МЕТОДОЛОГИЯ ОЦЕНКИ КАЧЕСТВА РЕЗУЛЬТАТОВ АНАЛИЗА ПАТТЕРНОВ

Мячин А.Л.

Национальный исследовательский университете «Высшая школа экономики»
Институт проблем управления им. В.А. Трапезникова РАН

Введение

К настоящему времени предложены множество методов поиска закономерностей в неоднородных данных, в связи с чем возникает необходимость использования количественных метрик для сопоставления конечных результатов. Несмотря на множество существующих методов оценки качества результатов кластеризации [Halkidi, Batistakis, Vazirgiannis 2001], обобщений для методов анализа паттернов [Myachin & Mirkin 2019; Myachin 2019] относительно немного. В настоящей работе описана целесообразность использования отдельных метрик и приведена их практическая реализация.

Предлагаемые методы

В качестве исходных данных исследуется n объектов $x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im})$. Общей задачей методов анализа паттернов является выявление качественно схожих групп объектов. Другими словами, объекты каждой группы должны иметь схожую структуру (независимо от разности в абсолютных значениях показателей), а объекты разных групп – существенно отличаться между собой.

При составлении моделей прогнозирования целесообразным может быть использование порядково-инвариантной паттерн-кластеризации [Мячин 2016] из-за возможности расположения показателей одной группы объектов, полученных на основе данного метода, в монотонно возрастающем порядке. При этом, в качестве метрики оценки качества предлагается использование коэффициента

$$R^2_{c^k_{ord\_inv}} = \frac{\sum_{i \in c^k_{ord\_inv}} (\widehat{x_i} - \overline{x_{c^k_{ord\_inv}}})^2}{\sum_{i \in c^k_{ord\_inv}} (x_i - \overline{x_{c^k_{ord\_inv}}})^2},$$

где: $c^k_{ord\_inv}$ – k-ая группа объектов, полученная при использовании порядково-инвариантной паттерн-кластеризации;

$\widehat{x_i}$ – значение, прогнозируемое значение для исследуемого объекта x_i;

$x_{c^k_{ord\_inv}}$ – среднее значение, рассчитываемое для k-ой группа объектов, полученной при использовании порядково-инвариантной паттерн-кластеризации.

При использовании данной метрики в задачах прогнозирования (при построении моделей для отдельных групп объектов, полученных на основе порядково-инвариантной паттерн-кластеризации), целесообразным является максимизация значений $R^2_{c^k_{ord\_inv}}$. При этом, разумеется, возникает вопрос об объединении различных групп $c^k_{ord\_inv}$ (и о последовательности данного объединения). В общем случае, требуется рассмотрение $\frac{|c_{ord\_inv}| * (|v_{ord\_inv}| - 1)}{2}$ парных сравнений. При этом первое объединение происходит при фиксировании группы с максимальным значением $|c^k_{ord\_inv}|$ (при наличии двух подобных групп – с минимальным значением $R^2_{c^k_{ord\_inv}}$). Выбор второй группы происходит при

максимизации значения $R^2_{c^k_{ord\_inv}}$ после объединения. Отметим, что при использовании $R^2_{c^k_{ord\_inv}}$ не предполагается разбиение изначально полученных групп $c^k_{ord\_inv}$.

При использовании других методов анализа паттернов целесообразным является обобщение понятий «компактности» и «отделимости» для использования коэффициента

$$\alpha = \frac{1}{|X|}\sum_{c_s \in V}\sum_{x_i \in c_s}\frac{\min_{c^* \in V/c_s}\{\frac{1}{|c^*|}\sum_{x_k \in c_s}||x_i - x_k||\} - \frac{1}{|c_s|}\sum_{x_k \in c_s}||x_i - x_k||}{\max\{\frac{1}{|c_s|}\sum_{x_k \in c_s}||x_i - x_k||, \min_{v^* \in V/c_s}\{\frac{1}{|c^*|}\sum_{x_k \in c_s}||x_i - x_k||\}\}},$$

где: c_s – группа объектов (паттерн) s;

$\frac{1}{|c^*|}\sum_{x_k \in c_s}||x_i - x_k||$ – среднее расстояние между объектами одного полученного паттерна и остальных ($c^* \neq c_s$).

Приведенный коэффициент описывает качество полученных результатов, однако не рекомендуется применять с методами анализа паттернов, использующими расстояние Хемминга (в связи с особенностями объединения объектов в группы).

Заключение

Представлены два метода оценки качества результатов анализа паттернов. Кратко описана методология применения. Отдельно отметим, что, согласно [Dubes & Jain 1976], не существует оптимального метода разбиения объектов на группы, и важным критерием оценки качества служит интерпретируемость конечных результатов. Также, согласно теореме о невозможности Клейнберга [Kleinberg 2002], невозможно создание универсального метода объединения исследуемых объектов (при соблюдении некоторых базовых условий). Однако, при работе с большими объемами данных весьма полезно введение отдельных метрик, позволяющих производить предварительные сравнения результатов различных разбиений объектов на группы. Подобные метрики могут существенным образом ускорять процесс выбора оптимального метода на практических задачах.

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.

Литература

[Алескеров и др. 2013] 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.

[Мячин 2016] Мячин А. Л. Анализ паттернов: порядково–инвариантная паттерн–кластеризация //Управление большими системами: сборник трудов. – 2016. – №. 61. – С. 41-59.

[Dubes & Jain 1976] Dubes R., Jain A. K. Clustering techniques: the user's dilemma //Pattern Recognition. – 1976. – Т. 8. – №. 4. – С. 247-260

[Halkidi, Batistakis, Vazirgiannis 2001] Halkidi M., Batistakis Y., Vazirgiannis M. On clustering validation techniques //Journal of intelligent information systems. – 2001. – Т. 17. – С. 107-145.

[Kleinberg 2002] Kleinberg J. An impossibility theorem for clustering //Advances in neural information processing systems. – 2002. – Т. 15.

[Myachin & Mirkin 2019] Myachin A., Mirkin B. Ordinal Equivalence Classes for Parallel Coordinates //Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20. – Springer International Publishing, 2019. – С. 525-533.