

© 2019 г. А.Л. МЯЧИН, канд. техн. наук (amyachin@hse.ru)

(Национальный исследовательский университет

«Высшая школа экономики», Москва,

Институт проблем управления им. В.А. Трапезникова РАН, Москва)

АНАЛИЗ ПАТТЕРНОВ В СИСТЕМЕ ПАРАЛЛЕЛЬНЫХ КООРДИНАТ НА БАЗЕ ПАРНОГО СРАВНЕНИЯ ПОКАЗАТЕЛЕЙ¹

Представлены основные свойства нового метода анализа паттернов в системе параллельных координат, результат которого не зависит от последовательности данных в исходной выборке анализируемых объектов. Доказано утверждение о том, что кластеры, полученные с использованием данного метода, не пересекаются. Показана возможность представления объектов одного кластера в виде монотонно возрастающих/убывающих функций.

Ключевые слова: анализ паттернов, порядково-инвариантная паттерн-кластеризация, кластерный анализ.

DOI: 10.1134/S0005231019010100

1. Введение

С развитием информационных технологий и накоплением больших объемов данных все более значимым становится создание методов, позволяющих автоматизировать процесс распознавания и выделения различных групп, обладающих схожими свойствами. К первым подобным работам можно отнести [1], предлагающую разбиение множества трех видов Ирисов по четырем признакам с использованием линейного дискриминантного алгоритма. Описанная в [1] база данных 150 Ирисов до сих пор является одной из наиболее популярных в машинном обучении, а предложенная Р. Фишером методология подробно рассматривается в ряде учебных пособий (например, [2, 3]). Дальнейшее развитие данное направление получило в алгоритмах кластерного анализа, основанных на использовании различных мер близости объектов [4, 5]. Однако потребности исследования все более сложных процессов (экономических, финансовых, социальных и иных) привели к необходимости учета не только близости значений, но и близости структуры самих данных, что, в свою очередь, привело к развитию методов их визуального представления и анализа (в общем русле которых находится данная работа).

Из более современных исследований, учитывающих отмеченные тенденции, укажем, в частности, на прикладные работы в банковской сфере [6, 7] (анализ показателей CAMEL [8–10]), менеджменте [11], макроэкономике [12].

¹ Статья 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». Автор выражает благодарность д-ру техн. наук, проф. Алескерову Ф.Т. за помощь в написании данной статьи.

В основе данного направления лежит понятие «паттерн», которое в различных областях знаний определяется по-разному. Основные определения приведены в [13]:

- 1) «как сущность явления, имеющего повторяющиеся черты»;
- 2) «как свойство повторяющихся компонентов, объединенных общей структурой»;
- 3) «как процесс, фиксирующий модель взаимодействия изучаемых объектов, включающего повторения».

В [2] под паттерном понимается «любые отношения, закономерности или структура, присущая некоторому набору данных», а под анализом паттернов — «процесс нахождения общих соотношений в наборе данных». Опираясь на предложенные формулировки, под паттерном, в данной работе будем понимать комбинацию определенных, качественно похожих признаков. В [2] отмечается, что для признания метода анализа паттернов эффективным, реализующий его алгоритм должен удовлетворять следующим условиям: возможность обработки больших объемов данных (что подразумевает относительно невысокую вычислительную сложность), робастность и статистическую устойчивость. В данной работе будем учитывать еще один критерий, затронутый в [14]: независимость конечных результатов от выбора исходной последовательности показателей. В связи с этим рассматриваются предложенные в [15] методы анализа паттернов: порядково-фиксированная и порядково-инвариантная паттерн-кластеризации. В работе обобщаются реализующие их алгоритмы, предлагаются методы распознавания паттернов, объединения схожих объектов в группы, а также исследуются ряд свойств данных групп, в том числе описываются удобные способы их визуализации и определения «средних/центральных» объектов кластера. Таким образом, цель работы — обобщение и структурирование разработанных алгоритмов анализа паттернов с не зависимым от последовательности входных данных конечным результатом, а также исследование основных свойств как самих алгоритмов, так и полученных на их основе групп объектов. Для удобства и цельности изложения, доказательства всех сформулированных в работе предложений вынесены в Приложение.

2. Анализ паттернов: основные понятия

Исследуется некоторое множество, состоящее из m объектов, по n показателям. Элементы множества будем обозначать через x_i , набор показателей конкретного объекта — $x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{in})$, где x_{ij} — j -й показатель i -го объекта. Основной задачей являются выделение и объединение качественно схожих объектов. Для визуализации используется система параллельных координат [14, 16, 17], которая состоит, как правило, из вертикальных линий (осей), отражающих значения показателей. На данных линиях отмечаются их фактические значения, после чего значения соединяются отрезками. В результате образуются кусочно-линейные функции, характеризующие анализируемые объекты.

Поясним идею метода объединения объектов в отдельные группы (кластеры) по виду кусочно-линейных функций на гипотетическом примере.

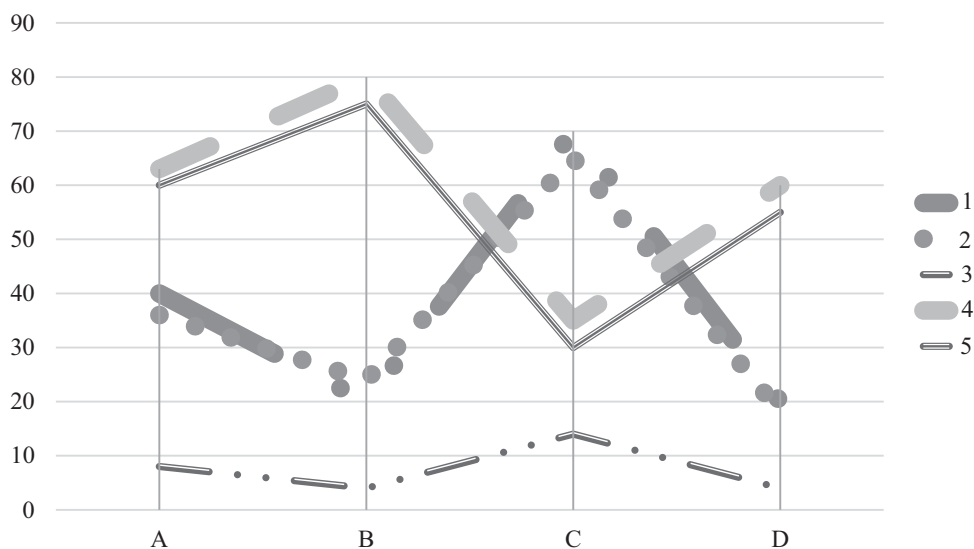


Рис. 1. Пример гипотетических объектов.

Пример. Исследуются пять объектов по четырем показателям (А, В, С и D), значения которых приведены в табл. 1.

Визуально объекты 1–5 представлены на рис. 1.

Из рисунка наглядно видно, что объекты 1 и 2 имеют схожие структуры. Данное утверждение верно и для объектов 4 и 5. Объект 3 рассмотрим отдельно. Несмотря на тот факт, что абсолютные значения показателей данного объекта сильно отличаются от объектов 1 и 2, структуры их схожи: показатели объекта 1 есть показатели объекта 3, умноженные на 5. Таким образом, все три объекта описываются кусочно-линейными функциями единой структуры (единым паттерном), что дает основание для их объединения по данному критерию.

Использование такого подхода оказалось удобным для решения широкого круга прикладных задач. В частности, в [6] проведен динамический анализ паттернов показателей 1018 банков России за 1999–2003 гг. Исходные данные базируются на фундаментальных характеристиках банков — показателях CAMEL (С – достаточность капитала, А – качество активов, М – менеджмент, Е – прибыль, L – ликвидность) с учетом некоторых дополнений.

Таблица 1. Пример с гипотетическими объектами

Объекты	А	В	С	Д
1	40	20	70	20
2	36	24	65	18
3	8	4	14	4
4	63	80	35	60
5	60	75	30	55

На базе квартальных данных за указанный период построено 19342 кусочно-линейные функции, что позволило выделить 151 типовых паттернов, причем первые 50 содержали 90,14% всех данных, а первые 13–52,19%. В [7] применен аналогичный подход для исследования 55 банков Турции. Базовая система показателей также основывалась на CAMEL, а результатом являлось формирование 27 паттернов.

Использование метода анализа паттернов в системе параллельных координат демонстрирует его эффективность для различных приложений [13, 16], однако сам метод критичен к последовательности параметров. Ниже представлены алгоритмы его реализации, основанные на парном сравнении показателей порядково-фиксированной и порядково-инвариантной паттерн-кластеризаций, позволившие устранить указанный недостаток.

3. Порядково-фиксированная и порядково-инвариантная паттерн-кластеризация

Приведем краткое описание порядково-фиксированной и порядково-инвариантной паттерн-кластеризаций, реализующих объединение объектов в группы/кластеры по виду их кусочно-линейных функций в системе параллельных координат (полное описание методов приведено в [15, 18]).

Как отмечено выше, исходные данные представлены множеством X , состоящим из m объектов, каждый из которых характеризуется n показателями. Образует кодировки c_i каждого объекта $x_i \in X$ путем парного сравнения его показателей согласно формуле

$$(1) \quad c_i = \sum_{s=1}^n 10^{s-1} z_i^{n-s},$$

где z_i^s определяется как

$$(2) \quad z_i^s = \begin{cases} 1, & \text{если } x_{is} < x_{is+1}, \\ 0, & \text{если } x_{is} = x_{is+1}, \\ 2, & \text{если } x_{is} > x_{is+1}. \end{cases}$$

Если $c_i = c_k$, объекты x_i и x_k объединяются в единый кластер, если нет — разделяются. Отметим, что значение z_i^s определяется выражением 2 путем сравнения рядом стоящих показателей x_{is} и x_{is+1} в заданной их последовательности. Поэтому в дальнейшем данный метод будем называть порядково-фиксированной паттерн-кластеризацией, а кластеры, полученные на основе данного метода, соответственно порядково-фиксированными паттерн-кластерами.

Вычислительная сложность N_{fix} данного метода можно выразить формулой

$$N_{fix} = \frac{m^2(m-1)(n-1)}{2}.$$

Поскольку N_{fix} является относительно невысокой, данный метод удобно применять для предварительного анализа данных.

Второй метод — порядково-инвариантная паттерн-кластеризация — разработан для получения не зависящих от исходной последовательности показателей результатов. Потребность в создании такого метода обусловлена, как отмечено выше, комментариями некоторых работ по параллельным координатам (в том числе [14]), указывающих на необходимость «крайне осторожного выбора последовательности анализируемых показателей», поскольку она влияет на вид формируемых паттернов и в общем случае другая последовательность может приводить к альтернативному характеру кусочно-линейных функций и, как следствие, к альтернативным результатам.

Алгоритм порядково-инвариантной паттерн-кластеризация аналогичен порядково-фиксированной и основан на сравнении кодировок объектов. Сами объекты в этом случае представлены в виде полных взвешенных ориентированных графов, вершины которых соответствуют анализируемым показателям, а значения соединяющих их ребер — результатам парных сравнений (предполагается сравнения вида «больше», «равно» и «меньше»). С использованием значений ребер, формируется дополнительная кодировка объекта:

$$(3) \quad c_i^{dop} = \sum_{s=1}^{n-2} \sum_{j=s+2}^n 10^{j-(s+2)} e_i^{sj},$$

где: e_i^{sj} — значение ребра графа, соединяющего s -ю и j -ю вершины, которое аналогично (2) определяется по формуле

$$(4) \quad e_i^{sj} = \begin{cases} 1, & \text{если } x_{is} < x_{is+1}, \\ 0, & \text{если } x_{is} = x_{is+1}, \\ 2, & \text{если } x_{is} > x_{is+1}. \end{cases}$$

Принцип формирования групп аналогичен порядково-фиксированной паттерн-кластеризации: если $c_i^{dop} = c_k^{dop}$, объекты x_i и x_k объединяются в единый кластер, в противном случае разбиваются по разным.

В дальнейшем кластер, сформированный на основе порядково-инвариантной паттерн-кластеризации, будем называть порядково-инвариантным паттерн-кластером.

Вычислительная сложность N_{inv} данного метода определяется как

$$N_{inv} = \frac{m^2 n (m-1)(n-1)}{4},$$

однако с использованием доказанного ниже утверждения 2 может быть снижена до сложности алгоритма сортировки.

Далее, определим максимальное число порядково-фиксированных и порядково-инвариантных паттерн-кластеров.

В рассматриваемых методах критерием объединения в единый кластер служит совпадение кодов, формируемых при парных сравнениях соответ-

ствующих показателей объекта. Используя известное выражение теории кодирования (при использовании алфавита кодирования, состоящего из β различных символов и длине кодовой последовательности γ , максимальное число V различных кодовых комбинаций равно β^γ), запишем:

$$V = R^N,$$

где R — число возможных значений, характеризующих результат парных сравнений соответствующих показателей объекта (определяемых формулами (2) и (4)).

Поскольку для порядково-фиксированной паттерн-кластеризации код формируется путем $n-1$ парных сравнений, число возможных порядково-фиксированных паттерн-кластеров V_{fix} определяется как

$$V_{fix} = R^{(n-1)} = 3^{(n-1)},$$

а число кластеров, формируемых в результате порядково-инвариантной паттерн-кластеризации, V_{inv} —

$$V_{inv} = R^{\frac{n(n-1)}{2}} = 3^{\frac{n(n-1)}{2}}.$$

Взяв отношение этих величин, получим оценку соотношения между максимально-возможными числами порядково-инвариантных и порядково-фиксированных паттерн-кластеров:

$$\frac{V_{inv}}{V_{fix}} = \frac{R^{\frac{n(n-1)}{2}}}{R^{(n-1)}} = 3^{\frac{n(n-1)}{2}}.$$

Замечание 1. Полученное выражение определяет максимально возможное число подкластеров, которое можно выделить в рамках порядково-фиксированного паттерн-кластера. Реальное их количество может оказаться существенно меньше. В частности, если порядково-фиксированный паттерн-кластер удовлетворяет условию утверждения 2 (см. раздел 4), то он сам является порядково-инвариантным паттерн-кластером и в силу утверждения 1 (см. ниже), никаких иных подкластеров не содержит.

4. Порядково-инвариантная паттерн-кластеризация: основные свойства

Рассмотрим три утверждения, демонстрирующие важные свойства описанного выше метода.

Утверждение 1. Кластеры, полученные с использованием порядково-инвариантной паттерн-кластеризации, не пересекаются.

Доказательство приведено в Приложении.

Можно привести следующее замечание.

Замечание 2. Утверждение 1 весьма важно, поскольку декларирует однозначность результатов рассматриваемого метода. Это означает, что произвольное расположение объектов в исходном множестве, равно как и произвольный порядок их показателей, а также использование порядково-инвариантной паттерн-кластеризации произвольное число раз на одних и тех же данных не влияет на результат кластеризации.

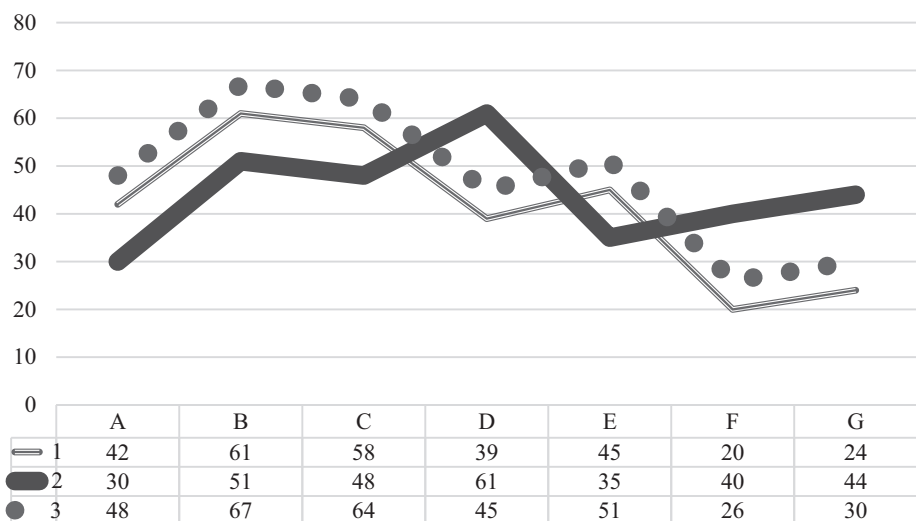


Рис. 2. Гипотетический пример двух порядково-инвариантных паттерн-кластеров.

Отметим, что если порядок показателей строго фиксирован и не меняется, то утверждение 1 верно и для порядково-фиксированных паттерн-кластеров.

Утверждение 2. Если существует последовательность расположения показателей, при которой их значения образуют строго монотонно возрастающую/убывающую последовательность для каждого объекта исходного множества X , то это множество представляет собой порядково-инвариантный паттерн-кластер.

Доказательство приведено в Приложении.

Можно привести замечание 3.

Замечание 3. Отмеченное свойство порядково-инвариантных паттерн-кластеров, определяемое утверждением 2, удобно использовать для предварительного сравнения различных групп объектов и визуального восприятия их отличительных особенностей. В качестве примера, рассмотрим множество из трех объектов, показатели которых и соответствующие кусочно-линейные функции приведены на рис. 2. Использование описанных выше методов позволяет выделить два порядково-инвариантных паттерн-кластера: Объект 1; Объект 3, Объект 2.

Расположим показатели таким образом, чтобы их значения образовали монотонно возрастающую последовательность для первого и третьего объектов, образующих порядково-инвариантный паттерн-кластер (см. рис. 3). Можно видеть, что в этом случае более четко выражен характер их различия с объектом 2.

Утверждение 3. Для объектов порядково-инвариантного паттерн-кластера, существует порядок расположения показателей, при котором их значения образуют монотонную неубывающую/невозрастающую последовательность для каждого объекта кластера.

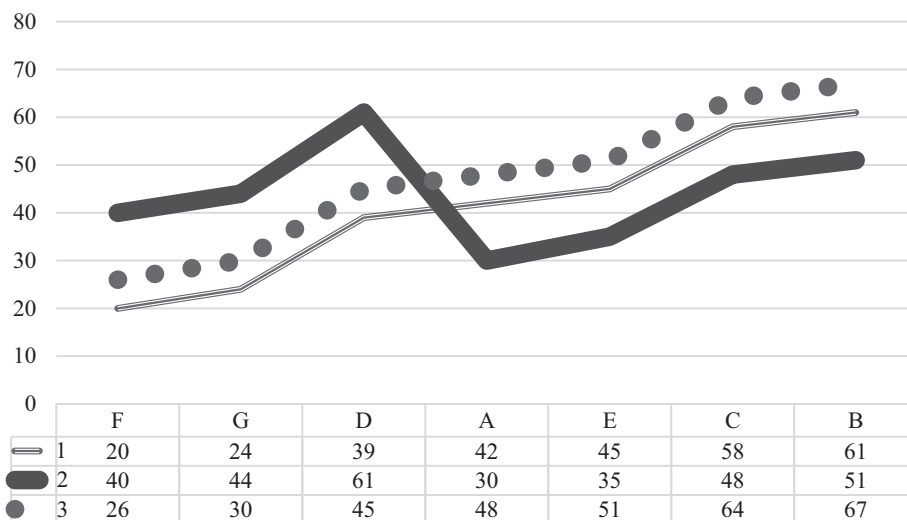


Рис. 3. Гипотетический пример двух порядково-инвариантных паттерн-кластеров.

Доказательство приведено в Приложении.

Можно привести замечание 4.

Замечание 4. Важным следствием утверждений 2 и 3 является возможность их использования для снижения вычислительной сложности порядково-инвариантной паттерн-кластеризации до сложности алгоритма сортировки.

5. Использование порядково-инвариантной паттерн-кластеризации при исследовании экономических, инновационных и образовательных показателей в РФ

Для иллюстрации предложенных методов, а также некоторых основных свойств рассмотрим пример реальных данных из [19], где исследуются показатели науки, образования и инновационной активности регионов РФ за 2007–2010 гг. На базе российского регионального инновационного индекса [20] были сформированы 6 блоков показателей: социально-экономические условия (А), образовательный потенциал (В), потенциал научно-технической деятельности (Х), результативность исследований и разработок (С), потенциал инновационной деятельности (D) и результативность инновационной деятельности (Е). В результате корреляционного анализа блок, научно-технический потенциал был исключен, и регионы РФ подразделялись на основании блоков А–Е.

Для начала воспользуемся порядково-инвариантной паттерн-кластеризацией для разбиения регионов на кластеры. Результатом являются 22 кластера, содержащих более 5 объектов, 12 кластеров — от 3 до 5 и 1 кластер, в который вошли «уникальные объекты». Примеры полученных кластеров представлены на рис. 4.

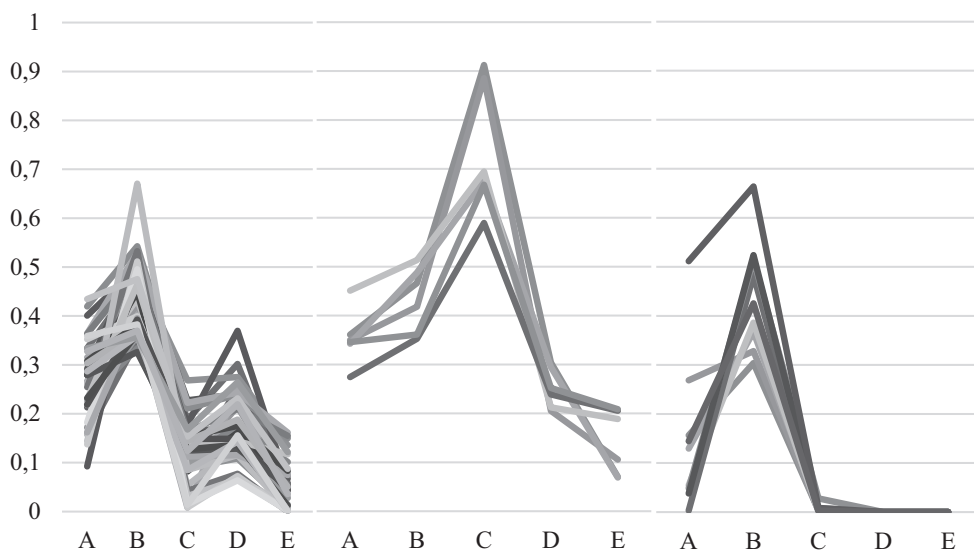


Рис. 4. Примеры кластеров, полученных при использовании порядково-инвариантной паттерн-кластеризации.

Далее, возьмем один из полученных в работе [20] кластеров и проверим, можно ли его отнести к порядково-инвариантным паттерн-кластерам (данные приведены в табл. 2, значения округлены до второго знака после запятой).

Каждому региону табл. 2 соответствует единая кодировка, полученная при помощи формул (1) и (3): «1122» и «222221» (дополнительная кодировка). Следовательно, полученный в [19] паттерн является порядково-инвариантным паттерн-кластером. Для данных регионов построим кусочно-линейные функции, см. рис. 5.

Теперь проиллюстрируем справедливость утверждения 3. Выделим первую строку табл. 2: (0,36; 0,47; 0,91; 0,3; 0,07). Данная строка описывает выбранные показатели Москвы в 2009 г. Расположим значения показателей в порядке возрастания (Е, D, А, В, С). Такому порядку расположения показателей соответствует паттерн, имеющий форму кусочно-линейной неубывающей функции. При данном расположении показателей согласно утверждению 3 кусочно-линейные функции других объектов также будут иметь вид неубывающей функции, что продемонстрировано на рис. 6. Несложно проверить

Таблица 2. Пример с гипотетическими объектами

Регион/года	А	В	С	Д	Е
Москва 2009	0,36	0,47	0,91	0,3	0,07
Москва 2010	0,35	0,42	0,89	0,21	0,11
Санкт-Петербург 2007	0,34	0,49	0,68	0,3	0,07
Санкт-Петербург 2009	0,45	0,51	0,7	0,21	0,19
Приморский край 2008	0,27	0,35	0,6	0,24	0,21
Приморский край 2009	0,35	0,36	0,67	0,25	0,21

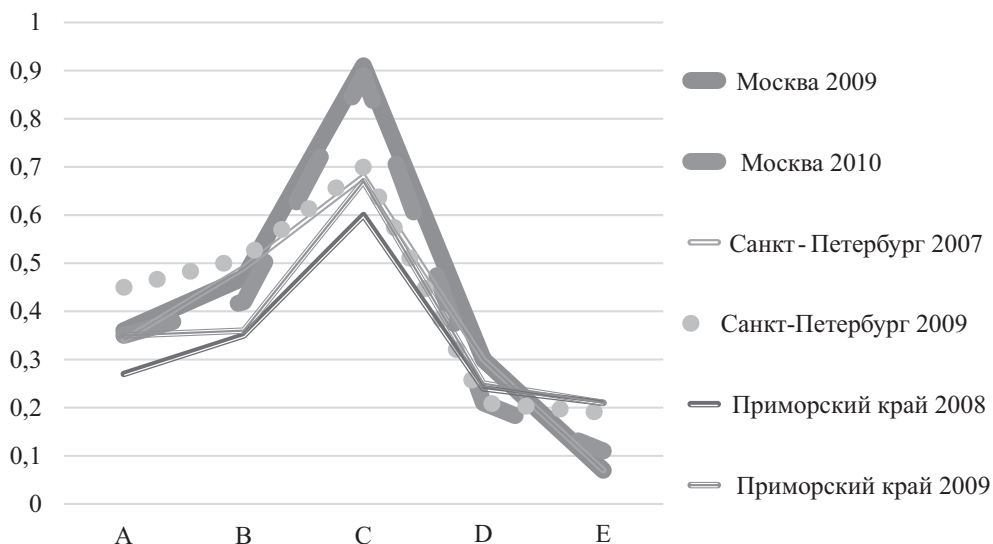


Рис. 5. Кусочно-линейные функции объектов, соответствующие порядку показателей A, B, C, D, E.

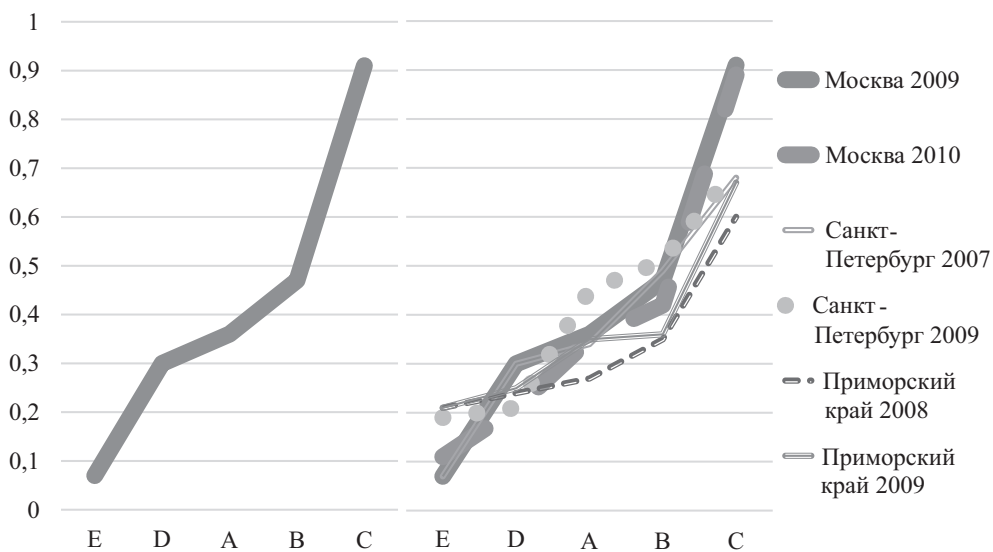


Рис. 6. Кусочно-линейные функции объектов, соответствующие порядку показателей A, B, C, D, E.

на данном примере и справедливость утверждения 2: для любого из трех рассматриваемых регионов (за два года каждый) существует последовательность расположения показателей (E, D, A, B, C), при которой формируется строго возрастающая последовательность, и, следовательно, рассматриваемые регионы формируют единый порядково-инвариантный паттерн-кластер.

6. Операции над объектами кластеров и их свойства

Рассмотрим некоторые математические операции над объектами порядково-инвариантных паттерн-кластеров.

1. Операция суммирования. Под суммой двух объектов $x_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1n})$ и $x_2 = (x_{21}, \dots, x_{2j}, \dots, x_{2n})$ будем понимать новый объект $x_s = x_1 + x_2$, значения показателей которого определяются как сумма соответствующих показателей объектов x_1 и x_2 : $x_s = (x_{11} + x_{21}, \dots, x_{1j} + x_{2j}, \dots, x_{1n} + x_{2n})$.

2. Произведение объекта на число. Под произведением объекта $x_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1n})$ порядково-инвариантного паттерн-кластера на действительное число α будем понимать новый объект $x_\alpha = \alpha x_1$, значения показателей которого определяются как произведение показателей исходного объектов x_1 на число α : $x_\alpha = (\alpha x_{11}, \dots, \alpha x_{1j}, \dots, \alpha x_{1n})$.

Утверждение 4. Если два объекта $x_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1n})$ и $x_2 = (x_{21}, \dots, x_{2j}, \dots, x_{2n})$ принадлежат одному порядково-инвариантному паттерн-кластеру, то и их сумма $x_s = x_1 + x_2 = (x_{11} + x_{21}, \dots, x_{1j} + x_{2j}, \dots, x_{1n} + x_{2n})$ также принадлежит этому кластеру.

Доказательство приведено в Приложении.

Утверждение 5. Если объект $x_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1n})$ принадлежит некоторому порядково-инвариантному паттерн-кластеру v_a^{inv} , то для любых положительных значений α ($\alpha > 0$) объект $x_\alpha = \alpha x_1$ также принадлежит данному кластеру.

Доказательство данного утверждения аналогично доказательству утверждения 4.

Сформулируем важные следствия утверждений 4 и 5.

Следствие 1. Если объекты $x_1, \dots, x_i, \dots, x_n$ принадлежат одному порядково-инвариантному паттерн-кластеру, то и «средний» («центральный») объект x_{st} вида

$$x_{st} = \frac{1}{n} \sum_{i=1}^n x_i$$

также принадлежит данному кластеру.

Следствие 2. Если объекты $x_1, \dots, x_i, \dots, x_n$ принадлежат одному порядково-инвариантному паттерн-кластеру, то и их линейная комбинация x_{lc} вида

$$x_{lc} = \sum_{i=1}^n \lambda_i x_i,$$

где $\lambda_i > 0 \mid i = 1, \dots, n$, также принадлежит данному кластеру.

7. Заключение

В работе приведено описание двух разновидностей перспективного метода анализа данных — анализа паттернов и описаны некоторые их свойства. Приведена вычислительная сложность соответствующих методов. Сформулированы и доказаны пять утверждений и два следствия, в том числе о непересечении порядково-инвариантных паттерн-кластеров, о представлении объектов данных кластеров в виде монотонно возрастающих/убывающих функций, а также о нахождении «среднего» («центрального») объекта кластера.

ПРИЛОЖЕНИЕ

Доказательство утверждения 1. Предположим, что существуют два различных (несовпадающих) кластера $v_a^{inv} \neq v_b^{inv}$, полученных в результате порядково-инвариантной паттерн-кластеризации некоторого множества объектов X . Согласно утверждению 1 $v_a^{inv} \cap v_b^{inv} = \emptyset$.

Доказательство проведем от противного. Допустим, что данное утверждение не выполняется и $v_a^{inv} \cap v_b^{inv} \neq \emptyset$. Это означает, что существует хотя бы один объект $x_i^* \in X : x_i^* \in v_a^{inv} \cap v_b^{inv}$.

Поскольку $v_a^{inv} \neq v_b^{inv}$, то существует и некоторая последовательность исходных показателей, при которой кусочно-линейные функции данных кластеров имеют различный вид. Данную последовательность обозначим через Y .

Согласно определению (алгоритму построения) порядково-инвариантных паттерн-кластеров вид кусочно-линейных функций всех объектов кластера v_a^{inv} должен совпадать при любой последовательности исходных показателей (включая последовательность Y). Поскольку $x_i^* \in v_a^{inv}$, то вид кусочно-линейных функций всех объектов кластера v_a^{inv} аналогичен виду кусочно-линейной функции объекта x_i^* . Повторив данное рассуждение для кластера v_b^{inv} , приходим к заключению, что вид кусочно-линейных функций всех объектов кластера v_b^{inv} также аналогичен виду кусочно-линейной функции объекта x_i^* . Таким образом, приходим к заключению, что вид кусочно-линейных функций всех объектов кластеров v_a^{inv} и v_b^{inv} совпадает (аналогичен виду кусочно-линейной функции объекта x_i^*). Однако в этом случае согласно определению (алгоритму построения) порядково-инвариантной паттерн-кластеризации кластеры v_a^{inv} и v_b^{inv} должны быть объединены в единый порядково-инвариантный паттерн-кластер, что противоречит сделанному предположению о том, что v_a^{inv} и v_b^{inv} — различные, несовпадающие кластеры. Следовательно, $v_a^{inv} \cap v_b^{inv} = \emptyset$. Утверждение 1 доказано.

Доказательство утверждения 2. Рассмотрим случай монотонно возрастающей последовательности показателей (аналогичное доказательство можно привести и для монотонно убывающей последовательности). Докажем утверждение 2, используя метод математической индукции (для чего проверим его справедливость для множества из двух и трех объектов и предположив, что оно выполняется для случая $k > 3$ объектов, докажем его справедливость для случая $k + 1$).

1. Пусть множество X содержит только два объекта: $x_1 = (x_{11}, \dots, \dots, x_{1j}, \dots, x_{1n})$ и $x_2 = (x_{21}, \dots, x_{2j}, \dots, x_{2n})$, причем $x_{11} < \dots < x_{1j} < \dots$

$\dots < x_{1n}$, и $x_{21} < \dots < x_{2j} < \dots < x_{2n}$. Исходя из метода порядково-инвариантной паттерн-кластеризации, описанного выше, для парного сравнения всех показателей одного объекта требуется $n(n-1)/2$ сравнений, результат которых однозначно определяется выражением (2). Поскольку значения показателей первого объекта расположены в порядке их возрастания ($x_{11} < \dots < x_{1j} < \dots < x_{1n}$), получаем, что $x_{1p} < x_{1q} \forall p < q$ и $x_{1r} > x_{1s} \forall r > s$. Для объекта x_2 можно сделать такой же вывод: $x_{2p} < x_{2q} \forall p < q$ и $x_{2r} > x_{2s} \forall r > s$. Отмеченная идентичность неравенств для первого и второго объектов определяет идентичность результатов сравнения, определяемых выражением (4), и, как следствие, идентичность кодовых последовательностей, определяемых выражением (3).

Поскольку для объектов x_1 и x_2 получены одинаковые кодировки, можно сделать вывод, что данные объекты могут быть объединены в единый порядково-инвариантный паттерн-кластер.

2. Дополним анализ объектом $x_3 = (x_{31}, \dots, x_{3j}, \dots, x_{3n})$, причем $x_{31} < \dots < x_{3j} < \dots < x_{3n}$.

Поскольку для показателей объекта x_3 выполняется условие $x_{31} < \dots < x_{3j} < \dots < x_{3n}$, то, опираясь на п. 1 доказательства, можно сделать следующие выводы:

- а) x_1 и x_3 принадлежат одному порядково-инвариантному паттерн-кластеру;
- б) x_2 и x_3 принадлежат одному порядково-инвариантному паттерн-кластеру.

Согласно утверждению 1 кластеры, полученные с использованием порядково-инвариантной паттерн-кластеризации, не пересекаются. Таким образом, все три объекта образуют единый порядково-инвариантный паттерн-кластер.

3. Предположив, что утверждение 2 справедливо для случая k объектов, проверим его справедливость для случая $k+1$. С этой целью дополним данное множество новым объектом $x_{(k+1)} = (x_{(k+1)1}, \dots, x_{(k+1)j}, \dots, x_{(k+1)n})$, где $x_{(k+1)1} < \dots < x_{(k+1)j} < \dots < x_{(k+1)n}$. По аналогии с п. 1 данный объект образует единый порядково-инвариантный паттерн-кластер с любым из k предыдущих объектов. Опираясь на утверждение 1, сделаем вывод, что все $k+1$ объектов принадлежат одному порядково-инвариантному паттерн-кластеру.

Утверждение 2 доказано.

Доказательство утверждения 3. Рассмотрим кластер v_a^{inv} , полученный в результате порядково-инвариантной паттерн-кластеризации. Выделим в нем произвольный объект x_i^* и расположим его показатели в неубывающем порядке: $x_{i1}^* \leq \dots \leq x_{ij}^* \leq \dots \leq x_{in}^*$. Графически это означает, что объект представлен в системе параллельных координат в виде неубывающей кусочно-линейной функции. Данную последовательность расположения показателей $(x_{i1}^*, \dots, x_{ij}^*, \dots, x_{in}^*)$ обозначим через P' .

В силу порядковой инвариантности паттерн-кластера v_a^{inv} кусочно-линейные функции всех входящих в него объектов имеют одинаковый вид для любой последовательности показателей, в том числе последовательности P' . Следовательно, при последовательности показателей P' все объекты пред-

ставлены в системе параллельных координат в виде неубывающих кусочно-линейных функций.

Утверждение 3 доказано.

Доказательство утверждения 4. Воспользуемся теоремой, доказанной в [15]: «Два объекта x_1 и x_2 , описанные векторами $x_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1n})$ и $x_2 = (x_{21}, \dots, x_{2j}, \dots, x_{2n})$ соответственно, принадлежат одному порядково-инвариантному паттерн-кластеру тогда и только тогда, когда они могут быть представлены полными взвешенными орграфами G_1 и G_2 с идентичными весами ребер $\dots$, соединяющих их соответственные вершины». При этом значения ребер определяются формулой (4), т.е. значениями парных сравнений соответствующих вершин.

Для исходных объектов x_1 и x_2 построим соответствующие орграфы G_1 и G_2 , а также орграф G_s суммарного объекта $x_s = x_1 + x_2$. Поскольку согласно условию эти объекты принадлежат одному порядково-инвариантному паттерн-кластеру, то значения соответствующих ребер (парных сравнений) орграфов G_1 и G_2 совпадают. Необходимо показать, что с этими значениями совпадают также и значения ребер орграфа G_s . Справедливость этого утверждения следует из известных свойств:

$$\begin{cases} x_{ij} > x_{ik} \\ x_{zj} > x_{zk} \end{cases} \Rightarrow (x_{ij} + x_{zj}) > (x_{ik} + x_{zk}),$$

$$\begin{cases} x_{ij} = x_{ik} \\ x_{zj} = x_{zk} \end{cases} \Rightarrow (x_{ij} + x_{zj}) = (x_{ik} + x_{zk}),$$

$$\begin{cases} x_{ij} < x_{ik} \\ x_{zj} < x_{zk} \end{cases} \Rightarrow (x_{ij} + x_{zj}) < (x_{ik} + x_{zk}),$$

где $j, k = 1, \dots, n$.

Из этого следует, что, во-первых, кодировка нового объекта x_s будет совпадать с кодировками объектов x_1 и x_2 и, во-вторых, значения ребер орграфа G_s будут совпадать со значениями ребер G_1 и G_2 .

Утверждение 4 доказано.

СПИСОК ЛИТЕРАТУРЫ

1. Fisher R.A. The use of multiple measurements in taxonomic problems // Ann. Eugenics. 1936. V. 7. P. 179–188.
2. Shawe-Taylor J., Cristianini N. Kernel methods for pattern analysis. Cambridge Univer. press, UK. 2004.
3. Duda R.O., Hart P.E., Stork, D.G. Pattern classification. John Wiley & Sons, 2012.
4. Mirkin B.G. Clustering for Data Mining: A Data Recovery Approach, 2005.
5. Mirkin B.G. Summary and semi-average similarity criteria for individual clusters, in: Models, Algorithms, and Technologies for Network Analysis. V. 59. N.Y: Springer, 2013. P. 101–126.
6. Алескеров Ф.Т., Солодков В.М., Челнокова Д.С. Динамический анализ паттернов поведения коммерческих банков России // Эконом. журн. Высш. шк. экономики. 2006. Т. 10. № 1. С. 48–62.

7. *Aleskerov F., Ersel H., Yolalan R.* Clustering Turkish Commercial Banks According to Structural Similarities // Yapi Credit Bank. Discussion Paper Ser. 1997. No. 97-02. 24 p.
8. *Cole R., Gunther J.* A CAMEL Rating's Shelf Life. Federal Reserve Bank of Dallas. December, 1995.
9. *Gardner M., Mills D.L.* Managing Financial Institutions: An Asset. Liability Approach. The Dryden Press, 1994.
10. *Golin J.* The Bank Credit Analysis Handbook: A Guide for Analysts, Bankers and Investors John Wiley & Sons, 2001.
11. *Aleskerov F., Ersel H., Gundes C., Yolalan R.* Multicriterial Method for Personnel Allocation among Bank Branches // Yapi Kredi Discus. Paper Ser. Istanbul. 1998. No. 98-01.
12. *Aleskerov F.T., Alper C.E.* Inflation, Money, and Output Growths: Some Observations // Bogazici Univer. Res. Paper. 1996. No. SBE 96-06.
13. *Алескеров Ф.Т., Белоусова В.Ю., Егорова Л.Г., Миркин Б.Г.* Анализ паттернов в статике и динамике, часть 1// Бизнес-информатика. 2013. Т. 3. С. 3.18.
14. *Few S.* Multivariate Analysis Using Parallel Coordinates. 2006. 9p. URL: https://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf (дата обращения 28.03.2018).
15. *Мячин А.Л.* Анализ паттернов: порядково-инвариантная паттерн-кластеризация // Управление большими системами. 2016. № 61. С.41–59.
16. *Inselberg A.* Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Tokyo: Springer, 2009.
17. *Inselberg A.* The Plane with Parallel Coordinates // Visual Comput. 1985. V. 1 (2). P. 69–91.
18. *Myachin A.L.* New methods of pattern analysis in the study of Iris Anderson-Fisher Data // 6 Int. Conf. Comput. Commun. Control (ICCCC). 2016. P. 97–102.
19. *Aleskerov F.T., Egorova L.G., Gokhberg L.M. et al.* Pattern Analysis in the Study of Science, Education and Innovation Activity in Russian Regions // Procedia Comput. Sci. 2013. V. 17. P. 687–694.
20. Рейтинг инновационного развития субъектов Российской Федерации: аналитический доклад / под ред. Л.М. Гохберга. М.: Нац. исслед. универ. «Высшая школа экономики», 2012.

Статья представлена к публикации членом редколлегии А.И. Михальским.

Поступила в редакцию 17.10.2017

После доработки 02.04.2018

Принята к публикации 08.11.2018