СТАТИСТИЧЕСКАЯ КЛАССИФИКАЦИЯ СТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ РАЗЛОЖЕНИЯ ВОЛЬДА

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

Рассматривается проблема статистической классификации реализаций стационарных в широком смысле временных рядов, различающихся по классам значениями своих параметров – коэффициентов авторегрессии из разложения Вольда.

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 случай стационарных временных рядов, а в качестве процедуры классификации выбран кластерный анализ.

I. Математическая модель

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 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 AP(p))

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

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:

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

Наряду с вектором коэффициентов авторегрессии θ_i^0 класс Ω_i также характеризуется своей априорной вероятностью:

$$P\left\{d_t^0 = i\right\} = \pi_i^0 > 0,\ i \in S,\ \sum_{i=1}^{L} \pi_i^0 = 1. \qquad (3)$$

II. Предварительные преобразования

Пусть имеется стационарный в широком смысле временной ряд. В модели (1) – (3) порядок авторегрессии полагается известным, но в нашем случае приходится дополнительно находить для него оценку $\hat{p}$. Эта задача сводится с помощью разложения Вольда

$$x_l + \sum_{j=1}^{\infty} \theta_j x_{l-j} = u_l,\ l \in Z$$

к оцениванию неизвестных коэффициентов авторегрессии θ_j (u_l удовлетворяет условиям (2)). Пошаговая процедура оценивания описана в [1].

Вернёмся к модели (1) – (3). Теперь будем полагать порядок авторегрессии p известным и равным полученной оценке $\hat{p}$. Пусть для коэффициентов авторегрессии выполняется следующее соотношение:

$$z^p + \sum_{j=1}^{p} \theta_{ij} z^{p-j} = 0 \qquad (4)$$

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 AP(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 (5), указанной в конце документа.

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 уравне-

ний 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,

$$\{Y_t, \hat{\sigma_t}\} = \arg \max_{\{\bar{\theta}, \sigma\}} \ln p(X_t; \bar{\theta}, \sigma), \qquad (6)$$

где $p(X_t; \bar{\theta}, \sigma)$ – плотность из (5), записанная для $\bar{\theta},\ \theta_i^0 := \bar{\theta}$. Отметим, что при решении задачи (6) наряду с Y_t строится также оценка $\hat{\sigma^2}$ для дисперсии.

III. Исследование полученных оценок

Теорема. Пусть в условиях модели (1) – (3) корни характеристических уравнений (4) лежат внутри единичного круга. Тогда при фиксированном векторе истинной классификации $D^0 = (d_1^0, \ldots, d_n^0)$ МП-оценки $Y_1, \ldots, Y_n$, полученные из (6) являются состоятельными оценками для соответствующих параметров авторегрессии:

$$Y_t \xrightarrow{P} \theta_{d_t^0}^0,\ T_t \to +\infty,$$

и имеют асимптотически нормальное распределение

$$\sqrt{T_t}(Y_t - \theta_{d_t^0}^0) \rightsquigarrow N_p(0_p, \sigma^2 R_p^{-1}(\theta_{d_t^0}^0, \sigma)),\ T_t \to +\infty.$$

Доказательство. Справедливость полученных соотношений следует из статистических свойств МП-оценок коэффициентов авторегрессии [1] и того факта, что при фиксированном $d_t^0 = i,\ i \in S$ исходное наблюдение X_t является реализацией длительности T_t временного ряда АР(p) с вектором коэффициентов авторегрессии $\theta_{d_t^0}^0$, а Y_t – МП-оценкой для $\theta_{d_t^0}^0$ из (6), полученной по $X_t,\ t = \overline{1, n}$.

IV. Кластер-процедура

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 $\theta_i^0,\ i \in S$.

Замечание. Как показано в [2], при больших длительностях наблюдений ($T_t \to +\infty$) из исходной выборки в приведённом выше алгоритме вместо МП-оценок можно использовать МНК-оценки:

$$Y_t = -\left(\sum_{l=p+1}^{T_t} X_{tl}^p (X_{tl}^p)'\right)^{-1} \sum_{l=p+1}^{T_t} x_{tl} (X_{tl}^p)',$$

$$X_{tl}^p = (x_{t,l-1}, \ldots, x_{t,l-p})' in R^p,\ t = \overline{1, n},$$

или оценки Юла – Уокера:

$$Y_t = -(\hat{R}_t^p)^{-1} \hat{r}_t^p, \hat{R}_t^p = (\hat{\rho}_{|k-l|}^t)_{k,l=1}^p,$$

$$\hat{r}_t^p = (\hat{\rho}_1^t, \ldots, \hat{\rho}_p^t)',\ t = \overline{1, n}.$$

V. Заключение

Рассмотрена проблема классификации стационарных временных рядов. Предложен 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.

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

$$\times (2\pi)^{-\frac{T_t - p}{2}} \sigma^{-(T_t - p)} \exp\left(-\frac{1}{2\sigma^2} \sum_{l=p+1}^{T_t} (x_{tl} + \theta_{i1}^0 x_{t,l-1} + \ldots + \theta_{ip}^0 x_{t,l-p})^2\right) \qquad (5)$$