УДК 519.237.8+519.246.8

Применение коэффициента временной корреляции для оценки монотонности многомерных временных рядов

Р.Д. Зайцев
Московский физико-технический институт (государственный университет)

В последние несколько лет активно развивается область машинного обучения, получившая в англоязычной литературе название Time Series Data Mining [1]. Хотя общепринятый перевод на русский в настоящее время не устоялся, можно понять, что речь идёт о задачах интеллектуального анализа, в которых объекты выборки представлены временными рядами. Временные ряды существенно многомерны, и к настоящему времени предложено довольно много разнообразных расстояний и мер различия, учитывающих разнообразные свойства рядов (такие, например, как периодичность, монотонность, автокорреляцию и спектральные характеристики) [2]. Однако обобщение многих мер на случай многомерных временных рядов до сих пор не достаточно проработано [3].

В рамках данной работы автор рассматривает класс непараметрических мер различия, основанных на оценке, с одной стороны, близости абсолютных значений рядов (например, Евклидово или DTW-расстояние [3]); и с другой стороны, учитывающих свойство монотонности рядов с помощью так называемого коэффициента временной корреляции (temporal correlation coefficient), предложенного Chouakria и Nagabhushan в 2007 [4]. В исходном варианте этот подход позволяет сравнивать монотонность одномерных временных рядов. Автором предложен подход по обобщению временной корреляции на случай многомерных рядов, который в дальнейшем может быть использован для улучшения результатов кластеризации при использовании многомерных непараметрических расстояний. Исследованы некоторые свойства временной корреляции, и преимущество этого подхода перед аналогичным – корреляцией рядов приращений.

Коэффициент временной корреляции (в дальнейшем - КВК) был предложен как мера оценки синхронности поведения (монотонности) одномерных временных рядов на соответствующих участках. Основой для подобного метода стало понятие временной дисперсии, использованное тем же автором для оценки потерь информации при сжатии многомерных временных рядов [5].

Для одномерных рядов X, Y длины T временная корреляция может быть оценена следующим образом:

$$cor_t = \frac{\sum_{t=2}^{T}(x_t - x_{t-1})(y_t - y_{t-1})}{\sqrt{\sum_{t=2}^{T}(x_t - x_{t-1})^2}\sqrt{\sum_{t=2}^{T}(y_t - y_{t-1})^2}}$$

КВК не превышает по модулю единицу, и показывает синхронность знака прироста двух рядов для соответствующих лагов, не учитывая близость значений ряда. Классическая линейная корреляция Пирсона, с другой стороны, берёт в расчёт только близость исходных значений, но не учитывает монотонности, что было также показано авторами [4].

В отличие от схожего подхода учёта монотонности – корреляции рядов приращений - информация об общем приросте при использовании временной корреляции не отбрасывается, что служит дополнительным основанием для использования КВК. Действительно, рассмотрим корреляцию рядов приращений X', Y', полученных из исходных рядов взятием разностей.

По определению, $cor(X', Y') = \frac{\mathrm{cov}(X', Y')}{\sqrt{D_{X'} D_{Y'}}}$.

Преобразовав выражение, получим

$$cor(X', Y') = \frac{\sum_{t=1}^{T}(x_{t+1} + x_t)(y_{t+1} - y_t) - n\delta_x\delta_y}{\sqrt{(\sum_{t=1}^{T}(x_{t+1} - x_t)^2 - n\delta_x^2)(\sum_{t=1}^{T}(y_{t+1} - y_t)^2 - n\delta_y^2)}}$$

где δ_x^2, δ_y^2 – средние приросты рядов. Это можно интерпретировать следующим образом – при таком подходе из общей временной корреляции вычитается корреляция трендов рядов.

Предложенная Chouakria и Nagabhushan [4] составная мера позволяет помимо близости исходных значений ряда учитывать и монотонность путём корректировки Евклидова расстояния или расстояния DTW на преобразованный коэффициент временной корреляции. В предыдущей работе автором показано, что в таком случае можно добиться существенного улучшения качества одномерных кластеризационных моделей [6].

В данной работе предложен способ обобщения КВК на многомерный случай с помощью простого усреднения коэффициентов по всем m параметрам:

$$CORT_t^m = \frac{1}{m}\sum_{i=1}^{m} cor_t^i$$

Итоговый коэффициент также не превышает единицу по модулю. Корень из суммы квадратов КВК по каждому атрибуту в данном случае представляется малоэффективным, поскольку значительная часть информации (например, об отрицательной временной корреляции) будет потеряна.

Для оценки метода была сконструирована синтетическая выборка, при этом использовался подход, схожий с использовавшимся для обоснования свойств одномерного КВК [4]. Всего было сгенерировано 200 трёхмерных рядов, где каждый компонент был представлен случайной функцией, принадлежащей одному из трёх семейств:

$$f_1 = gt + 2t + 3 + \varepsilon$$
$$f_2 = mean(gt) - gt + 2t + 3 + \varepsilon$$
$$f_3 = 4gt - 3 + \varepsilon$$

где ε – случайная ошибка ~ N(0,1), t – время, а g – заранее сгенерированная общая последовательность длиной 40 из равномерного целочисленного распределения [3;8]. При этом, как можно видеть, типы f_1 и f_3 строго монотонны в смысле временной корреляции, так как включают компонент g с одинаковым знаком. Тип f_2 строго антимонотонный для остальных двух типов. Были составлены следующие семейства трёхмерных рядов:

$$F_1 = [f_1; f_2; f_3] - 100\,рядов$$
$$F_2 = [f_2; f_1; f_2] - 40\,рядов$$
$$F_3 = [f_2; f_1; f_1] - 40\,рядов$$
$$F_4 = [f_3; f_2; f_1] - 20\,рядов$$

Таким образом, в выборке представлены все возможные случаи монотонности и антимонотонности компонентов для рассматриваемых исходных семейств. Действительно, у F_1 с F_2 все компоненты антимонотонны, как и у F_2 с F_4; у F_1 с F_3 и F_3 с F_4 монотонен один компонент, и антимонотонны остальные; у F_2 с F_3 монотонны два компонента и, наконец, у F_1 с F_4 монотонны все компоненты.

Распределение попарных коэффициентов временной корреляции для отдельных групп и для всей выборки показано на рис. 1, и рис. 2. Видно, что обобщённый таким образом КВК позволяет различить группы соответствия рядов между собой, причём величина дисперсии коэффициентов в каждой группе зависит от величины случайной ошибки ε при генерации выборки. Так, при полной монотонности КВК лежит в пределах [0.7,0.9], и этим объясняется высота четвёртого пика – туда попали также все попарные коэффициенты для рядов из одной и той же группы. Небольшой пик в единице – диагональные элементы матрицы КВК. При полной антимонотонности КВК лежит в пределах [-0.85,-0.65], при двух антимонотонных и двух монотонных компонентах – [-0.3,-0.1] и [0.1,0.35] соответственно.

При этом основной недостаток усреднения КВК – невозможность тонкого различия внутренней структуры рядов. Например, итоговый многомерный КВК будет совпадать в следующих случаях:

$$\frac{(cor_t^1 = 0.3) + (cor_t^2 = 0.3) + (cor_t^3 = 0.3)}{3} = 0.3 = \frac{(1) + (1) + (-1)}{3}$$

В дальнейшем планируется разработка метода обобщения, который поможет устранить указанный недостаток с помощью более гибкого учёта внутренней структуры сравниваемых многомерных рядов.

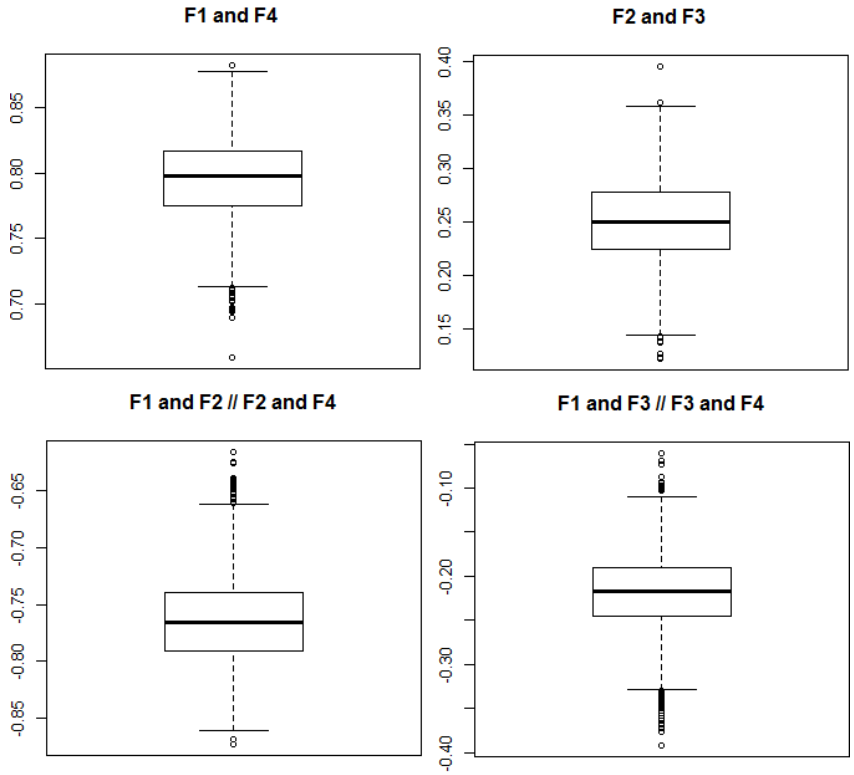

Рис. 1. Диаграммы размаха (boxplots) для каждой из групп

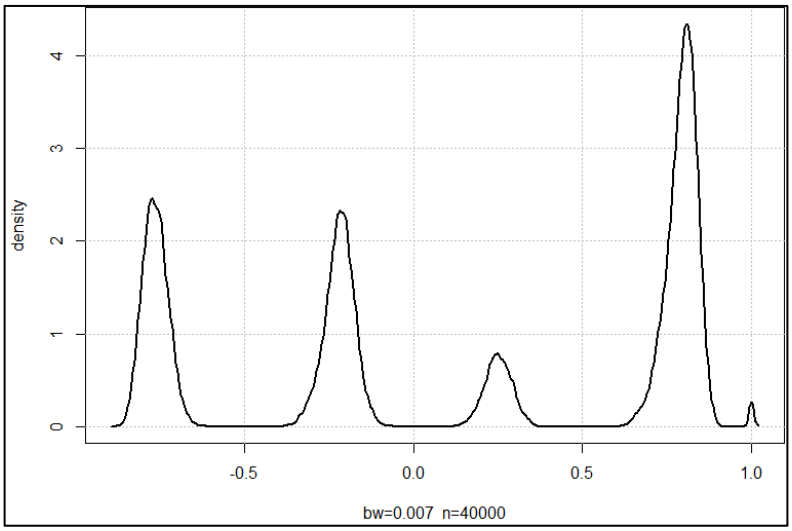

Рис. 2. Сглаженная эмпирическая плотность всех попарных КВК синтетической выборки

Литература

1. *Esling P., Agon C.* Time-series data mining //ACM Computing Surveys (CSUR). 2012. – Т. 45. – №. 1. – С. 12.
2. *Montero P., Vilar J.* TSclust: An R Package for Time Series Clustering //Journal of Statistical Software. 2014. № 62(1), 1–43.
3. *Aggarwal C. C., Reddy C. K.* (ed.). Data clustering: algorithms and applications. – CRC Press, 2013.
4. *Chouakria A. D., Nagabhushan P. N.* Adaptive dissimilarity index for measuring time series proximity //Advances in Data Analysis and Classification. 2007. – Т. 1. – №. 1. – С. 5-21.
5. *Chouakria-Douzal A.* Compression technique preserving correlations of a multivariate temporal sequence //International Symposium on Intelligent Data Analysis. Springer Berlin Heidelberg, 2003. – С. 566-577.
6. *Зайцев Р. Д., Бритков В. Б.* Применение языка R для многомерной кластеризации временных рядов с целью анализа динамики научно-технического развития //Труды Второй молодежной научной конференции «Задачи современной информатики». 2015. - С. 92-98.