

Выявление групп объектов, описанных набором многомерных временных рядов

*Ивахненко А. А., Каневский Д. Ю., Рудева А. В.,
Стрижов В. В.*

Москва, Вычислительный центр РАН
strijov@ccas.ru

Далеко не во всех прикладных задачах стандартные способы оценивания корреляции между временными рядами имеют «разумную» содержательную интерпретацию. В данной работе предполагается, что исходная информация представлена матрицей «объект–показатель–время», и предлагается методика выделения групп объектов, демонстрирующих сходное поведение относительно заданного показателя, либо относительно заданного набора показателей.

Методика основана на построении матриц сходства между всеми парами временных рядов. Сходство (или расстояние) определяется числом размеченных интервалов совместного роста или спада. Используется понятие «разметки», определенное в [1], и алгоритм автоматической разметки временных рядов [2]. По матрице сходства однозначно определяются группы объектов со сходным поведением. Группы объектов задаются матрицей смежности и представляются в виде множества графов [3].

Описан алгоритм, который отыскивает все группы сходных объектов. Для этого алгоритм последовательно выполняет следующие операции: разметку основных рядов, построение матрицы сходства временных рядов, нахождение групп объектов для фиксированного показателя, нахождение групп объектов для произвольного набора показателей.

Разметка временных рядов

Дано декартово произведение множества объектов, множества показателей и множества моментов времени $\{x_{ijt}\}_{i,j,t=1}^{I,J,T}$. Элементом x_{ijt} этого произведения является значение j -ого показателя на i -ом объекте в момент времени t .

Временной ряд является размеченным, если каждому его элементу поставлен в соответствие знак из алфавита \mathcal{A} . Примером алфавита может служить множество $\mathcal{A} = \{U, D, N\}$, где символы U и D интерпретируются как увеличение и уменьшение значения показателя данного объекта в данный момент времени, символ N соответствует отказу от классификации. Таким образом каждому временному ряду $\{x_{ijt}\}_{t=1}^T$ поставлены в соответствие ряды $\{u_{ijt}\}_{t=1}^T$ и $\{d_{ijt}\}_{t=1}^T$, где $u_{ijt} \in \{U, N\}$, $d_{ijt} \in \{D, N\}$, интерпретируемые как ряды, размеченные интервалами роста и падения значений показателя на данном объекте.

Приведем пример условий, определяющих разметку временного ряда. Элементу из упорядоченной последовательности $x_{ijt_1}, \dots, x_{ijt_N}$ ставится в соответствие знак \bar{U} , если выполнено:

$$\left\{ \begin{array}{l} \bar{N} \leq N \in \mathbb{Z}; \\ x_{ijt_1} < x_{ijt_2}; \\ x_{ijt_{N-1}} < x_{ijt_N}; \\ \bar{D} < \left((x_{ijt_N} - x_{ijt_1}) - \sum_{\forall \tau \in T_D} (x_{ijt_\tau} - x_{ijt_{\tau+1}}) \right) (x_{ijt_N} - x_{ijt_1})^{-1}, \\ T_D = \{ \tau : x_{ijt_{\tau+1}} < x_{ijt_\tau} \}; \\ x_{ijt_1} \leq x_{ijt_\tau}, \quad \tau = 1, \dots, N; \\ \bar{L} \leq (x_{ijt_N} - x_{ijt_1}). \end{array} \right.$$

где \bar{N} , \bar{D} , \bar{L} — заданные параметры. Каждому элементу из упорядоченной последовательности $x_{ijt_1}, \dots, x_{ijt_N}$ ставится в соответствие знак \bar{D} , если выполнены вышеприведенные условия с учетом замены x_{ijt} на $(-x_{ijt} - \max_t x_{ijt})$. В противном случае ставится знак \bar{N} .

Вычисление функции сходства временных рядов

Для каждой пары рядов $\{u_{ijt}\}_{t=1}^T$ и $\{d_{kjt}\}_{t=1}^T$ вычисляется функция сходства $\rho_{ikj}^{\bar{U}}$, где $i, k = 1, \dots, I$ — номера объектов, а значение $j \in \{1, \dots, J\}$ — номер показателя. Значение этой функции равно числу интервалов, удовлетворяющих следующим условиям.

Рядам с индексами i, k поставлены в соответствие наборы $\{\mathfrak{I}_{i\xi j}^{\bar{U}}\}$ интервалов $\{\mathfrak{I}_{i\xi j}^{\bar{U}}\}, \{\mathfrak{I}_{k\zeta j}^{\bar{U}}\}$, составленные из последовательно идущих индексов знаков \bar{U} соответствующих рядов, где порядковые номера интервалов $\xi, \zeta \in \Xi = \{1, \dots, \lfloor \frac{T}{2} \rfloor\}$, а элементы $\mathfrak{I}_{i\xi j}^{\bar{U}}, \mathfrak{I}_{k\zeta j}^{\bar{U}} \subset \{1, \dots, T\}$. Значение функции сходства $\rho_{ikj}^{\bar{U}}$ равно числу интервалов, удовлетворяющих условию

$$\rho_{ikj}^{\bar{U}} = \# \left\{ (\xi, \zeta) \mid |\mathfrak{I}_{i\xi j}^{\bar{U}} \cap \mathfrak{I}_{k\zeta j}^{\bar{U}}| + \bar{\Delta} \geq |\mathfrak{I}_{i\xi j}^{\bar{U}} \cup \mathfrak{I}_{k\zeta j}^{\bar{U}}| \right\},$$

где $\bar{\Delta}$ — заданный параметр.

Функция сходства $\rho_{ikj}^{\bar{D}}$ для каждой пары рядов $\{d_{ijt}\}_{t=1}^T$ и $\{d_{kjt}\}_{t=1}^T$ вычисляется так же.

Полученные функции сходства $\rho_{ikj}^{\bar{U}}$ и $\rho_{ikj}^{\bar{D}}$ задают трехиндексную матрицу сходства $S^{\bar{U}\bar{D}} = \{\rho_{ikj}^{\bar{U}} + \rho_{ikj}^{\bar{D}}\}_{i,k,j=1}^{I,I,J}$. Каждый элемент этой матрицы является значением функции сходства, определенной на всех возможных парах объектов и на всех показателях.

Поиск группы объектов для одного показателя

Все действующие группы, включающие объекты, действующие сходным образом и имеющие на фиксированном показателе расстояние более \bar{R} , однозначно заданы полносвязным графом на матрице смежности S . Эта матрица получается путем порогового отсека значения элементов матрицы S^{UD} :

$$S = \{\rho_{ikj}\}_{i,k,j=1}^{I,I,J}, \text{ где } \rho_{ikj} = \begin{cases} 1, & \text{если } (\rho_{ikj}^{\text{U}} + \rho_{ikj}^{\text{D}}) \geq \bar{R}; \\ 0, & \text{в другом случае.} \end{cases}$$

Полносвязные графы, полученные на трехиндексной матрице сходства S , представимы в виде множества $\{G_{j\gamma}\}_{j,\gamma=1}^{J,\Gamma_j}$, где сходно действующая на j -й акции группа с номером $\gamma \in \{1, \dots, \Gamma_j\}$ есть набор объектов $G_{j\gamma} = \{g_{j\gamma 1}, \dots, g_{j\gamma l_{j\gamma}}\}$, где $g \in \{1, \dots, I\}$ — номер объекта.

Группы объектов для нескольких показателей

Группа объектов P_s сходно действует по нескольким показателям, если существует непустое пересечение элементов групп $G_{j\gamma}$ такое, что выполняются условия

$$P_s = \bigcap_{\substack{j \in \mathcal{J} \subseteq \{1, \dots, J\}, \\ \gamma \in \{1, \dots, \Gamma_j\}}} G_{j\gamma},$$

где $|\mathcal{J}| \rightarrow \max$, $|P_s| \rightarrow \max$, s — номер полученной группы.

Примером использования вышеприведенной методики может служить задача поиска группы магазинов торговой сети, в которых набор товаров продается схожим образом. Другим примером может служить поиск групп участников биржевых торгов, действующих синхронно на наборе финансовых инструментов.

Работа выполнена при поддержке РФФИ, проекты № 07-07-00181, № 07-07-00372.

Литература

- [1] Рудаков К. В., Чехович Ю. В. Алгебраический подход к проблеме синтеза обучаемых алгоритмов выделения трендов // Доклады РАН. — 2003. — Т. 388, № 1. — С. 33–36.
- [2] Васин Е. А., Костенко В. А., Коваленко Д. С. Автоматическое построение алгоритмов, основанных на алгебраическом подходе, для распознавания предаварийных ситуаций динамических систем // Искусственный интеллект. — 2006. — № 2. — С. 130–134.
- [3] Тамм У. Т. Теория графов. — Москва: Мир, 1988. — 314 с.