МЕТОДЫ СТАТИСТИЧЕСКОЙ КЛАССИФИКАЦИИ СТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ

Микулич Г. В., Жук Е. Е.
Кафедра математического моделирования и анализа данных,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: aragornguga@gmail.com, zhukee@mail.ru

Рассматриваются различные подходы к проблеме статистической классификации стационарных в широком смысле временных рядов: классификация коэффициентов авторегрессии и ковариационных функций.

Введение

Задача статистической классификации является одной из основных прикладных задач математической статистики. При этом, данные, подлежащие классификации, во многих случаях находятся в формате временных рядов, например: биомедицинские измерения (электрокардиограмма, кровяное давление), данные о погоде, цены акций на бирже и др. В данной работе изучается частный случай стационарных временных рядов, поскольку исследование нестационарных временных рядов может быть сведено к исследованию стационарных [1]. Рассматривается два подхода к решению задачи, а также построены решающие правила для алгоритмов, например, дискриминантного или кластерного анализа.

I. Математическая модель

Различие подходов заключается в разном выборе признаков для классификации: в первом случае это коэффициенты авторегрессии модели АР(p), а во втором – ковариационные функции реализаций стационарных временных рядов. Обозначим подход через классификацию коэффициентов авторегрессии как первый, а через ковариационные функции – как второй. Опишем обе математические модели, принимая во внимание то, что разделение на классы неодинаково для разных подходов.

Пусть наблюдается случайная выборка $X^n = (X_1, \ldots, X_n)$ объёма n из независимых в совокупности случайных векторов наблюдений, принадлежащих к $L \geq 2$ классам $\Omega_1, \ldots, \Omega_L$. Наблюдение X_t принадлежит к классу со случайным ненаблюдаемым номером $d_t^0 \in S$, $S = \{1, \ldots, L\}$, $t = \overline{1,n}$ и при фиксированном номере класса $d_t^0 = i$, $i \in S$ является:

1. Реализацией длительности T_t ($X_t = (x_{t1} \ldots, x_{tT_t})' \in R^{T_t}$, $'$ – символ транспонирования) временного ряда авторегрессии $x^i = \left\{x_l^i\right\}_{l=-\infty}^{+\infty}$ порядка $p \geq 1$ (модель АР(p))

$$x_l^i + \theta_{i1}^0 x_{l-1}^i + \ldots + \theta_{ip}^0 x_{l-p}^i = u_l^i,\ l \in Z, \quad (1)$$

где $Z = \{0, \pm 1, \pm 2, \ldots\}$, $\theta_i^0 \in R^p$ –вектор авторегрессии для i-го класса, а $\left\{u_l^i\right\}_{l=-\infty}^{+\infty}$– независимые в совокупности нормальные случайные величины с нулевым математическим ожиданием и одинаковой дисперсией σ^2 для всех классов Ω_i:

$$E\left\{u_l^i\right\} = 0,\ D\left\{u_l^i\right\} = \sigma^2,\ l \in Z,\ i \in S. \quad (2)$$

2. Реализацией длительности T_t стационарного временного ряда с нулевым математическим ожиданием и ковариационной функцией $\sigma_i(h)$:

$$E\left\{x_{tj} | d_t^0 = i\right\} = 0, j = 1, 2, \ldots$$

$$\sigma_i(h) = \sigma_i(-h) = E\left\{x_{tj}, x_{t,j+h} | d_t^0 = i\right\},$$

$$h = 0, 1, 2, \ldots, i \in S. \quad (3)$$

Наряду с вектором коэффициентов авторегрессии θ_i^0 для первого случая и ковариационными функциями $\{\sigma_i(\cdot)\}_{i \in S}$ для второго случая, классы Ω_i также характеризуется своей априорной вероятностью:

$$P\left\{d_t^0 = i\right\} = \pi_i^0 > 0,\ i \in S,\ \sum_{i=1}^{L} \pi_i^0 = 1. \quad (4)$$

II. Классификация в пространстве оценок коэффициентов авторегрессии

Используем модель (1),(2),(4). Преобразуем исходную выборку X^n в выборку $Y^n = (Y_1, \ldots, Y_n)$, где $Y_t \in R^p$, $t = \overline{1,n}$ – МП-оценка для p-вектора коэффициентов авторегрессии $\theta_{d_t^0}^0 \in R^p$, построенная по наблюдению $X_t \in R^{T_t}$, являющемуся реализацией длительности T_t одного из временных рядов АР(p) из (1).

Для построения МП-оценки Y_t воспользуемся тем фактом, что наблюдение $X_t \in R^{T_t}$ при фиксированном $d_t^0 = i$, $i \in S$ является нормальным T_t-вектором с нулевым математическим ожиданием $E\left\{X_t | d_t^0 = i\right\} = 0_{T_t}$, и плотностью распределения вероятностей

$$p(X_t; \theta_i^0, \sigma) = n_p(X_t^p | 0_p, R_p(\theta_i^0, \sigma)) \times$$

$$\times (2\pi)^{-\frac{T_t - p}{2}} \sigma^{-(T_t - p)} \times \quad (5)$$

$$\times \exp\left(\frac{-1}{2\sigma^2}\sum_{l=p+1}^{T_t}(x_{tl}+\theta_{i1}^0 x_{t,l-1}+\ldots+\theta_{ip}^0 x_{t,l-p})^2\right).$$

В формуле (5) $n_p(y|\mu,\Sigma)$ – плотность p-мерного нормального распределения, $X_t^p = (x_{t1},\ldots,x_{tp})^{`} \in R^p$, $R_p(\theta_i^0,\sigma) = E\left\{X_t^p(X_t^p)^{`} \mid d_t^0 = i\right\}$ – невырожденная [2] ковариационная матрица, элементами которой являются автоковариации $(\rho_{|k-l|}(\theta_i^0,\sigma))_{k,l=1}^p$, определяемые системой уравнений Юла – Уокера [3]:

$$\sum_{j=1}^{p}\theta_{ij}^0\rho_j(\theta_i^0,\sigma)=\sigma^2;$$

$$\sum_{j=1}^{p}\theta_{ij}^0\rho_{|k-j|}(\theta_i^0,\sigma)+\rho_k(\theta_i^0,\sigma)=0,\ k=1,2,\ldots$$

Согласно методу максимального правдоподобия,

$$\{Y_t,\hat{\sigma_t}\}=\arg\max_{\{\bar{\theta},\sigma\}}\ln p(X_t;\bar{\theta},\sigma),$$

где $p(X_t;\bar{\theta},\sigma)$ – плотность из (5), записанная для $\bar{\theta}$, $\theta_i^0 := \bar{\theta}$. Получаем следующее решающее правило:

$$\hat{d}_t=\arg\min_{i\in S}|Y_t-\hat{\theta}_i|,\ t=\overline{1,n}.$$

Для решения задачи дискриминантного анализа по вектору истинной классификации строим оценки для "центров"классов (в случае кластерного анализа, например, с помощью алгоритма L-средних, вектор истинной классификации заменяется на оценки, полученные на соответствующем шаге алгоритма):

$$\hat{\theta}_i=\left(\sum_{t=1}^{n}\delta_{d_t,i}\right)^{-1}\sum_{t=1}^{n}\delta_{d_t,i}Y_t,\ i\in S$$

III. Классификация в пространстве ковариационных функций

Используем модель (3),(4). Введем обозначения:

$$\overline{\sigma_i}=(\sigma_i(0),\sigma_i(1),\ldots,\sigma_i(N))^T\in R^{N+1}- \qquad (6)$$

вектор, образованный из первых $N+1$ ковариаций (значений ковариационной функции, определяющих класс Ω_i. Исходному наблюдению-реализации $x_t = (x_t1,\ldots,x_{t,T_t})^T \in R^{T_t}$ длительности T_t поставим в соответствие наблюдение в пространстве оценок ковариационных функций:

$$y_t=(y_{t0},y_{t1},\ldots,y_{tN})^T\in R^{N+1},$$

$$t=1,\ldots,n,n+1,\ldots,$$

где

$$y_{th}=\frac{1}{T_t-h}\sum_{j=i}^{T_t-h}(x_{tj}x_{t,j+h})$$

– непараметрическая оценка ковариаций с лагом h, построенная по x_t.

Определим решающее правило:

$$d(y;\{\hat{\sigma}_i\}_{i\in S})=\arg\min_{i\in S}|y-\hat{\sigma}_i|,y\in R^{N+1}$$

РП относит наблюдение к такому классу, к “центру” которого оно ближе. В роли “центров” классов здесь выступают $(N+1)$ – вектора из (6), а в качестве меры близости используется метрика Евклида. Строим оценки для "центров"классов:

$$\hat{\sigma}_i=\left(\sum_{t=1}^{n}\delta_{d_t^0,i}y_t\right),i\in S$$

IV. Пример: процедура кластер-анализа

Построим процедуру кластер-анализа в пространстве МП-оценок параметров авторегрессии, основанную на алгоритме L -средних.

1. По исходной выборке X^n из (6) находится выборка МП-оценок Y^n, из которой в качестве начальных приближений $\left\{\hat{\theta}_i^{(0)}\right\}$, $i \in S$ для «центров» $\left\{\theta_i^0\right\}$, $i \in S$ классов $\{\Omega_i\}$ выбираются какие-либо L наблюдений $Y_{j_1},\ldots,Y_{j_L}$, $j_i \in \{1,\ldots,n\}$; $j_i \neq j_k$, $i \neq k \in S$;
2. На l-м шаге ($l = 0,1,2,\ldots$) производится классификация выборки Y^n:

$$(\hat{d}_t^{(l)})=\arg\min_{i\in S}|Y_t-\hat{\theta}_i^{(l-1)}|,\ t=\overline{1,n},$$

т.е. строится оценка $\hat{D}^{(l)} = (\hat{d}_1^{(l)},\ldots,\hat{d}_n^{(l)})^{`}$ для D^0, и уточняются оценки для «центров» классов:

$$\hat{\theta}_i^{(l)}=\left(\sum_{t=1}^{n}\delta_{\hat{d}_t^{(l)},i}\right)^{-1}\sum_{t=1}^{n}\delta_{\hat{d}_t^{(l)},i}Y_t,\ i\in S$$

где $\delta_{i,j}$ – символ Кронекера;
3. Итерационный процесс останавливается при достижении на l-м шаге ($2 \leq l < \infty$) равенства $\hat{D}^{(l)} = \hat{D}^{(l-1)}$, и его результатом являются оценки $\hat{D} := \hat{D}^{(l)}$ для вектора истинной классификации D^0 и $\hat{\theta} := \hat{\theta}^{(l)} \in R^{Lp}$ для составного вектора θ^0 параметров авторегрессии θ_i^0, $i \in S$.

V. Заключение

Рассмотрена проблема классификации стационарных временных рядов. Предложены различные подходы к её решению, приведён алгоритм кластерного анализа в пространстве оценок максимального правдоподобия параметров авторегрессии, основанный на алгоритме L-средних.

1. Харин, Ю. С. Теория вероятностей, математическая и прикладная статистика : учеб. пособие / Ю. С. Харин, Н. М. Зуев, Е. Е. Жук. – Минск: БГУ, 2011. – 464 с.
2. Бокс, Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М. : Мир, 1974. – 406 с.
3. Андерсон, Т. Статистический анализ временных рядов / Т. Андерсон. – М. : Мир, 1976. – 760 с.