УДК 681.513.8

КЛАССИФИКАЦИЯ ОБЪЕКТОВ, ЗАДАННЫХ ВРЕМЕННЫМИ РЯДАМИ

В.А. Павлов

Открытый международный университет развития человека «Украина»
vpavlo@bk.ru

Розглянуто задачу класифікації об'єктів, заданих часовими рядами. Для оцінки взаємодії часових рядів запропоновано ряд критеріїв селекції статистичних причинно-наслідкових структур. Критерії використовують значення крос-кореляційної функції. Визначено механізм вибору критерію і значення порогів селекції, що найкращим чином дозволяють вирішити задачу класифікації об'єктів.

Ключові слова: *класифікація, кроскореляційна функція , причинно-наслідкові структури, критерій селекції.*

The paper considers the classification problem of objects that are represented **by** time series. To assess the interaction of time-series a number of selection criteria of statistical causal structures are offered. The criteria use cross-correlation function values computed using advanced and lagged shift index. The technique for determining selection criterion and selection thresholds that solve the problem best, was found.

Keywords: classification, cross-correlation function, causal structure , selection criteria

Рассмотрена задача классификации объектов, заданных временными рядами. Для оценки взаимодействия временных рядов предложен ряд критериев селекции статистических причинно-следственных структур. Критерии используют значения кросс-кореляционой функции, рассчитанной при опережающем и запаздывающем индексе сдвига. Определен механизм выбора критерия и значения порогов селекции наилучшим образом разрешающие задачу классификации объектов.

Ключевые слова: классификация, кросс-кореляционная функция, причинно-следственные структуры, критерий селекции.

Вступление

Задачи классификации объектов имеют широкое распространение в различных областях человеческой деятельности. Аппарат решения задач классификации интенсивно развивается. Настоящая работа посвящена разработке нового подхода к классификации объектов, заданных временными рядами.

1.Постановка задачи

Рассматривается задача классификации, в которой объект характеризуется не одиночными измерениями (точками) в многомерном пространстве признаков, а их множествами, представляющими собой реализации временных рядов.

Для простоты рассматривается случай двух классов. Объекты A_i и B_j классов A и B характеризуются реализациями $x_{pj}, j = 1,\ldots,n$ временных рядов (процессов) X_p, $p=1,..,m$.

Ряды признаки X_p, затруднительно непосредственно использовать для оценки меры близости объектов к определенному классу. Поэтому далее предложим процедуру формирования вторичных признаков, позволяющих применить привычные меры близости и алгоритмы классификации. Данная процедура должна использовать целесообразные критерии и параметры селекции для получения наилучшего качества результатов классификации.

2. Содержание работы

Эффективность предложенного далее подхода зависит от степени выполнения следующих предположений:

1. Временные ряды X_p, характеризующие объекты классификации представляют собой взаимосвязанные процессы.
2. Взаимосвязь процессов X_p в значительной мере должна характеризоваться линейным эффектом.

Обозначим реализации рядов X_p, характеризующих объекты A_i , $i=1,...,N_A$ класса А и объекты B_i, $i=1,...,N_B$ класса В, как $x_{pj}^{A_i}$, $j = 1,\ldots,n_{pA_i}$ и $x_{pj}^{B_i}$, $j = 1,\ldots,n_{pB_i}$.

Так как предлагаемый ниже механизм не налагает ограничений на различия в длинах реализаций процессов, то без потери общности упростим индексацию, считая длины реализаций для всех объектов одинаковыми $n_{pA_i} = n_{qB_j} = n$.

Логично считать, что при выполнении предположений метода одним из характерных признаков классифицируемых объектов могут быть статистические причинно-следственные структуры (ПСС) процессов X_p. А отличия в ПСС объектов различных классов в определенных случаях могут быть основой для их классификации. Необходимо определить критерии и параметры селекции ПСС, дающие наилучшие результаты классификации.

Рассмотрим реализации x_{pj} $j = 1,\ldots,n$ процессов X_p, для произвольного объекта выборки данных, пока не акцентируя его принадлежность к классу. Выберем реализации x_{pj} и x_{qj} . $j = 1,\ldots,n$ некоторых процессов X_p, X_q и

построим кросс-корреляционную функцию (ККФ) с опережающим индексом сдвига k для каждого из них.

На рис.1 и рис. 2 показано формирование рядов для расчета ККФ $R_k^{X_{p(t+k)}X_{qt}}$ и $R_k^{X_{q(t+k)}X_{pt}}$.

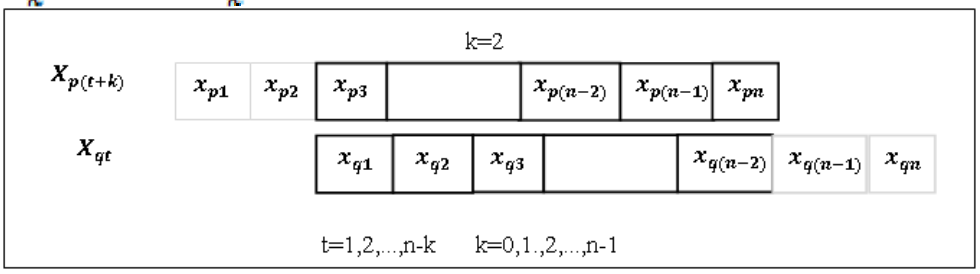

Рис.1 ККФ с опережающим индексом сдвига *k* для X_p

По выделенным элементам рядов рассчитываем значения $R_k^{X_{p(t+k)}X_{qt}}$.

$$R_k^{X_{p(t+k)}X_{qt}} = \frac{cov(X_{p(t+k)}X_{qt})}{s_p s_q}, \quad (1)$$

где $cov(X_{p(t+k)}X_{qt})$–выборочная ковариация, s_p , s_q- несмещенные выборочные оценки дисперсии выделенных рядов.

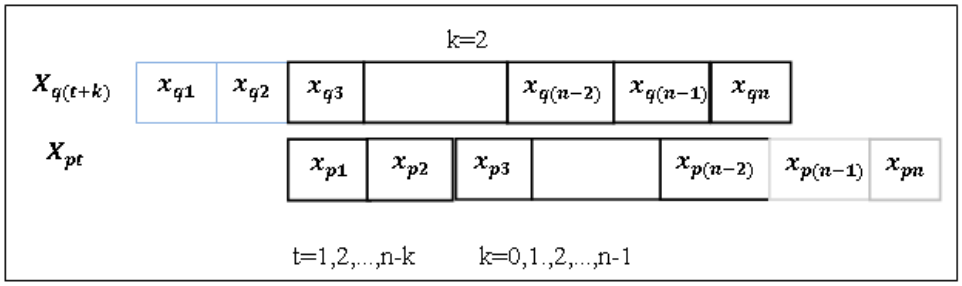

Рис.2 ККФ с опережающим индексом сдвига *k* для X_q

По выделенным элементам рядов рассчитываем значения $R_k^{X_{q(t+k)}X_{pt}}$.

$$R_k^{X_{q(t+k)}X_{pt}} = \frac{cov(X_{q(t+k)}X_{pt})}{s_p s_q}, \quad (2)$$

где $cov(X_{q(t+k)}X_{pt})$–выборочная ковариация и s_p , s_q- несмещенные выборочные оценки дисперсии выделенных рядов.

Расчет ККФ через выборочные значения ряда:

$$R_k^{X_{p(t+k),qt}} = \frac{\sum_{t=1}^{n-k} x_{p(t+k)}x_{qt}/(n-k) - \sum_{t=1}^{n-k} x_{p(t+k)}/(n-k) \cdot \sum_{t=1}^{n-k} x_{qt}/(n-k)}{\sqrt{\left[\frac{1}{(n-k-1)}\sum_{t=1}^{n-k}(x_{p(t+k)} - \frac{1}{(n-k)}\sum_{t=1}^{n-k} x_{p(t+k)})^2\right] \cdot \left[\frac{1}{(n-k-1)}\sum_{t=1}^{n-k}(x_{qt} - \frac{1}{(n-k)}\sum_{t=1}^{n-k} x_{qt})^2\right]}}, \quad (3)$$

$$R_k^{X_{q(t+k),pt}} = \frac{\sum_{t=1}^{n-k} x_{q(t+k)} x_{pt} /(n-k) - \sum_{t=1}^{n-k} x_{q(t+k)} /(n-k) \cdot \sum_{t=1}^{n-k} x_{pt} /(n-k)}{\sqrt{\left[\frac{1}{(n-k-1)} \sum_{t=1}^{n-k} (x_{q(t+k)} - \frac{1}{(n-k)} \sum_{t=1}^{n-k} x_{q(t+k)})^2\right] \cdot \left[\frac{1}{(n-k-1)} \sum_{t=1}^{n-k} (x_{pt} - \frac{1}{(n-k)} \sum_{t=1}^{n-k} x_{pt})^2\right]}} \quad (4)$$

Величины ККФ при каждом k характеризуют силу линейной связи между значениями рядов с данным сдвигом, а следовательно, степень прогностической способности соответствующей линейной модели. Таким образом, отличия в значениях $R_k^{X_{p(t+k)}X_{qt}}$ и $R_k^{X_{q(t+k)}X_{pt}}$ возможно использовать для установления направления статистической причинно-следственной связи значений ряда при данном k. Грубо говоря, при $\left|R_k^{X_{p(t+k)}X_{qt}}\right| > \left|R_k^{X_{q(t+k)}X_{pt}}\right|$ линейный прогноз $X_{p(t+k)}$ по X_{qt} будет более точный, чем $X_{q(t+k)}$ по X_{pt} и при данном k статистически $X_q \to X_p$

Однако суть причинно-следственных отношений неоднозначна. Очевидно, что при различных k соотношения между $R_k^{X_{p(t+k)}X_{qt}}$ и $R_k^{X_{q(t+k)}X_{pt}}$ могут меняться и соответственно меняется направление влияния процессов. Поэтому для оценки «взаимоотношений» процессов X_p и X_q и построения причинно-следственных структур (ПСС) целесообразно ввести ряд критериев, как для отдельных значений k, так и интегральных, для характеристики преимущественного направления влияния. В зависимости от цели задачи (прогноз, классификация, анализ объекта) возможно предложить ряд таких критериев. Ниже ограничимся наиболее простыми критериями селекции ПСС.

Упростим запись ККФ обозначив $R_k^{X_{p(t+k)}X_{qt}} = R_k^{X_{pt}X_{qt}}$ и $R_k^{X_{q(t+k)}X_{pt}} = R_k^{X_{qt}X_{pt}}$. Тогда далее рассмотрим для селекции ПСС следующие критерии:

$$Cr_1 = Cr_{k=1} = cr_{pq_1} - cr_{qp_1} = \left|R_1^{X_{pt}X_{qt}}\right| - \left|R_1^{X_{qt}X_{pt}}\right| \quad (5)$$

$$Cr_2 = Cr_{max} = cr_{pq_{max}} - cr_{qp_{max}} = \max_k \left|R_k^{X_{qt}X_{pt}}\right| - \max_k \left|R_k^{X_{pt}X_{qt}}\right| \quad (6)$$

$$Cr_3 = Cr_{n/2} = cr_{pq_{n/2}} - cr_{qp_{n/2}} = \sum_k^{n/2} \left|R_k^{X_{pt}X_{qt}}\right| - \sum_k^{n/2} \left|R_k^{X_{qt}X_{pt}}\right| \quad (7)$$

$$Cr_4 = Cr_n = cr_{pq_n} - cr_{qp_n} = \sum_k^{n-l_{pq}} \left|R_k^{X_{pt}X_{qt}}\right| - \sum_k^{n-l_{pq}} \left|R_k^{X_{qt}X_{pt}}\right| \quad (8)$$

где l_{pq} – один из искомых параметров, в простом случае $l_{pq} = l$ - параметр алгоритма.

С точки зрения чувствительности при идентификации однонаправленных и двунаправленных статистических причинно-следственных связей процедуру определения ПСС целесообразно параметризировать. Далее в качестве параметров процедуры введем два порога:

1. Δ_1 - порог нечувствительности определения направления связи ПСС .

2. Δ_2- порог значения cr, при котором правило процедуры действительно.

Для характеристики ПСС системы процессов $\mathbf{X_1, X_2, ..., X_m}$ введем матрицу ПСС [1]:

	X_1	X_2	...	X_m
X_1 ←	γ_{11}	γ_{12}	⋯	γ_{1m}
X_2 ←	γ_{21}	γ_{22}	⋯	γ_{2m}
...	⋯	⋯	⋯	⋯
X_m ←	γ_{m1}	γ_{m2}	⋯	γ_{mm}

Рис.3. Матрица ПСС процессов $\mathbf{X_1, X_2, ..., X_m}$

для которой элементы $\gamma_{pq} = 1$ если $X_p \leftarrow X_q$ и $\gamma_{pq} = 0$, если в данном направлении связи нет и $\gamma_{qp} = 1$, если $X_p \rightarrow \tilde{O}_q$ и $\gamma_{qp} = 0$, если в данном направлении связи нет.

Определим правило процедуры установления направления связи (и соответствующие значения элемента матрицы ПСС) следующим образом:

1. При $\max(cr_{pq}, cr_{qp}) \geq \Delta_2$:

если $Cr > \Delta_1$, то $X_q \rightarrow X_p$, если $Cr < -\Delta_1$, то $X_q \leftarrow X_p$,

если $|Cr| < \Delta_1$, то имеем двустороннюю связь $X_q \leftarrow X_p$, $X_q \rightarrow X_p$.

2. При $\max(cr_{pq}, cr_{qp}) < \Delta_2$ и $|Cr| < \Delta_1$ связи между X_p и X_q нет.

Случай $\max(cr_{pq}, cr_{qp}) < \Delta_2$, $|Cr| \geq \Delta_1$ требует дополнительного исследования.

В результате применения конкретного критерия при конкретных значениях порогов получим соответствующую данному выбору матрицу ПСС.

Пусть матрицы ПСС A_k объектов класса A и матрицы ПСС B_k объектов класса B имеют вид:

$$A_k = \begin{matrix} a_{11}^k & \dots & a_{1m}^k \\ \dots & \dots & \dots \\ a_{m1}^k & \dots & a_{mm}^k \end{matrix}, \quad B_k = \begin{matrix} b_{11}^k & \dots & b_{1m}^k \\ \dots & \dots & \dots \\ b_{m1}^k & \dots & b_{mm}^k \end{matrix}. \qquad (9)$$

Учитывая предположение о ПСС, как характеристике классов объектов, заданными временными рядами, необходимо выбрать критерий Cr_i и значения

порогов селекции ПСС, обеспечивающие наиболее близкие матрицы ПСС на объектах одного класса, в соответствии с минимумом внутриклассовой дисперсии:

$$Cr_{cl-} = \sum_{g}^{N_A}\sum_{h}^{N_A}(\sum_{j}^{m}\sum_{i}^{m}(a_{ij}^{g} - a_{ij}^{h})^2 + \sum_{g}^{N_B}\sum_{h}^{N_B}(\sum_{j}^{m}\sum_{i}^{m}(b_{ij}^{g} - b_{ij}^{h})^2, \quad (10)$$

и наиболее различающиеся ПСС на объектах различных классов в соответствии с максимумом межклассовой дисперсии:

$$Cr_{cl+} = \sum_{g}^{N_A}\sum_{h}^{N_b}(\sum_{j}^{m}\sum_{i}^{m}(a_{ij}^{g} - b_{ij}^{h})^2. \quad (11)$$

Тогда выбор конкретного критерия и порогов селекции ПСС задачи классификации возможно рассчитывать согласно следующих дисперсионных критериев:

$$Cr_{cl} - \max_{Cr_i,\Delta_1,\Delta_2}\left(\frac{\alpha}{Cr_{cl-}} + \beta \cdot Cr_{cl+}\right), \quad (12)$$

где коэффициенты α и β балансируют требования совпадения и различения матриц ПСС или аналог критерия разделимости классов [2]

$$Cr_W = \max_{Cr_i,\Delta_1,\Delta_2}\frac{Cr_{cl+}}{Cr_{cl-}}. \quad (13)$$

Процедура максимизации критериев может учитывать разбиение выборки объектов классификации: расчет параметров для всех вариантов критериев селекции на обучающей выборке и выбор наилучшего варианта по максимуму дисперсионного критерия (12) или (13) на проверочной выборке.

Меру близости объекта к классу, возможно принять как сумму отличий матрицы ПСС классифицируемого объекта от матриц ПСС объектов каждого класса. Естественные алгоритмы классификации: определение наилучшего граничного значения меры близости, разделяющего объекты классов (дискриминантный анализ), алгоритм ближайшего соседа или алгоритм взвешенного по объектного голосования в каждом классе.

3. Выводы

Предложен алгоритм классификации объектов, заданных временными рядами, основанный на построении меры близости матриц причинно-следственных связей объектов классификации**.**

Литература

1. Павлов В.А. Причинно-следственный анализ в системах процессов / Проблемы информационных технологий. – 2009. - №5. - С. 8-14.
2. Дуда Р., Харт П. Распознавание образов и анализ сцен / пер. с англ. - М.: «Мир», 1976.- 511с.