решения задачи прогнозирования. Задачу прогнозирования будем решать для каждой переменной Y_l, $l = 1, \ldots, m$, отдельно, т.е. решающую функцию $\bar{F}$ будем задавать набором функций $\{\bar{f}_1, \ldots, \bar{f}_l, \ldots, \bar{f}_m\}$, где под функцией $\bar{f}_l$ будем понимать отображение $B \to \to D_{y_l}$. Построение функции $\bar{f}_l$ осуществляется в два этапа.

На первом этапе по временно́му ряду q организуется обучающая выборка в виде таблицы данных $v_d = \{x_j^{\mu-d}, y_l^{\mu}\}, j = 1, \ldots, n; \mu = R + 1, \ldots, N$. По таблице v_d строится логическая решающая функция $\bar{f}^d \sim \langle \alpha_d, r(\alpha_d) \rangle$ с разбиением $\alpha_d = \{E_d^1, \ldots \ldots, E_d^t, \ldots, E_d^{M_d}\}$ множества D на M_d подобластей. Введем новую переменную Z_d с множеством значений $D_{z_d} = \{z_d^1, \ldots, z_d^t, \ldots, z_d^{M_d}\} = \{1, \ldots, t, \ldots, M_d\}$ следующим образом: если предыстория $(x_1^{\mu-d}, \ldots \ldots, x_n^{\mu-d}) \in E_d^t$, то $z_d^t = t$. Указанные выше вычисления проводятся для всех номеров предысторий $d = 1, \ldots, R$. Результатом первого этапа является набор разбиений $\{\alpha_1, \ldots, \alpha_d, \ldots, \alpha_R\}$ и соответствующий ему набор переменных $\{Z_1, \ldots, Z_d, \ldots \ldots, Z_R\}$. Это позволяет для любого момента времени t_μ, $\mu = R + 1, \ldots, N$, n-мерную предысторию длины R $(x^{\mu-R}, \ldots, x^{\mu-d}, \ldots, x^{\mu-1})$ представить в виде одномерной последовательности $(z_R^{\mu}, \ldots, z_d^{\mu}, \ldots \ldots, z_1^{\mu})$, $x^{\mu-d} \in D$, $z_d^{\mu} \in D_{z_d}$.

На втором этапе n-мерному временно́му ряду $q = \{x_j^{\nu}\}$, $\nu = 1, \ldots, N$; $j = 1, \ldots, n$, ставится в соответствие таблица данных $v = \{z_d^{\mu}, y_l^{\mu}\}$, $d = 1, \ldots, R$; $\mu = R + 1, \ldots, N$. По таблице v строится логическая решающая функция $\bar{f}_l$, которая позволяет прогнозировать значение переменной Y_l, $l = 1, \ldots, m$, по предыстории длины R.

4. Прикладные задачи. Метод прогнозирования был использован для решения трех прикладных задач.

1) В книге [4] описана марковская модель прогноза состояния планктона озера Байкал. В этой модели использовалась однородная марковская последовательность случайных переменных x_ν со значениями в множестве $\{0, 1\}$. Событие $x_\nu = 0$ означает, что ν-й год нормальный (немелозирный), а событие $x_\nu = 1$ – что ν-й год аномальный (мелозирный). Ниже приводится бинарная последовательность, соответствующая периоду с 1946 по 1977 г.: 1000 1001000 100 11001000 1001001010. На основе анализа данной последовательности необходимо было предложить решающую функцию и проверить ее эффективность при прогнозировании состояния планктона с 1978 по 1983 г., которым соответствует последовательность 11 0010. Марковская модель прогноза дает такое решение 0 11 000, т.е. три ошибки.

Предложенный в данной работе метод прогнозирования выделил три логических закономерности (длина предыстории $R = 4$):

"если $(x^{\mu-4} = 1) \wedge (x^{\mu-1} = 1)$, то $\tilde{y} = 0$";

"если $(x^{\mu-4} = 1) \wedge (x^{\mu-1} = 0)$, то $\tilde{y} = 1$";

"если $(x^{\mu-4} = 0)$, то $\tilde{y} = 0$",

где μ – номер прогнозируемого символа в общей последовательности. Использование данной решающей функции для прогнозирования на год вперед (μ = 33, 34, 35, 36, 37, 38) дало такой результат: 11 0010, т.е. одну ошибку.

2) Ниже приводится одномерная символьная последовательность, которая представляет собой условную классификацию 11-летних циклов солнечной активности по цюрихской нумерации (с 1501 по 1964 г.). Каждый 11-летний цикл характеризуется одним из 7 символов: *AI*, *BJ*, *BK*, *CL*, *CM*, *DM*, *DN*. Для указанного периода символьная последовательность имеет вид *BK*, *DM*, *BJ*, *DM*, *AI*, *CM*, *BK*, *DM*, *BK*, *DN*, *BJ*, *CM*, *BK*, *DM*, *AI*, *CM*, *BJ*, *DM*, *BK*, *DN*, *AI*, *CL*, *AI*, *CL*, *BJ*, *DM*, *AI*, *CM*, *BK*, *CL*, *AI*, *CM*, *BK*, *DN*, *BJ*, *DN*, *BK*, *DM*, *BJ*, *DN*, *AI*, *CL*. Для четных циклов символы *A* и *B* обозначают типы зависимости высоты цикла от продолжительности ветви роста; *A*, *I*, *J*, *K* – типы зависимости высоты цикла от отношения продолжительностей ветвей спада и роста. Для нечетных циклов соответствующие символы есть *C*, *D* и *L*, *M*, *N*. Задача состоит в том, чтобы обнаружить закономерности в чередовании типов циклов для прогнозирования типов последующих циклов. В частности, предсказать 43-й символ указанной последовательности. Исходная информация и постановка задачи предложена сотрудником Института солнечной и земной физики СО РАН. Построенная решающая функция $\bar{F}$ представляется списком 18 логических закономерностей. Для прогнозирования 43-го символа ($\mu = 43$) оказалась истинной следующая закономерность: "если $(x^{\mu-1} = CL \vee DN) \wedge (x^{\mu-6} = \overline{BJ}) \wedge (x^{\mu-8} = AI \vee BJ)$, то $\tilde{y} = x^{\mu} = BJ$". Предсказанное значение символа совпадало с истинным значением.

3) Решалась задача прогноза среднемесячной температуры воздуха в Ояшинском районе Новосибирской области. Для построения решающей функции был использован одномерный числовой ряд, представляющий измерения среднемесячной температуры в течение 17 лет. Предполагается, что для успешного прогнозирования достаточно учитывать предысторию из 12 месяцев. Функция

$\bar{F}$, представленная списком из 21 закономерности, была использована для прогнозирования среднемесячной температуры на каждый месяц вперед в течение последующих двух лет. В качестве примера приведем закономерность, которая оказалась истинной при прогнозировании температуры января второго года: "если $(x^{\mu-12} \leq -19.2) \wedge$ $\wedge (x^{\mu-6} > -12.4) \wedge (x^{\mu-5} \leq 15.8) \wedge (x^{\mu-2} > -14.3)$, то $\tilde{y} = -23.8°C$". Истинное значение температуры $y =$ $= -23.1°C$. Оценка стандартного отклонения, вычисленная по 24 месяцам прогноза, равна 0.5°C.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 93–012–466).

СПИСОК ЛИТЕРАТУРЫ

1. *Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981. 160 с.

2. *Лбов Г.С., Старцева Н.Г.* В кн.: Методы анализа данных (Выч. системы, В. 111). Новосибирск, 1986. С. 3 - 10.

3. *Старцева Н.Г.* В кн.: Анализ данных и знаний в экспертных системах (Выч. системы. В. 134). Новосибирск, 1990. С. 67 - 84.

4. Прогнозирование экологических процессов / Под ред. Кожовой О.М., Паутовой В.Н., Черепанова В.В. Новосибирск: Наука, 1986. С. 210.