

УДК 519.2:004.9

Л. О. Кириченко, А. Е. Ткаченко, Т. А. Радивилова

КЛАСТЕРИЗАЦИЯ ЗАШУМЛЕННЫХ ВРЕМЕННЫХ РЯДОВ

Аннотация. Проведен сравнительный анализ кластеризации зашумленных временных рядов выборки. Выборка для кластеризации содержала временные ряды различных типов, среди которых присутствовали нетипичные объекты. Кластеризация проводилась методами *k*-средних и DBSCAN с использованием различных функций расстояния для временных рядов.

Ключевые слова: кластеризация, временной ряд, функция расстояния, метод *k*-средних, метод DBSCAN.

Введение и цель

Одной из актуальных задач машинного обучения является задача кластеризации объектов. Кластеризация временных рядов используется как самостоятельная исследовательская техника, а также как часть более сложных методов интеллектуального анализа данных, такие как обнаружение правил, классификация, выявление аномалий и т.д. [1-4]. В задаче кластерного анализа временных рядов требуется разбить множество объектов на относительно небольшое число кластеров так, чтобы критерий качества группировки принял наилучшее значение. Под критерием качества обычно понимается некоторый функционал, зависящий от разброса объектов внутри кластеров и расстояний между ними. Способы задания расстояния или меры различия между объектами также являются различными [2-5].

Одной из задач кластеризации временных рядов, является выделение в отдельный кластер аномальных объектов. [6,7]. Это не является простым заданием, особенно в условиях зашумленности временных рядов. Целью данной работы является проведение сравнительного анализа кластеризации зашумленных временных рядов с нетипичными объектами с использованием нескольких методов кластеризации и различных функций расстояния.

Методы исследования

Метод *k*-средних является одним из широко используемых методов кластеризации [2,5]. Вначале мы задаем количество кластеров и соответствующие центроиды для каждого из них. Центроиды – это заданные главные объекты, относительно которых распределяются объекты по кластерам, первоначально они могут быть выбраны случайно. Далее распределяем все объекты по кластерам согласно близости к центрам. Итеративность метода заключается в том, что после каждого распределения относительно центроидов

мы их пересчитываем и повторяем весь процесс с начала. Продолжаем процесс до тех пор, пока центроиды не перестанут меняться.

С точки зрения вычислительной сложности алгоритм довольно прост. Недостатком является то, что количество кластеров не меняется и результат зависит от начальных центроидов. Это значит, что мы можем получить в кластере такие объекты, которые на самом деле не являются близкими к их центроиду.

Метод DBSCAN (Density-based spatial clustering of applications with noise). Суть метода – распределить достаточно близкие объекты по кластерам относительно плотности распределения объектов [1,2]. Первоначально задается радиус близости и минимальное количество точек, которые должны находиться внутри этого радиуса. Достаточно близкими или плотно расположенными являются объекты, которые находятся на расстоянии меньше заданного радиуса. Они выделяются в кластеры. Шумом являются объекты, которые не схожи ни с какими из выделенных. Объекты, которые являются шумом, распределяются по кластерам следующим образом: если в заданном радиусе от шумного объекта нет ни одного объекта, то он определяется в отдельный кластер. Если в радиусе находится один или несколько шумных объектов, то они объединяются в один кластер.

Одним из достоинств метода DBSCAN является возможность выделять нетипичные объекты выборки. Одним из недостатков является то, что объекты, которые являются шумом, могут определяться как принадлежащие какому-либо кластеру.

Функции расстояния. При кластеризации временных рядов необходимо использовать особые метрики. Одной из наиболее простых и популярных метрик является метрика Эвклида:

$$E(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

где X, Y – временные ряды длины n .

Существуют ситуации, когда два временных ряда в целом схожи, но сильно отличаются для некоторых значений времени. Если необходимо считать эти ряды похожими, можно использовать основную метрику и функцию Complexity Invariance Distance (CID):

$$CID(X, Y) = D(X, Y) \times CF(X, Y), \quad CF(X, Y) = \frac{\max\{CE(X), CE(Y)\}}{\min\{CE(X), CE(Y)\}},$$

$$CE(X) = \sum_{i=2}^n \sqrt{(x_i - x_{i-1})^2},$$

где X, Y – временные ряды длины n , $D(X, Y)$ – основная метрика.

Если необходимо сравнивать временные ряды разной длины, то можно использовать метрику Minimum Jump Cost (MJC). Основная идея метрики – необходимо найти сумму минимальных «скачков» между рядами.

Пусть имеется два ряда x и y длины N и M соответственно. Берем начальную точку $x(i)$, $i = 0$. Далее находим такую компоненту j ряда y , чтобы $(x(i) - y(j))^2 \rightarrow \min$, причем $i < j$. На следующей итерации $i = j + 1$,

$$MJC(X, Y) = \sum_i c_{\min}^i, c_{\min}^i = \min \left(c_{t_x}^{t_y}, c_{t_x}^{t_y+1}, \dots, c_{t_x}^{t_y+N} \right),$$

где $c_{t_x}^{t_y}$ – это всевозможные «скачки», которые мы вычисляем по формуле

$$c_{t_x}^{t_y+\Delta} = \left(x_{t_x} - y_{t_y+\Delta} \right)^2.$$

Функционалы качества. Для проверки качества кластеризации необходимо проверить, насколько похожие объекты находятся в одном кластере и насколько разные объекты в разных кластерах. Функционал качества – это некоторая функция, которая характеризует степень приближенности результатов кластеризации к идеальному решению.

Сумма внутрикластерных расстояний. Это сумма расстояний между объектами, которые находятся в одном кластере:

$$F_0 = \frac{\sum_{i=1}^n \sum_{j=i+1}^n \rho(x_i, x_j) u(x_i, x_j, y_i, y_j)}{\sum_{i=1}^n \sum_{j=i+1}^n u(x_i, x_j, y_i, y_j)},$$

где n – количество объектов, $\rho(x_i, x_j)$ – заданная функция расстояния между объектами, $u(x_i, x_j, y_i, y_j)$ – функция принадлежности объектов x_i к кластеру y_i .

Сумма межкластерных расстояний. Это сумма расстояний между объектами, которые находятся в разных кластерах:

$$F_1 = \frac{\sum_{i=1}^n \sum_{j=i+1}^n \rho(x_i, x_j) (1 - u(x_i, x_j, y_i, y_j))}{\sum_{i=1}^n \sum_{j=i+1}^n (1 - u(x_i, x_j, y_i, y_j))}.$$

Если стоит задача сравнить несколько полученных разбиений на кластеры, то лучшим будет тот, у которого будет минимальным $\frac{F_0}{F_1}$.

Результаты исследования

В работе был проведен численный эксперимент для исследования применения методов k-means и DBSCAN к модельным временным рядам с аддитивным белым шумом. Выборка, на которой проводилась кластеризация, состояла из m временных рядов различного типа: гармонические реализации, параболические реализации и «всплески». Для аддитивного зашумления временных рядов были использованы реализации белого шума с нормальным

распределением $N(0,\sigma)$. Дисперсия нормального распределения изменялась и имела значения $\sigma^2=\{0.5, 0.75, 1.0, 1.25\}$. На рис.1 представлены некоторые типичные «чистые» и зашумленные реализации для кластеризации, $\sigma^2=1$.

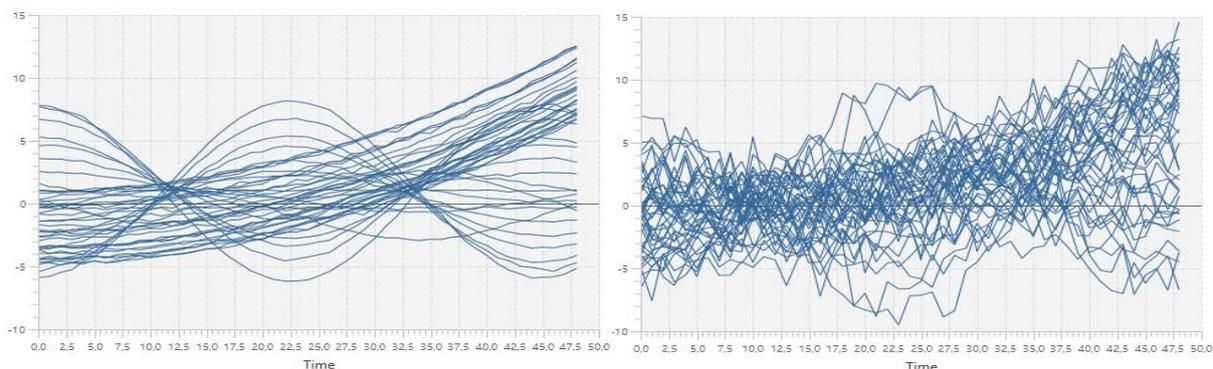


Рис.1 Типичные реализации

Как хороший результат кластеризации, было ожидаемо, что на выходе мы получим минимум 3 кластера. В первый кластер попадут ряды, сгенерированные с помощью на основе гармонических колебаний, во второй – квадратичные кривые и отдельно будут вынесены в третий кластер аномальные объекты типа всплесков.

Для исследований были выбраны методы кластеризации DBSCAN и k-means. В методе k-means было задано 3 центра. Входные параметры метода DBSCAN подбирались экспериментально. Для сравнения близости временных рядов были использованы метрики MJC и метрика Эвклида с функцией CID. Результаты эксперимента оценивались как визуальной проверкой попадания объектов в нужный кластер, так и с помощью функционалов качества.

Кластеризация «чистых» временных рядов.

Особенностью заданной выборки является наличие нетипичных реализаций (всплесков). Согласно полученным результатам, вынести такие объекты в отдельный кластер успешно получилось только с помощью метода DBSCAN, не смотря на то, что для в методе k-means одним из начальных центров задавался нетипичный объект. В результатах, полученных методом k-means, в одном кластере были смешаны временные ряды разной формы, также среди них были и нетипичные объекты. Среди выбранных метрик для сравнения временных рядов наилучшие результаты были получены с помощью метрики Эвклида с функцией CID. В табл.1 представлены количественные показатели качества кластеризации. Надо отметить, что численные эксперименты показали, что малые значения величины F_0/F_1 для метода k-means соответствуют неправильному распределению объектов по кластерам.

Таблица 1. Показатели качества кластеризации для «чистых» рядов

Метрика	Метод	F_0/F_1
---------	-------	-----------

Euclidean+ CID	K-means	0.425
	DBSCAN	0.336
MJC + CID	K-means	0.194
	DBSCAN	0.490

Кластеризация зашумленных данных

В данном случае результаты аналогичны результатам, полученным для чистых данных. Наилучшие показатели у метода DBSCAN с метрикой Эвклида и CID. Во всех остальных случаях в одном кластере оказываются ряды, которые отличаются по форме.

Таблица 2 демонстрирует изменение количественных показателей качества при возрастании дисперсии шума для кластеризации. Значение $\sigma^2=0$ соответствует выборке с «чистыми» реализациями. Очевидно, что несмотря на достаточно большой уровень шума, метод DBSCAN показывает корректное разбиение на кластеры. Это позволяет использовать данный метод для кластеризации реальных данных, которые обычно являются зашумленными.

Таблица 2. Показатели качества кластеризации для «чистых» рядов

σ^2	0	0.5	0.75	1	1.25
F_0/F_1	0.336	0.627	0.667	0.694	0.774

Выводы

В работе была проведена кластеризация зашумленных временных рядов различных типов. Были использованы методы DBSCAN и k-средних с различными функциями расстояния. Лучшие результаты показал метод DBSCAN с евклидовой метрикой и CID-функцией.

Анализ результатов кластеризации временных рядов позволяет определить ключевые различия между методами: если можно определить количество кластеров и не требуется отделять нетипичные временные ряды, метод k-средних показывает довольно хорошие результаты; если нет информации о количестве кластеров и существует задача выделения нетипичных рядов, целесообразно использовать метод DBSCAN.

Список литературы

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.J.: Time-series clustering. A Decade Review Information systems 53, 16-38 (2015).
2. Aggarwal, C., Reddy, C.: Data Clustering: Algorithms and Applications. CRC Press (2013).
3. Liao, T.W. Clustering of time series data – a survey. Pattern Recognition, 38 (11), 1857-1874 (2005).
4. Rani, S., Sikka, G.: Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications 52 (15), 1-9 (2012). doi: 10.5120/8282-1278

-
5. Grabusts, P., Borisov, A.: Clustering methodology for time series mining (2009). *Scientific Journal of Riga Technical University* 40, 81-86 (2009).
 6. Barreto, G., Aguayo, L.: Time Series Clustering for Anomaly Detection Using Competitive Neural Networks. In: *Proceeding WSOM '09 Proceedings of the 7th International Workshop on Advances in Self-Organizing Maps*, St. Augustine, FL, USA, 28-36 (2009).
 7. Nascimento, E.S., Tavares, O.L., Souza, A.F.: A Cluster-based Algorithm for Anomaly Detection in Time Series Using Mahalanobis. In: *ICAI'2015 International Conference on Artificial Intelligence 2015*, Las Vegas, USA 622-628 (2015).

УДК 519.2:004.9

Кириченко Л.О., Ткаченко А.Е., Радивилова Т.А. **Кластеризация зашумленных временных рядов** // Системные технологии. Региональный межвузовский сборник научных работ. - Выпуск(??).-Днепр, 2019. –С. ??-??.

Проведен сравнительный анализ кластеризации зашумленных временных рядов выборки. Выборка для кластеризации содержала временные ряды различных типов, среди которых присутствовали нетипичные объекты. Кластеризация проводилась методами k-средних и DBSCAN с использованием различных функций расстояния для временных рядов.

Библ.7 , рис. 1, табл.1 .

УДК 519.2:004.9

Кіриченко Л.О., Ткаченко А.Є., Радівілова Т.А. **Кластеризація зашумлених часових рядів** // Системні технології. Регіональний міжвузівський збірник наукових робіт. – Випуск (??).-Дніпро, 2019. –С. ??-??.

Проведено порівняльний аналіз кластеризації зашумлених часових рядів вибірки. Вибірка для кластеризації містила часові ряди різних типів, серед яких були присутні нетипові об'єкти. Кластеризація проводилася методами k-середніх і DBSCAN з використанням різних функцій відстані для часових рядів.

Бібл.7, рис. 1, табл.1.

UDC 519.2:004.9

Kirichenko L., Tkachenko A., Radivilova T. **Clustering Noisy Time Series** // System technologies.- N(??).-Dnipro, 2019. –P. ??-??.

A comparative analysis of clustering noisy time series is carried out. The clustering sample contained time series of various types, among which there were atypical objects. Clustering was performed by k-means and DBSCAN methods using various distance functions for time series.

Ref.7, fig. 1, tab.1.

Ткаченко Анастасия Евгеньевна – магистр кафедры прикладной математики Харьковского национального университета радиоэлектроники

Кириченко Людмила Олеговна – д.т.н., профессор кафедры прикладной математики Харьковского национального университета радиоэлектроники.

Радивилова Тамара Анатольевна – к.т.н., доцент каф. инфокоммуникационной инженерии Харьковского национального университета радиоэлектроники