ПАТТЕРНЫ ДАННЫХ В СИСТЕМЕ ПАРАЛЛЕЛЬНЫХ КООРДИНАТ[97]

Мячин А.Л.

Национальный исследовательский университет «Высшая школа экономики»,
Россия, г. Москва, ул. Мясницкая, д.20,
Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная д.65
amyachin@hse.ru

Аннотация: Показана возможность использования новых методов анализа паттернов для поиска закономерностей в данных. Описаны некоторых свойства методов анализа паттернов, основанных на алгоритмах сортировки. Рассмотрена методология корректировки конечных результатов при наличии погрешностей в данных, характеризующих исследуемую выборку объектов.

Ключевые слова: паттерн, анализ паттернов, кластерный анализ.

Введение

С накоплением большого количества разнородных данных весьма актуальным становится совершенствование существующих и создание новых алгоритмов, позволяющих выявлять закономерности в больших массивах информации. Среди множества существующих методов можно выделить три важных подхода: классификация, кластеризация и анализ паттернов. При отсутствии возможности установления конечного количества групп, на которое должны быть разбиты исследуемые объекты, а также при сомнениях в определении типичных представителей каждой группы, целесообразно не использовать методы классификации данных, а переходить к кластеризации и анализе паттернов.

Методам кластеризации посвящено множество обзоров, к примеру [8]. К наиболее известным методам можно отнести k-means [7] и его модификации (к примеру, k-means++), DBSCAN, иерархическая кластеризация, Mean Shift и др.

Анализ паттернов также является весьма популярным методам анализа данных, успешно примененный в различных задачах, к примеру, при анализе государственной состоятельности стран мира [2], в банковской сфере [1], политологии [6] и др.

1 Анализ паттернов и кластерный анализ: краткое сопоставление методов

Как и описано во введение, довольно часто в самых различных областях имеется необходимость поиска закономерностей в данных и объединение схожих объектов в группы. Если мы не знаем

[97] *Работа подготовлена в результате проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) и с использованием средств субсидии в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".*

конечное количество подобных групп, а также ее типичных представителей, целесообразным представляется использование методов кластеризации или анализа паттернов для получения некоторого разбиения исходного множества объектов.

Методам кластерного анализа посвящено множество обзоров, подробно описывающих различные методы выявления кластеров в исходном наборе данных. Само понятие «кластер» в литературе хорошо определено. Из [3]: «Под кластером обычно понимается часть данных (в типичном случае – подмножество объектов или подмножество переменных, или подмножество объектов, характеризуемых подмножеством переменных), которая выделяется из остальной части наличием некоторой однородности ее элементов».

Одним из современных методов анализа паттернов, продемонстрировавший свою эффективность при решении множества прикладных задач, а также протестированный на классических тестовых данных (таких, как Iris Data Set, Wine Data Set, Balance Scale Data Set и др.) является порядково-инвариантная паттерн-кластеризация [4]. Данный метод основан на парном сравнении показателей и не зависит от выбора последовательности входных данных. Основные свойства порядково-инвариантной паттерн-кластеризации подробно представлены в [4], одно из которых описывается следующим утверждением: «Для любых объектов, объединенных в единый кластер на основе порядково-инвариантной паттерн-кластеризации, существует порядок расположения показателей, образующий для данных объектов монотонную неубывающую/невозрастающую последовательность».

Приведенное выше утверждение открывает возможность для существенного уменьшения вычислительной сложности порядково-инвариантной паттерн-кластеризации, позволяя использовать классические методы сортировки данных для получения порядково-инвариантных паттерн-кластеров (групп объектов, полученных с использованием данного метода).

Основное отличие методов анализа паттернов от методов кластерного анализа является способ объединения объектов: в первом случае рассматривается структура объектов по выбранным показателям, не делая упор на их абсолютные значения. Продемонстрируем данное свойство на следующем примере.

Таблица 1. Гипотетический пример для общего сопоставления методов анализа паттернов и кластерного анализа

Объект	A	B	C
X1	20	60	20
X2	18	62	18
X3	30	10	30
X4	28	8	28
X5	2	6	2
X6	1,8	6,2	1,8

Для наглядности, приведем два различных способы визуального представления данных из табл. 1. При использовании методов анализа паттернов, как правило, используется система параллельных координат [9].

Как видно на рис. 1, хорошо выделяются три группы объектов: {X1, X2}, {X3, X4}, {X5, X6}. Во многих практических задачах подобное разбиение и потребуется, и, с использованием множества методов кластерного анализа возможно получение данного результата.

Однако, если посмотреть на объекты X1 и X4, можно заметить, что их структура весьма схожа: значения показателей объекта X1 есть значения показателей X4, помноженные на 10. Данное наблюдение верно и для объектов X2 и X6. Таким образом, несмотря на существенные различия в абсолютных значениях показателей, приведенные выше объекты имеют схожую структуру данных, и задачей анализа паттернов, в данном случае, является разбиение на две группы: {X1, X2, X5, X6}, {X3, X4}.

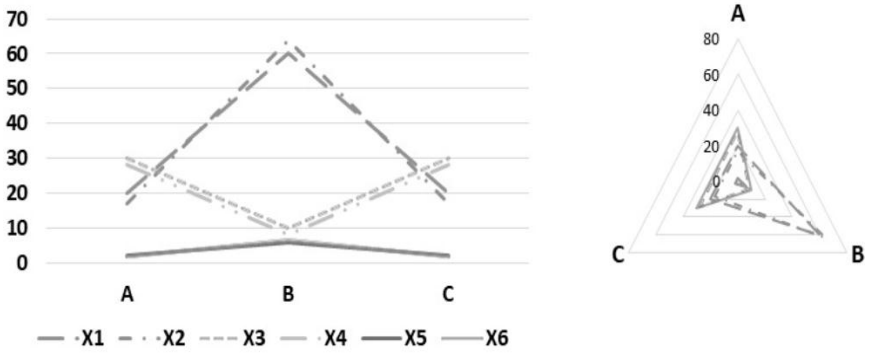

Рис. 1. Гипотетический пример: визуализация данных (слева – представление в 3-мерной системе параллельных координат; справа – многоугольники оценки).

2 Корректировка результатов с использованием центроидов

Очевидно, что методы анализа паттернов, основанные на парном сравнении показателей (и не только данные методы), могут быть чувствительны к наличию погрешностей в исходных данных. Их округление до определенной точности не всегда способно привести к ожидаемым результатам. Также, весьма частой проблемой при использовании различных методов кластеризации/анализа паттернов является вопрос выявления и интерпретации небольших групп объектов. В связи с этим, целесообразным является разработка методов, позволяющих корректировать конечные результаты разбиения исходного множества объектов при предположении о наличии погрешностей в исходных данных. Одним из возможных решений является корректировка результатов с использованием центроидов порядково-инвариантных паттерн-кластеров, подробно описанный в [5]. Приведем краткое описание.

В [5] доказано, что объект x_N (центроид), образованный как:

$$x_N = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \quad (1)$$

принадлежит к той же группе, что и объекты, на основе значений которых он образован. В данном случае, n – общее количество объектов, принадлежащих соответствующей группе. Очевидно, что значения показателей объекта x_N будут средними значениями показателей по группе, т.е. $x_N = (1/n\sum_{i=1}^{n} x_{i1}, 1/n\sum_{i=1}^{n} x_{i2},..., 1/n\sum_{i=1}^{n} x_{ik})$. Для каждого исследуемого объекта рассчитывается расстояние до центроида каждой полученной в результате использования порядково-инвариантной паттерн-кластеризации группы с использованием формулы

$$d(x_i, v_{inv}) = \sqrt{\sum_{i=1}^{k} \left(x_{ij} - x_{Nj}\right)^2} \,, \quad (2)$$

где v_{inv} – соответствующая группа объектов, полученная с использованием порядково-инвариантной паттерн-кластеризации.

В результате, для каждого объекта будет рассчитано z (согласно количеству полученных при использовании порядково-инвариантной паттерн-кластеризации групп) расстояний $d(x_i, v_{inv})$. Корректировка результатов происходит с использованием формулы

$$w_i = \min(d(x_i, v_{inv\_1}), d(x_i, v_{inv\_2}),..., d(x_i, v_{inv\_z})). \quad (3)$$

Объект x_i относим к группе v_{inv} при $w_i = d(x_i, v_{inv})$. Таким образом, минимизируется расстояние от каждого объекта до центроида группы, к которой данный объект принадлежит.

Литература

1. Алескеров Ф. Т., Солодков В. М., Челнокова Д. С. Динамический анализ паттернов поведения коммерческих банков России // Экономический журнал Высшей школы экономики. – 2006. – Т. 10. – №. 1.

*2. Ахременко А.*С., Мячин А.Л. Паттерн-анализ и кластеризация в исследовании государственной состоятельности: «адаптивная оптика» к политической науке // Политическая наука. – 2019 (в печати).

3. Миркин Б. Г. Методы кластер-анализа для поддержки принятия решений: обзор. – 2011.

3. Мячин А. Л. Анализ паттернов в системе параллельных координат на базе парного сравнения показателей //Автоматика и телемеханика. – 2019. – №. 1. – С. 138-152.

4. Мячин А.Л. Определение центроидов для повышения точности порядково-инвариантной паттерн-кластеризации // Управление большими системами: сборник трудов. – 2019. – №78. – С.6-22.
5. Aleskerov F., Nurmi H. A method for finding patterns of party support and electoral change: An analysis of British general and Finnish municipal elections // Mathematical and Computer Modelling. – 2008. – Т. 48. – №. 9-10. – С. 1385-1395.
6. 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. – Т. 28. – №. 1. – С. 100-108.
7. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review // ACM computing surveys (CSUR). – 1999. – Т. 31. – №. 3. – С. 264-323.
8. Inselberg A. The plane with parallel coordinates // The visual computer. – 1985. – Т. 1. – №. 2. – С. 69-91.