519.6

МЕТОДЫ АНАЛИЗА ПАТТЕРНОВ, ОСНОВАННЫЕ НА ИНТЕРВАЛЬНЫХ ОЦЕНКАХ

А.Л. Мячин
Национальный исследовательский университет «Высшая школа экономики»
Россия, 101001, Москва, Мясницкая ул., 20
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: amyachin@hse.ru

Ключевые слова: паттерн, интервальные оценки, анализ паттернов, порядково-интервальная паттерн-кластеризация.

Аннотация: Представлена методология поиска закономерностей в данных при использовании интервальных оценок. Описана алгоритмическая реализация и отдельные свойства порядково-интервальной паттерн-кластеризации (в т.ч. однозначность определения полученных на основе данного метода групп объектов; отсутствие пересечений в подобных группах; особенности упорядочения показателей в полученных группах). Изучена вычислительная сложность, а также возможность ее уменьшения при введении дополнительных ограничений. Приведены рекомендации по целесообразности использования порядково-интервальной паттерн-кластеризации.

1. Введение

К настоящему времени предложены множества методов поиска закономерностей в данных [9, 7], позволяющие работы со шкалами разных типов. Однако, использование интервальных оценок сопряжено с рядом сложностей, включая: ограничение в количестве методов (в сравнении с точными/округленными значениями); возможное ухудшения конечных результатов; при наличии возможности предпочтение отдается в пользу преобразования интервальных оценок в точные. С другой стороны, накопление данных большой размерности может приводить и к накоплению ошибок, связанных как с погрешностью измерений, так и с человеческим фактором. При этом, если используемая методология также предполагает определенную погрешность, конечный результат, даже при корректных вычислениях, может быть ошибочным.

В научной литературе с каждым днем появляется все больше отечественных и зарубежных исследований, посвященных методам анализа интервальных оценок (к примеру, 2 480 000 результатов по запросу «Interval Clustering Algorithm» по состоянию на 16.01.2024). Существующие методы кластеризации на основе интервальных оценок варьируются от алгоритмов, адаптированных для работы с интервальными данными, до новаторских подходов, разрабатываемых для учета диапазонов значений при группировке данных. Выделим несколько известных методов:

- Interval k-means clustering. Является расширением широко известного метода k-means [4]. Результат получается на основе минимизации суммы квадратов внутрикластерных расстояний;
- Possibilistic c-Means Clustering. Данный метод обобщает классический алгоритм c-means [3], допуская определенный уровень «гибкости» при определении принадлежности исследуемых объектов к конкретным кластерам;

- Иерархическая кластеризация данных, описанная интервальными оценками. Адаптация классических алгоритмов иерархической кластеризации [5] с использованием Ward's method или Complete Linkage.

В работе предлагается дополнительный способ поиска закономерностей в данных при использовании интервальных оценок, относящийся, согласно определению из [6], к методам анализа паттернов.

2. Порядково-интервальная паттерн-кластеризация

2.1. Описание метода

Метод, представленных ниже, относится к анализу паттернов, активно развивающийся в отечественной [1, 2] и зарубежной литературе в последние годы [8]. Под паттерном понимается «комбинация определенных качественно похожих признаков» [6]. Отличительной особенностью данного метода является возможность объединения объектов, имеющих схожую структуру, но отличающихся по абсолютным значениям выбранной базовой системы показателей. Пример работы метода изображен на рис. 1.

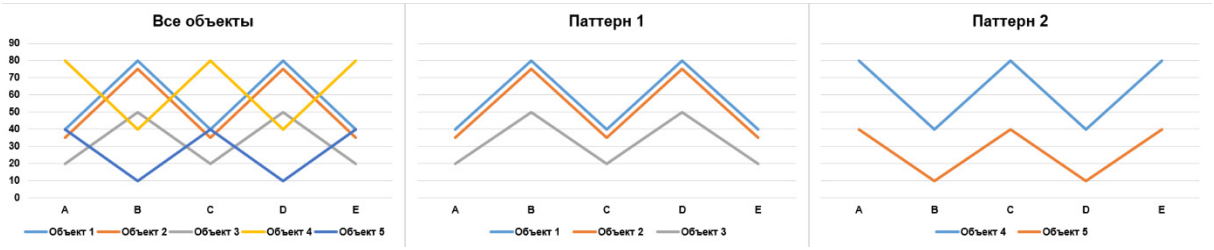

Рис. 1. Пример использования методов анализа паттернов. Слева направо: исходные данные, представленные в системе параллельных координат; паттерн 1; паттерн 2.

Приведем описание нового метода анализа паттернов с использование интервальных оценок. Имеется множество объектов $y_k \in Y$, каждому из которых поставлен в взаимно однозначное соответствие вектор $y_k = \left([y_{k1} - \xi_{k1}, y_{k1} + \xi_{k1}], [y_{k2} - \xi_{k2}, y_{k2} + \xi_{k2}], \dots, [y_{kj} - \xi_{kj}, y_{kj} + \xi_{kj}], \dots, [y_{kn} - \xi_{kn}, x_{kn} + \xi_{kn}]\right)$. Задачей является нахождение объектов, качественно похожих друг на друга. Другими словами, на основе выбранной метрики, необходимо выявление паттернов в исследуемом множестве объектов.

Возможным решением является применение обобщения правила Борда на случай интервальных оценок совместно с порядково-инвариантной паттерн-кластеризацией [6]. С этой целью, для любых $y_k \in Y$ рассчитывается значение ω_k согласно формуле

$$\omega_k = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} 10^{j-2} \phi_{kj}^s$$

где

$$\begin{cases} \phi_{kj}^s = 1, \text{если } [y_{kj} - \xi_{kj}, y_{kj} + \xi_{kj}] \prec [y_{ks} - \xi_{ks}, y_{ks} + \xi_{ks}] \\ \phi_{kj}^s = 0, \text{если } [y_{kj} - \xi_{kj}, y_{kj} + \xi_{kj}] \infty [y_{ks} - \xi_{ks}, y_{ks} + \xi_{ks}] \\ \phi_{kj}^s = 2, \text{если } [y_{kj} - \xi_{kj}, y_{kj} + \xi_{kj}] \succ [y_{ks} - \xi_{ks}, y_{ks} + \xi_{ks}] \end{cases}$$

$$s = 1, \dots, \frac{n(n-1)}{2} - 1.$$

При этом, будем считать что интервал $[y_{kj} - \xi_{kj}, y_{kj} + \xi_{kj}]$ доминирует $[y_{ks} - \xi_{ks}, y_{ks} + \xi_{ks}]$ если $y_{kj} - \xi_{kj} > y_{ks} + \xi_{ks}$.

Для выявления паттернов критерием служит нулевое расстояние Хемминга между объектами. Другими словами, $y_\alpha, y_\beta \in p_i => r(y_\alpha, y_b) = 0$, где $r(y_\alpha, y_b)$ – расстояние Хемминга между объектами y_α и y_β.

Метод поиска паттернов с использованием алгоритма, описанного выше, будем называть «порядково-интервальной паттерн-кластеризацией», а паттерны, полученные на его основе – «порядково-инвариантными паттерн-кластерами».

2.2. Основные свойства

Порядково-интервальная паттерн-кластеризация обладает некоторыми свойствами, использование которых целесообразно как для понимания конечных результатов, так и конкретных случаев, в которых потребуется использование альтернативных методов.

Свойство 1. При использовании порядково-интервальной паттерн-кластеризации принадлежность исследуемых объектов к порядково-интервальным паттерн-кластерам определяется однозначно.

Свойство 2. Для любой группы объектов, сформированных на основе порядково-интервальной паттерн-кластеризации, возможно упорядочить показатели при соблюдении условий: $[x_{ij} - \varepsilon_{ij}, x_{ij} + \varepsilon_{ij}] \prec [x_{ij+1} - \varepsilon_{ij+1}, x_{ij+1} + \varepsilon_{ij+1}] \forall\, y_{ij+1} + \varepsilon_{ij+1} < y_{ij} + \varepsilon_{ij})$; или $[x_{ij} - \varepsilon_{ij}, x_{ij} + \varepsilon_{ij}] \infty [x_{ij+1} - \varepsilon_{ij+1}, x_{ij+1} + \varepsilon_{ij+1}] \forall\, y_{ij+1} + \varepsilon_{ij+1} = y_{ij} + \varepsilon_{ij}))$; или $[x_{ij} - \varepsilon_{ij}, x_{ij} + \varepsilon_{ij}] \succ [x_{ij+1} - \varepsilon_{ij+1}, x_{ij+1} + \varepsilon_{ij+1}] \forall\, y_{ij+1} + \varepsilon_{ij+1} > y_{ij} + \varepsilon_{ij})$.

Данное свойство демонстрирует нецелесообразность использования предложенного метода при $\phi_{kj}^{s} = 0 \;\forall\, s = 1, \dots, \frac{n(n-1)}{2} - 1$. Однако, на практических задачах данная ситуация является относительно редкой.

Свойство 3. При $|Y| = m$ для алгоритмической реализации порядково-интервальной паттерн-кластеризации порядок необходимых вычислений составляет $m^3 n^3$.

Замечание 1. Вычислительная сложность порядково-интервальной паттерн-кластеризации на практике можно существенным образом снизить при помощи алгоритмов сортировки данных.

3. Заключение

Представлен новый метод, позволяющий определять паттерны данных с использованием интервальных оценок. Предложено название «порядково-интервальная паттерн-кластеризация». Алгоритмическая реализация основана на применении правила Борда на случай интервальных оценок. Описаны некоторые свойства рассматриваемого метода.

Работа выполнена при поддержки Международного центра анализа и выбора решений Национального исследовательского университета «Высшая школа экономики», а также Лаборатории теории выбора и анализа решений Института проблем управления им. В.А. Трапезникова РАН.

Список литературы

1. Алескеров Ф. Т. и др. Анализ паттернов в статике и динамике. Часть 2: Примеры применения к анализу социально-экономических процессов // Бизнес-информатика. 2013. №. 4 (26). С. 3-20.
2. Мячин А.Л. Анализ паттернов: диффузионно-инвариантная паттерн-кластеризация // Проблемы управления. 2016. №. 4. С. 2-9.

3. Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters // Journal of Cybernetics. 1973. P. 32-57.
4. Hartigan J.A., Wong M.A. Algorithm AS 136: A k-means clustering algorithm // Journal of the royal statistical society. Series C (applied statistics). 1979. Vol. 28, No. 1. P. 100-108.
5. Murtagh F., Contreras P. Algorithms for hierarchical clustering: an overview // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012. Vol. 2, No. 1. P. 86-97.
6. Myachin A.L. Pattern analysis in parallel coordinates based on pairwise comparison of parameters // Automation and Remote Control. 2019. Vol. 80. P. 112-123.
7. Qiu J., et al. A survey of machine learning for big data processing // EURASIP Journal on Advances in Signal Processing. 2016. Vol. 2016. P. 1-16.
8. Shawe-Taylor J., Cristianini N. Kernel methods for pattern analysis. Cambridge university press, 2004.
9. Xu R., Wunsch D. Survey of clustering algorithms // IEEE Transactions on Neural Networks. 2005. Vol. NN-16, No. 3. P. 645-678.