Метод кластерного анализа разнотипных временных рядов

В. Б. Бериков[1], И. А. Пестунов[2,*], М. К. Герасимов[1]

[1]Институт математики СО РАН, Новосибирск, Россия
[2]Институт вычислительных технологий СО РАН, Новосибирск, Россия
*Контактный e-mail: pestunov@ict.nsc.ru

Рассматривается задача разбиения множества многомерных временных рядов на группы похожих подмножеств (кластеров). Каждый временной ряд представляет собой описание характеристик некоторых объектов, изменяющихся с течением времени, при этом характеристики могут быть как количественными, так и качественными. В работе предложен способ задания меры различия между временными рядами с использованием деревьев решений. Также предложен алгоритм кластеризации временных рядов, использующий полученные матрицы различий.

Ключевые слова: многомерный временной ряд, кластерный анализ, деревья решений.

Введение

Одной из актуальных задач математического моделирования на основе анализа данных является задача кластеризации объектов, информация о которых представлена в виде многомерных разнотипных временных рядов. Под разнотипными рядами понимаются ряды, которые содержат наборы измерений вещественных (количественных), порядковых, номинальных или булевых (качественных) переменных. Такого рода задачи возникают при построении моделей объектов в трудноформализуемых областях исследований, например в медицине, когда требуется дать описание типичных групп пациентов со сходной динамикой развития заболевания на основе данных об изменениях клинических показателей и диагностических признаков. Типизация пациентов позволяет, в частности, разрабатывать методики лечения, оптимальные для каждой группы.

В задаче кластерного анализа требуется разбить множество объектов, описываемых набором некоторых переменных (либо матрицей попарных расстояний), на относительно небольшое число кластеров так, чтобы критерий качества группировки принял наилучшее значение. Число кластеров может быть как выбрано заранее, так и не задано (в последнем случае оптимальное количество групп должно быть определено автоматически). Под критерием качества обычно понимается некоторый функционал, зависящий от разброса объектов внутри группы и расстояний между группами.

В широко распространенных алгоритмах кластеризации (алгоритм k-средних, алгоритм построения дендрограммы и т. д.) так или иначе используются различные способы задания расстояния или меры различия между объектами и их группами. При выборе

конкретного способа определения расстояния исследователь, как правило, полагается на свои знания и опыт решения аналогичных задач. Определение расстояния или меры различия между временными рядами имеет дополнительные трудности: ряды могут быть разной длины, состоять из разнотипных компонентов, иметь большую размерность. Кроме того, предполагается наличие зависимостей между наблюдаемыми характеристиками в различные моменты времени.

Существует несколько основных подходов к кластеризации временных рядов [1]. Первый подход к определению расстояния/различия между временными рядами основан на простом сравнении наблюдений, в том числе и с помощью трансформации временной шкалы — DTW-алгоритмы [2], или некоторых статистик (среднего значения, коэффициентов корреляции), вычисленных по данным. Особенность другого направления состоит в сравнении параметров моделей процессов, реализацией которых являются наблюдаемые временные ряды (см., например, [3, 4]). В случае разнотипных данных применение этих подходов затруднительно из-за невозможности введения метрики в разнотипном пространстве.

В настоящей работе предлагается метод определения меры различия между временными рядами, основанный на деревьях решений в качестве аппроксимирующих функций. Данный подход дает возможность решения указанных выше проблем. Методы анализа временных рядов, основанные на деревьях решений [5], обладают положительными особенностями: не требуют априорных предположений о данных, могут обрабатывать как количественные, так и качественные характеристики, просты для понимания и интерпретации, могут сочетаться с другими методами принятия решений, такими как регрессионный и кластерный анализ. Методы кластеризации разнотипных объектов с использованием деревьев решений впервые предложены в работах [6, 7]. В работе [8] на основе байесовского подхода проведено теоретическое обоснование критерия качества таксономического дерева решений. В предлагаемой работе методы, основанные на деревьях решений, применяются для кластеризации разнотипных временных рядов.

Статья организована следующим образом. В разд. 1 дана постановка задачи и представлена мера различия между временными рядами, использующая деревья решений. В разд. 2 рассмотрены некоторые теоретические свойства предлагаемой меры. В разд. 3 описан алгоритм кластеризации временных рядов и приведены результаты численных экспериментов. В заключении изложены основные результаты работы и сформулированы направления дальнейших исследований.

1. Мера различия между временными рядами, основанная на деревьях решений

Пусть наблюдаются N объектов $a^1, \ldots, a^N$, изменяющихся во времени. Предположим, что в последовательные моменты времени $t^i, \ldots, T^i$ для объекта a^i проведены измерения значений количественной переменной Y. Обозначим через y_t^i значение переменной Y для объекта a^i в момент времени t.

Также предполагается, что наблюдения за объектами представлены некоторым набором переменных $\mathbf{X} = (X_1, \ldots, X_n)$, которые, возможно, влияют на целевую переменную Y. Объединим измерения значений этих переменных для объекта a^i в момент времени t в набор $\mathbf{x}_t^i$. В общем случае компоненты $\mathbf{X}$ могут быть количественными,

принимающими значения в R, или качественными переменными с конечным числом значений. Другими словами, имеются N разнотипных временных рядов

$$Y^i = \langle \mathbf{x}^i_{t^i}, y^i_{t^i}, \ldots, \mathbf{x}^i_{T^i}, y^i_{T^i} \rangle.$$

Через n^i обозначим длину временного ряда Y^i: $n^i = T^i - t^i + 1$, $i = 1, \ldots, N$. Отметим, что в общем случае $n^i \neq n^j$ для $i \neq j$.

В данной работе предполагается, что процессы, генерирующие данные, неизвестны и различны для отдельных временных рядов в том смысле, что существуют неизвестные функции $f_1, \ldots, f_K$, такие что

$$y^i_t = f_{\varphi(i)}(y^i_{t-1}, \ldots, y^i_{t-L}, \mathbf{x}^i_t, \ldots, \mathbf{x}^i_{t-L}) + \varepsilon^i_t,$$

где $i = 1, \ldots N$; L — известный лаг; $t = t^i + L, \ldots, T^i$; $\varphi(i) \in \{1, \ldots, K\}$; шумы ε^i_t являются независимыми непрерывными случайными величинами с нулевыми математическими ожиданиями и дисперсиями $D(\varepsilon^i_t) = \sigma^2$.

Задача состоит в определении подходящей меры различия между временными рядами и ее обосновании. Требуется также разработать алгоритм группировки временных рядов на основе введенного расстояния.

Зафиксируем индекс i и рассмотрим временной ряд Y^i. Обозначим

$$z^i_t = (y^i_{t-1}, \ldots, y^i_{t-L}, \mathbf{x}^i_t, \ldots, \mathbf{x}^i_{t-L}).$$

Тогда

$$y^i_t = f_{\varphi(i)}(z^i_t) + \varepsilon^i_t.$$

Используя обучающую выборку $V^i = \{(z^i_t, y^i_t)\}$, где $t = t^i + L, \ldots, T^i$, мы можем построить аппроксимацию g^i функции $f_{\varphi(i)}$. Для этого используем деревья решений [5, 9]; пример дерева изображен на рис. 1.

В вершинах дерева проверяются высказывания относительно некоторых переменных в определенный отсчет времени назад (относительно текущего момента). Цепочка проверяемых высказываний ведет из корня дерева в терминальную вершину, которой приписано прогнозируемое значение. Детали алгоритма построения дерева решений для прогнозирования временного ряда по его предыстории можно найти в [5, 9].

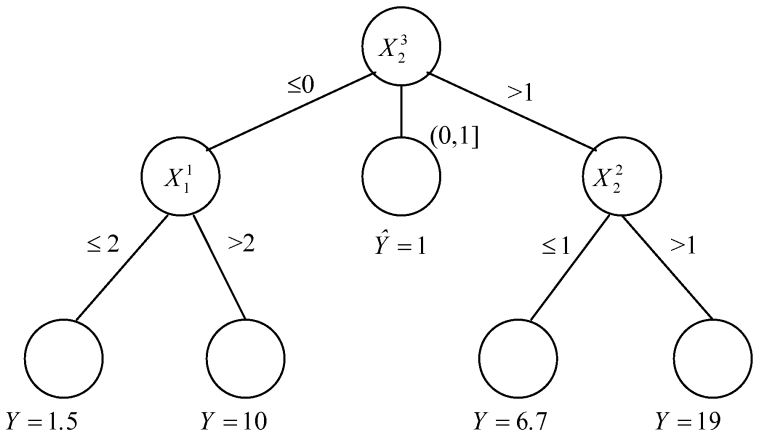

Рис. 1. Пример дерева решений

Аналогично для временного ряда Y^j имеется обучающая выборка $V^j = \{(z_t^j, y_t^j)\}$, где $t = t^j + L, \ldots, T^j$. Применим построенное решающее дерево g^i на обучающей выборке V^j.

Рассмотрим среднеквадратическую ошибку решающего дерева g^i:

$$\mu^{ij} = \frac{\sum\limits_{t=t^j+L}^{T^j} \left(g^i(z_t^j) - y_t^j\right)^2}{n^j - L}. \tag{1}$$

Используем следующее предположение: ошибка прогноза μ^{ij} будет меньше, если временные ряды Y^i и Y^j порождены одной и той же функцией, чем если бы они были порождены различными функциями. Следовательно, чем меньше ошибка μ^{ij}, тем более вероятно, что временные ряды Y^i и Y^j порождены одной и той же функцией и, значит, могут быть отнесены к одному кластеру.

Определим *меру различия* между временными рядами Y^i и Y^j как среднее значение величин μ^{ij} и $\mu^{j\,i}$:

$$\rho(i,j) = \begin{cases} \dfrac{\mu^{ij} + \mu^{j\,i}}{2}, & \text{если } i \neq j, \\ 0, & \text{иначе.} \end{cases}$$

2. Свойства меры различия

Введенная мера различия обладает следующими очевидными свойствами:

1) неотрицательность $\rho(i,j) \geq 0$;
2) тождественность $\rho(i,i) = 0$;
3) симметрия $\rho(i,j) = \rho(j,i)$.

Введем обозначения: r^- — максимальное математическое ожидание меры различия между временными рядами, принадлежащими одному кластеру:

$$r^- = \max_{i,j|\varphi(i)=\varphi(j)} E[\rho(i,j)];$$

r^+ — минимальное математическое ожидание меры различия между временными рядами, принадлежащими разным кластерам:

$$r^+ = \min_{i,j|\varphi(i)\neq\varphi(j)} E[\rho(i,j)];$$

$\xi^i(z)$ — ошибка аппроксимации функции $f_{\varphi(i)}$ деревом g^i в точке z:

$$\xi^i(z) = g^i(z) - f_{\varphi(i)}(z);$$

$\eta^{ij}(z)$ — разница между функциями $f_{\varphi(i)}$ и $f_{\varphi(j)}$ в точке z:

$$\eta^{ij}(z) = f_{\varphi(i)}(z) - f_{\varphi(j)}(z).$$

Заметим, что если $\varphi(i) = \varphi(j)$, то $\eta^{ij}(z) = 0\ \forall z$.

Утверждение 1. *Пусть* $\sup\limits_{i,z} |\xi^i(z)| = \xi$, $\inf\limits_{i,j,z|\varphi(i)\neq\varphi(j)} |\eta^{ij}(z)| = \eta$. *Тогда если* $\eta \geq 2\xi$, *то* $r^- \leq r^+$.

Доказательство. Слагаемые в числителе (1) обозначим как $\delta_t = \big(g^i(z_t^j) - y_t^j\big)^2$. Тогда

$$\delta_t = \big(g^i(z_t^j) - f_{\varphi(j)}(z_t^j) - \varepsilon_t^j\big)^2 = \big(g^i(z_t^j) - f_{\varphi(i)}(z_t^j) + f_{\varphi(i)}(z_t^j) - f_{\varphi(j)}(z_t^j) - \varepsilon_t^j\big)^2 =$$

$$= \big(\xi^i(z_t^j) + \eta^{ij}(z_t^j) - \varepsilon_t^j\big)^2 = \big(\xi^i(z_t^j) + \eta^{ij}(z_t^j)\big)^2 - 2\varepsilon_t^j\big(\xi^i(z_t^j) + \eta^{ij}(z_t^j)\big) + \big(\varepsilon_t^j\big)^2.$$

Рассмотрим математическое ожидание случайной величины δ_t:

$$E[\delta_t] = E\big[\big(\xi^i(z_t^j) + \eta^{ij}(z_t^j)\big)^2\big] - 2E\big[\varepsilon_t^j\big(\xi^i(z_t^j) + \eta^{ij}(z_t^j)\big)\big] + E\big[\big(\varepsilon_t^j\big)^2\big].$$

Так как ε_t^j и $\xi^i(z_t^j)$ — независимые случайные величины, то

$$E\big[\varepsilon_t^j\xi^i(z_t^j)\big] = E\big[\varepsilon_t^j\big]E\big[\xi^i(z_t^j)\big] = 0.$$

Аналогично, ε_t^j и $\eta^{ij}(z_t^j)$ — независимые случайные величины, следовательно,

$$E\big[\varepsilon_t^j\eta^{ij}(z_t^j)\big] = E\big[\varepsilon_t^j\big]E\big[\eta^{ij}(z_t^j)\big] = 0.$$

Отсюда получаем

$$E[\delta_t] = E\big[\big(\xi^i(z_t^j) + \eta^{ij}(z_t^j)\big)^2\big] + \sigma^2.$$

Предположим, что $\varphi(i) = \varphi(j)$. Тогда $\eta^{ij}(z_t^j) = 0$. Поэтому

$$E[\delta_t] = E\big[\big(\xi^i(z_t^j)\big)^2\big] + \sigma^2 \le \xi^2 + \sigma^2.$$

Следовательно,

$$E\big[\mu^{ij}\big] = \frac{\sum\limits_{t=t^j+L}^{T^j} E\big[\delta_t\big]}{n^j - L} \le \xi^2 + \sigma^2 \Rightarrow E\big[\rho(i,j)\big] \le \xi^2 + \sigma^2 \quad \forall i,j|\varphi(i) = \varphi(j).$$

В результате имеем $r^- \le \xi^2 + \sigma^2$.

Теперь предположим, что $\varphi(i) \ne \varphi(j)$. В этом случае, учитывая цепочку неравенств $|\eta^{ij}(z_t^j)| \ge \eta \ge 2\xi \ge 2|\xi^i(z_t^j)| \ge |\xi^i(z_t^j)|$, получим

$$|\xi^i(z_t^j) + \eta^{ij}(z_t^j)| \ge \big||\eta^{ij}(z_t^j)| - |\xi^i(z_t^j)|\big| = |\eta^{ij}(z_t^j)| - |\xi^i(z_t^j)| \ge \eta - \xi \ge \xi \Rightarrow$$

$$\Rightarrow E[\delta_t] \ge \xi^2 + \sigma^2 \Rightarrow E\big[\rho(i,j)\big] \ge \xi^2 + \sigma^2 \quad \forall i,j|\varphi(i) \ne \varphi(j).$$

В результате имеем $r^+ \ge \xi^2 + \sigma^2$.

Объединяя полученные неравенства, получим $r^- \le r^+$. Утверждение доказано.

Из утверждения следует, что если решающие деревья достаточно точно аппроксимируют исходные порождающие функции, при этом сами порождающие функции достаточно сильно различаются между собой, то временные ряды из одного кластера будут в среднем более похожи друг на друга (с точки зрения предложенной меры различия), чем временные ряды из разных кластеров.

3. Алгоритм кластеризации временных рядов

После вычисления мер различия между всеми парами временных рядов получим $N \times N$-матрицу $M = (m_{ij})$, где $m_{ij} = \rho(i, j)$. Далее матрица M используется на входе иерархического агломеративного алгоритма кластеризации для итогового разбиения на кластеры. Требуемое число кластеров — заданный параметр K.

Итак, общая схема алгоритма кластеризации временных рядов выглядит следующим образом:

- Для каждого временного ряда Y^i на основе обучающей выборки V^i строится решающее дерево g^i.
- Для каждого решающего дерева g^i вычисляются среднеквадратические ошибки μ^{ij} на обучающих выборках V^j.
- Для всех временных рядов вычисляются попарные меры различия $\rho(i, j)$. Определяется матрица M.
- Используя матрицу попарных мер различия M, стандартным алгоритмом иерархической кластеризации строим разбиение исходных временных рядов на кластеры.

Эффективность предложенного подхода была исследована методами статистического моделирования. Использовалась следующая схема моделирования.

- Фиксируется число кластеров K, число объектов N, начальные и конечные моменты времени измерений каждого временного ряда t_i и T_i, истинное разбиение на кластеры φ_i, $i = 1, \ldots, N$, порождающие функции $F_1, \ldots, F_K$, тип шума и его дисперсия σ^2.
- Генерируются N временных рядов.
- Предложенным в работе способом строится разбиение временных рядов на кластеры.
- Определяется качество группировки.

В качестве примера приведем результаты серии экспериментов со следующими параметрами: число кластеров $K = 3$, число временных рядов $N = 15$, $t_i = 1$, $T_i = 30$, $\varphi_i = \lceil i/5 \rceil$ для всех i. Использовались линейные порождающие функции от трех переменных: $f_k(u_1, u_2, u_3) = a_k u_1 + b_k u_2 + c_k u_3 + d_k$, $k = 1, 2, 3$; коэффициенты приведены в таблице.

Распределение шума ε_t^i — стандартное нормальное. Дисперсия шума изменялась от 0.1 до 0.9 с шагом 0.01. Для каждого фиксированного значения дисперсии эксперимент повторялся 40 раз, затем результаты усреднялись.

Временные ряды генерировались по формуле

$$Y_t^i = a_{\varphi(i)} Y_{t-1}^i + b_{\varphi(i)} Y_{t-2}^i + c_{\varphi(i)} Y_{t-3}^i + d_{\varphi(i)} + \varepsilon_t^i.$$

Построены аппроксимирующие решающие деревья для каждого временного ряда, вычислены попарные меры различия временных рядов, из них составлена матрица M.

Коэффициенты

k	a_k	b_k	c_k	d_k
1	0.2	0.3	0.4	0.5
2	0.1	−0.3	0.4	0.1
3	−0.2	0.3	0.4	0.5

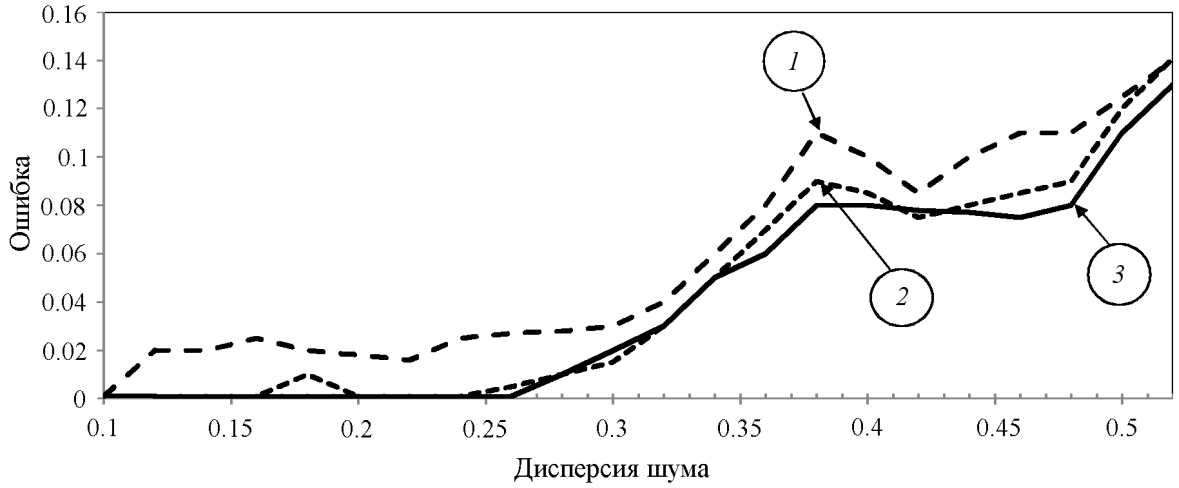

Рис. 2. Результаты эксперимента: ошибка для различных вариантов расстояния между кластерами (*1* — по ближайшему соседу, *2* — по дальнему соседу, *3* — по среднему расстоянию)

В качестве алгоритма кластеризации использовался иерархический агломеративный алгоритм с различными вариантами вычисления расстояния между кластерами (метод ближайшего соседа, метод дальнего соседа, среднее расстояние).

Качество метода оценивалось исходя из доли пар рядов, ошибочно причисленных к одному кластеру или к разным кластерам. На рис. 2 представлены результаты эксперимента. Видно, что при умеренном уровне шума ошибка кластеризации достаточно мала и слабо зависит от варианта определения расстояния между группами.

Заключение

Предложен метод кластеризации многомерных временных рядов, основанный на деревьях решений, которые использованы в качестве аппроксимирующих функций при определении меры различия между рядами. Исследованы свойства введенного расстояния. В частности, доказано, что если решающие деревья достаточно точно аппроксимируют порождающие ряды функции и при этом сами порождающие функции сильно различаются между собой, то временные ряды из одного кластера будут в среднем более похожи друг на друга, чем временные ряды из разных кластеров.

Разработанный метод, в отличие от других существующих, позволяет анализировать многомерные разнотипные временные ряды, состоящие из количественных и качественных характеристик, изменяющихся с течением времени и влияющих друг на друга.

В качестве направлений дальнейших исследований можно указать рассмотрение различных вариантов расстояния между рядами. Перспективным представляется использование дополнительной информации о задаче, имеющей вид экспертных логических высказываний, получаемых из некоторой базы знаний, с последующим вычислением расстояний между высказываниями [10]. Для улучшения качества (“устойчивости”) решений планируется использовать ансамбль алгоритмов кластеризации [11].

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты № 14-07-00249а, 14-07-00851а), РНФ (грант № 14-14-00453) и фонда В. Потанина.

Список литературы / References

[1] **Aggarwal, C., Reddy, C.** Data Clustering: Algorithms and Applications. CRC Press, 2013. 652 p.

[2] **Meesrikamolkul, W., Niennattrakul, V., Ratanamahatana, C.** Shape-based clustering for time series data // Proc. 16th Pacific-Asia Conf., PAKDD 2012, Kuala Lumpur, Malaysia, May 29 – June 1, 2012. Part I. P. 530–541.

[3] **Corduas, M., Piccolo, D.** Time series clustering and classification by the autoregressive metric // Comput. Statistics & Data Analysis. 2008. Vol. 52(4). P. 1860–1872.

[4] **Ghassempour, S., Girosi, F., Maeder, A.** Clustering Multivariate Time Series Using Hidden Markov Models // Intern. J. Environ. Res. Publ. Health. 2014. Vol. 11(3). P. 2741–2763.

[5] **Лбов Г.С., Бериков В.Б.** Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Изд-во Ин-та математики, 2005. 218 с.
Lbov, G.S., Berikov, V.B. Stability of decision functions in problems of pattern recognition and analysis of heterogeneous information. Novosibirsk: Izd-vo In-ta Matematiki, 2005. 218 p. (in Russ.)

[6] **Лбов Г.С., Пестунова Т.М.** Группировка объектов в пространстве разнотипных признаков // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 141–149.
Lbov, G.S., Pestunova, T.M. Grouping of objects in the space of heterogeneous features // Analysis of Non-numeric Information in Sociological Researches. Moskow: Nauka, 1985. P. 141–149. (in Russ.)

[7] **Лбов Г.С., Пестунова Т.М.** Построение дерева разбиений в задаче группировки объектов с использованием логических функций // Вычисл. системы. 1986. Вып. 117. С. 63–77.
Lbov, G.S., Pestunova, T.M. Construction of partition tree in the problem of grouping of objects with use of logical functions // Vychisl. Sistemy. 1986. Vyp. 117. P. 63–77. (in Russ.)

[8] **Berikov, V.B.** Grouping of objects in a space of heterogeneous variables with the use of taxonomic decision trees // Pattern Recognition and Image Analysis. 2011. Vol. 21, No. 4. P. 591–598.

[9] **Бериков В.Б., Пестунов И.А., Герасимов М.К.** Анализ совокупности разнотипных временных рядов с использованием логических решающих функций // Вычисл. технологии. 2012. Т. 17, №. 5. С. 12–22.
Berikov, V.B., Pestunov, I.A., Gerasimov, M.K. Analysis of a set of heterogeneous time series with use logical decision functions // Comput. Technologies. 2012. Vol. 17, No. 5. P. 12–22. (in Russ.)

[10] **Vikent'ev, A.A.** Distances and degrees of uncertainty in many-valued propositions of experts and application of these concepts in problems of pattern recognition and clustering // Pattern Recognition and Image Analysis. 2014. Vol. 24, No. 4. P. 489–501.

[11] **Berikov, V.B.** Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters. 2014. Vol. 38. P. 99–106.

Поступила в редакцию 30 декабря 2014 г.

Method for clustering of heterogeneous time series

Berikov, Vladimir B.[1], Pestunov, Igor A.[2,*], Gerasimov, Maxim K.[1]

[1]Sobolev Institute of Mathematics of the SB RAS, Novosibirsk, 630090, Russia
[2]Institute of Computational Technologies SB RAS, Novosibirsk, 630090, Russia
[*]Corresponding author: Pestunov, Igor A., e-mail: pestunov@ict.nsc.ru

Purpose. The paper addresses the problem of partitioning of a set of multidimensional time series on groups of similar subsets (clusters). Each time series represents characteristics (qualitative or quantitative) of an object that changes in time. By assumptions, the data generating mechanism is unknown and may vary across the set of time series in the sense that the observed values of individual time series depend on one of the unobserved generative functions.

Methodology. In this paper, we suggest a way to define a measure of difference between time series with the help of decision trees as approximation functions. The proposed dissimilarity measure satisfies some useful properties such as non-negativity, identity, and symmetry.

Findings. We suggest a mathematical model of data generating mechanism and prove that if we have good approximations of initial well-distinguished generative functions then time series from same clusters are more similar to each other (in the sense of the proposed dissimilarity measure) than series from different clusters.

Originality\value. The suggested approach makes it possible to determine distance/dissimilarity measure between time series with heterogeneous components, different lengths, large sizes and dimensions along with the interdependencies between observation values at different time points. The approach does not rely on prior assumptions about the data. It is simple to understand and interpret and can be combined with other decision making techniques such as regression analysis and clustering. The algorithm of time series clustering that utilizes the obtained dissimilarity matrix is also suggested.

Keywords: multidimensional heterogeneous time series, cluster analysis, decision trees.

Acknowledgements. This work was financially supported by RFBR (projects No. 14-07-00249a, 14-07-00851a), RSF (grant No. 14-14-00453) and V. Potanin Foundation.

Received 30 December 2014