УДК 51-74
ББК 32.973.26-018.2

АНАЛИЗ ПАТТЕРНОВ: ПОРЯДКОВО-ИНВАРИАНТНАЯ ПАТТЕРН-КЛАСТЕРИЗАЦИЯ[1,2]

Мячин А. Л.[3]
(*Национальный исследовательский университет «Высшая школа экономики», Москва, ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Представлены новые алгоритмы выделения паттернов анализируемых наборов данных на основе методов порядково-фиксированной и порядково-инвариантной паттерн-кластеризации. Приведено описание предлагаемых методов и оценки вычислительной сложности. Рассмотрены примеры, демонстрирующие их особенности и поясняющие работу соответствующих процедур кластеризации. Сформулирована и доказана теорема о взаимосвязи кластеров, полученных в результате использования порядково-инвариантной паттерн-кластеризации с полными взвешенными орграфами. Этот результат делает возможным использование теории графов для исследования свойств полученных кластеров.

---

[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», а также при поддержке Лаборатории теории выбора и анализа решений Института проблем управления им. В.А. Трапезникова РАН.*
[2] *Автор выражает благодарность д.т.н., проф. Алескерову Ф.Т. и к.т.н. Рубчинскому А.А. за помощь в написании данной статьи.*
[3] *Алексей Леонидович Мячин* (amyachin@hse.ru).

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

1. Введение

В настоящее время понятие «паттерн» широко используется в самых различных сферах деятельности. Несмотря на то, что это понятие применялось в задачах обработки информации впервые более 50 лет назад [8], современное толкование этого термина сложилось относительно недавно. Согласно [1], «под паттерном понимается такая комбинация определённых, с точностью до погрешности, значений некоторого подмножества признаков, что объекты с этими значениями достаточно сильно отличаются от других объектов».

Ранее метод анализа паттернов успешно зарекомендовал себя при решении прикладных задач в ряде областей: анализе банковской сферы [2, 5], макроэкономики [4], политологии [6].

Предлагаются к рассмотрению две модификации метода анализа паттернов: порядково-фиксированная и порядково-инвариантная паттерн-кластеризация. Обе основаны на парном сравнении выбранных показателей, однако порядково-фиксированная паттерн-кластеризация использует одну, заранее заданную их последовательность, в то время как порядково-инвариантная паттерн-кластеризация рассматривает все возможные перестановки показателей. Приведены примеры, поясняющие предложенные методы.

2. Описание метода анализа паттернов

Метод анализа паттернов базируется на выявлении схожести показателей, характеризующих внутреннюю структуру исследуемых объектов. На его основе формируются кластеры, схожие по некоторой, заранее выбранной метрике, причём объекты различных кластеров существенно отличаются. При наличии информации о количественных значениях показателей исследуемых объектов, измеренной в последовательные момен-

ты времени, возможно построение траекторий развития как динамических групп, так и отдельных объектов, а также выявление неявных взаимосвязей исходных показателей.

Приведём общее описание метода.

Пусть имеется некоторое множество исследуемых объектов X, содержащее k элементов. Каждый объект $x_i \in X$ характеризуется набором m показателей, что будем записывать в векторной форме как $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$, где x_{ij} – значение j-го показателя i-го объекта.

С целью иллюстрации метода воспользуемся графическим представлением объектов в системе параллельных координат [7]. Данное представление использует m параллельных, обычно вертикальных и равномерно распределённых координатных осей, каждая из которых отражает один из выбранных показателей. Тогда каждый объект $x_i \in X$ изображается в виде кусочно-линейной функции с вершинами на параллельных осях. Далее, вводится некоторое множество кластеров Y, состоящее из номеров (имён, меток), а также функция $d(x_i, x_l)$, позволяющая оценить меру близости объектов x_i и x_l. Исходной задачей анализа является разбиение множества X на v непересекающихся подмножеств (кластеров), содержащих близкие по метрике d объекты. При этом каждому объекту подмножества приписывается имя (метка, номер кластера) $y_i \in \boldsymbol{Y}$.

В качестве иллюстрации рассмотрим пример данных из [2]. Исследуется множество, состоящее из трех гипотетических банков по трем показателям (A, B и C), описанным в таблице 1.

Таблица 1. Показатели гипотетических банков, используемые в анализе паттернов

	A	B	C
Банк 1	50	20	40
Банк 2	55	10	45
Банк 3	10	60	20

По приведённым показателям для каждого банка построим кусочно-линейные функции (рис. 1).

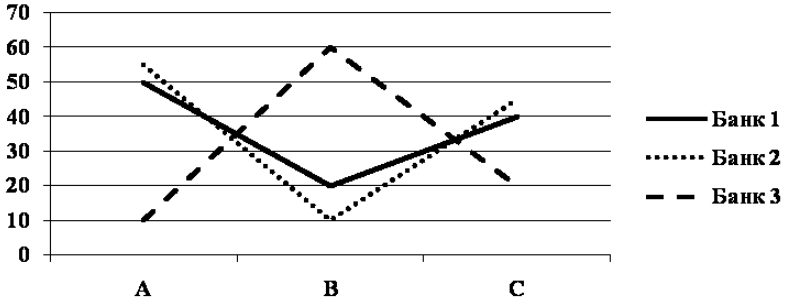

Рис. 1. Кусочно-линейные функции гипотетических банков

Рис. 1 демонстрирует, что банки 1 и 2 придерживаются схожих стратегий по выбранным показателям, тогда как стратегия банка 3 существенно отличается. В результате получим два подмножества. В первое входят банки 1 и 2, во второе – банк 3.

Следует отметить, что с целью разбиения исходного множества X на непересекающиеся подмножества возможно совместное использование методов анализа паттернов и методов кластерного анализа [9, 10].

3. Порядково-фиксированная паттерн-кластеризация

Опишем предлагаемый алгоритм порядково-фиксированной паттерн-кластеризации объектов множества X. С этой целью учтем характер парных отношений смежных показателей, а именно: объекту $x_i \in X$ ставится в соответствие последовательность символов $r_i^1, r_i^2, \ldots, r_i^j, \ldots, r_i^{m-1}$, где r_i^j определяется формулами

(1) $r_i^j = 1$ при $x_{ij} < x_{ij+1}$,

(2) $r_i^j = 0$ при $x_{ij} = x_{ij+1}$,

(3) $r_i^j = 2$ при $x_{ij} > x_{ij+1}$.

Отметим, что область значений, которую могут принимать базовые показатели исследуемого объекта, принадлежит множеству действительных чисел. С другой стороны, множество значений парных отношений, как правило, перечисляемо, т.е. дискретно и ограничено. Это дает возможность рассматривать последовательность символов $r_i^1, r_i^2, \ldots, r_i^{m-1}$ как позиционный код $r_i^1, r_i^2, \ldots, r_i^{m-1}$ некоторого числа. Такая трактовка делает удобным построение и оптимизацию автоматической процедуры кластеризации, поскольку позволяет, в частности, заменить операцию посимвольного сравнения кодов на арифметическую операцию сравнения двух чисел. В частности, возможно рассматривать последовательность $r_i^1, r_i^2, \ldots, r_i^{m-1}$ в качестве позиционного десятичного кода числа q_i, формируемого посредством формулы

$$(4) \quad q_i = \sum_{j=1}^{m-1} 10^{j-1} r_i^{m-j}.$$

Такое представление привычно и удобно в использовании, а однозначность кодирования определяется однозначностью десятичной системы счисления.

Учитывая сделанное выше замечание, исследуемые объекты x_i будем характеризовать вектором значений базовых показателей $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ и позиционным кодом $r_i = r_i^1 r_i^2 \ldots r_i^{m-1}$, характеризующим значения парных отношений смежных показателей.

Процедуру кластеризации реализуем посредством оценки меры близости формируемых кодов, для чего может быть использовано расстояние Хемминга[1]:

$$(5) \quad d(r_i, r_l) = \sum_{j=1}^{m-1} \left| r_i^j - r_l^j \right|.$$

Реализованный алгоритм предполагает, что:

1) при $d(r_i, r_l) = 0$ объекты x_i и x_l относятся к одному кластеру;

---

[1] *Имеется в виду общий случай расстояния Хемминга для кодовых последовательностей одинаковой длины произвольного алфавита* [3].

2) при $d(r_i, r_l) \neq 0$ объекты x_i и x_l относятся к разным кластерам.

Таким образом, исходное множество объектов X разбивается на ряд кластеров, число которых обозначим как v_{fix}. Дадим следующее определение.

Определение 1. Кластеризацию, проведенную описанным выше методом с заданной (т.е. с заранее фиксированной) последовательностью показателей, будем называть порядково-фиксированной паттерн-кластеризацией, а кластеры, полученные в результате ее проведения, соответственно, порядково-фиксированными паттерн-кластерами.

Определим максимальное число кластеров, на которое может быть разбито исходное множество при использовании указанной выше процедуры порядково-фиксированной паттерн-кластеризации – v^{max}_{fix}. С этой целью учтем, что рассматриваемая процедура кластеризации основана на сопоставлении кодовых последовательностей $r_i^1, r_i^2, \ldots, r_i^j, \ldots, r_i^{m-1}$, формируемых посредством выражений (1)–(3) для каждого объекта. Длина этих последовательностей равна $(m-1)$. Выражения (1)–(3) определяют также и три возможных значения, которые может принимать каждый символ. Подставляя, получим, что число различных кодовых комбинаций равно $3^{(m-1)}$, и оно определяет максимально возможное число формируемых кластеров (когда каждый кластер содержит только один объект):

$$(6) \quad v_{fix}^{\max} = 3^{(m-1)}.$$

Вычислительную сложность Z_{fix} порядково-фиксированной паттерн-кластеризации можно оценить следующим образом. Для каждого из k исследуемых объектов x_i требуется вычислить при помощи формул (1)–(3) $(m-1)$ значений r_i^j. Далее, при формировании кластеров необходимо сравнить k расстояний Хемминга d. Известно, что возможное количество парных сочетаний равно $k(k-1)/2$. Таким образом:

$$(7) \quad z_{fix} = \frac{k^2(k-1)(m-1)}{2}.$$

Продемонстрируем работу предложенного метода, используя данные примера 1. Банки 1–3 описываются векторами

x_1 = (50; 20; 40); x_2 = (55; 10; 45); x_3 = (10; 60; 20) соответственно. Таким образом, работа каждого банка охарактеризована тремя показателями, для сопоставления которых необходимо провести 2 парных сравнения. Для Банка 1:

$x_{11} > x_{12} \Rightarrow r_1^1 = 2$; $x_{12} < x_{13} \Rightarrow r_1^2 = 1$.

Следовательно, формируемая кодовая последовательность имеет вид: $r_1 = (2, 1)$.

Для Банка 2: $x_{21} > x_{22} \Rightarrow r_2^1 = 2$; $x_{22} < x_{23} \Rightarrow r_2^2 = 1 \Rightarrow r_2 = (2, 1)$.

Для Банка 3: $x_{31} < x_{32} \Rightarrow r_3^1 = 1$; $x_{32} > x_{33} \Rightarrow r_3^2 = 2 \Rightarrow r_3 = (1, 2)$.

Соответствующие им десятичные представления находятся как

$$q_1 = \sum_{j=1}^{2} 10^{j-1} r_1^{m-j} = 21; \qquad q_2 = \sum_{j=1}^{2} 10^{j-1} r_2^{m-j} = 21;$$

$$q_3 = \sum_{j=1}^{2} 10^{j-1} r_3^{m-j} = 12.$$

Для разделения банков по различным кластерам необходимо оценить расстояние Хемминга d между полученными кодовыми последовательностями:

$d(r_1, r_2) = 0$; $d(r_1, r_3) \neq 0$; $d(r_2, r_3) \neq 0$.

Полученные результаты демонстрируют идентичность кодировок паттернов банка 1 и 2. Таким образом, исходное множество из трех банков разделяется на 2 подмножества: первое образуют Банки 1 и 2, второе – Банк 3 (что согласуется с приведенным в разделе 2 результатом).

4. Порядково-инвариантная паттерн-кластеризация

В этом разделе предлагается иной метод анализа паттернов – порядково-инвариантная паттерн-кластеризация и рассматриваются отдельные его свойства. Для демонстрации лежащей в его основе идеи вновь рассмотрим множество, состоящее из трех гипотетических банков, однако с несколько иными значениями показателей, представленными в таблице 2.

Таблица 2. Показатели гипотетических банков

	A'	B'	C'
Банк 1	50	10	50
Банк 2	50	0	70
Банк 3	50	20	30

Используя значения таблицы 2, вновь построим кусочно-линейные функции (рис. 2).

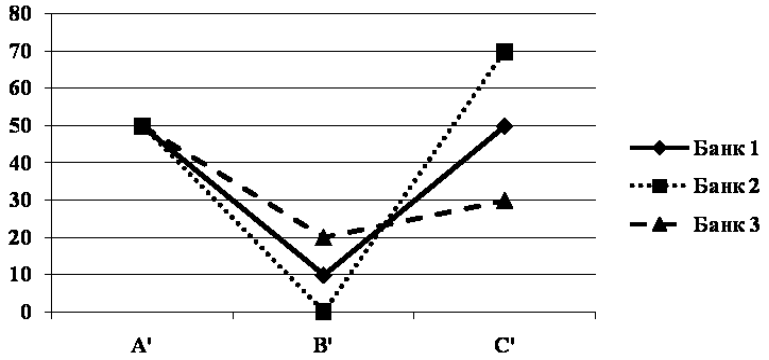

Рис. 2. Кусочно-линейные функции Банков 1, 2 и 3

Все три графика имеют схожий вид, что свидетельствует о возможности объединения их в единый кластер. Проверим этот вывод, используя процедуру порядково-фиксированной паттерн-кластеризации:

$$x_{12} < x_{13} \Rightarrow r_1^2 = 1; \qquad x_{22} < x_{23} \Rightarrow r_2^2 = 1;$$

$$x_{32} < x_{33} \Rightarrow r_3^2 = 1.$$

Соответственно:

$$q_1 = \sum_{j=1}^{2} 10^{j-1} r_1^{m-j} = 21; \qquad q_2 = \sum_{j=1}^{2} 10^{j-1} r_2^{m-j} = 21;$$

$$q_3 = \sum_{j=1}^{2} 10^{j-1} r_3^{m-j} = 21 \quad \text{и}$$

$d(r_1, r_2) = 0$; $d(r_1, r_3) = 0$; $d(r_2, r_3) = 0$.

Таким образом, согласно полученным результатам все три исследуемых банка можно объединить в единый кластер.

Однако проверим, сохранится ли данный вывод, если анализировать значения показателей A', B', C' в иной последовательности, например A', C', B' или B', A', C'.

Для наглядности построим кусочно-линейные функции (рис. 3).

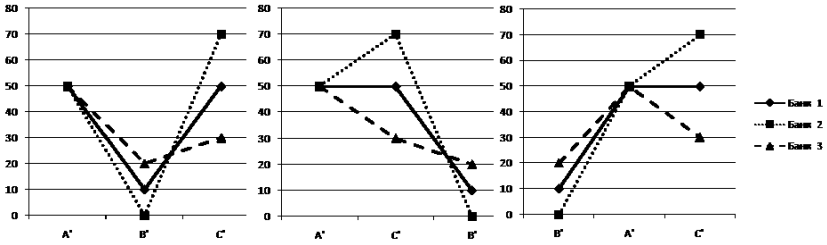

Рис. 3. Кусочно-линейные функции для различных последовательностей показателей A', B' и C'

Рисунок наглядно демонстрирует, что кусочно-линейные функции теряют свою схожесть при изменении последовательности анализируемых показателей. Аналогичный вывод дает и метод порядково-фиксированной паттерн-кластеризации, примененный для случая рассматриваемых последовательностей показателей.

Основная идея метода порядково-инвариантной паттерн-кластеризации состоит в поиске и объединении в единые кластеры объектов, которые не меняют свою принадлежность кластеру при любом изменении последовательности анализируемых показателей.

Отметим, что само существование таких кластеров далеко не очевидно. Поэтому продемонстрируем саму возможность порядково-инвариантной паттерн-кластеризации, используя данные примера 1. Построим кусочно-линейные функции для различных последовательностей показателей A, B и C (рис. 4–6).

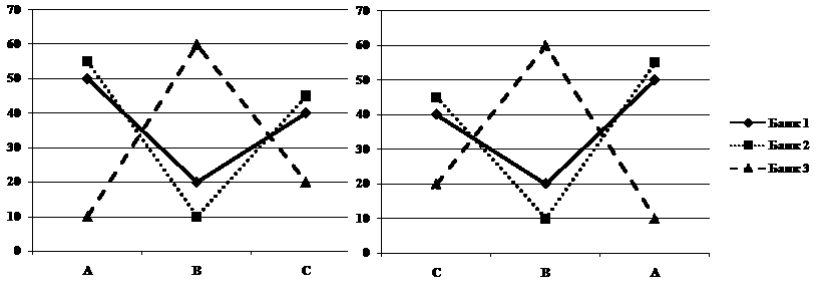

Рис. 4. Кусочно-линейные функции показателей «ABC» и «CBA»

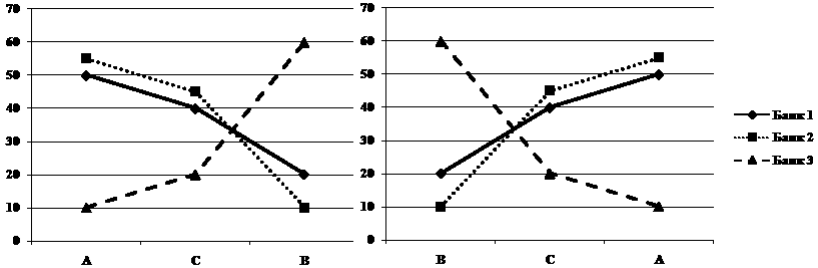

Рис. 5. Кусочно-линейные функции показателей «ACB» и «BCA»

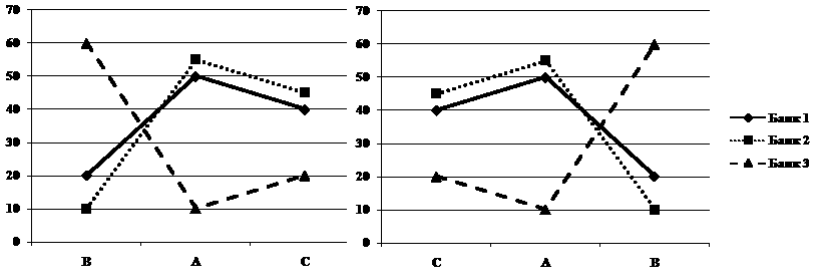

Рис. 6. Кусочно-линейные функции показателей «BAC» и «CAB»

Приведенный рисунок наглядно демонстрирует, что Банки 1 и 2 продолжают демонстрировать схожий, а Банк 3 – иной

характер кусочно-линейных функций, вне зависимости от последовательности рассматриваемых показателей.

Определение 2. Процедуру кластеризации, результат которой не зависит от последовательности показателей исследуемых объектов, будем называть «порядково-инвариантной паттерн-кластеризацией», а кластеры, полученные в результаты ее применения, соответственно, порядко-инвариантными паттерн-кластерами.

Перейдем к непосредственному описанию метода. Как и ранее, исходные объекты определены множеством X: $|X| = k$. Каждый из объектов $x_i \in X$, $i = 1, 2, \ldots, k$, представлен системой m показателей, что записывается в векторной форме как $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. Значения символов r_i^j однозначно определены формулами (1)–(3).

Объекты представлены графически в системе параллельных координат. Учтем, что различная последовательность анализируемых показателей приводит в общем случае к построению различных кусочно-линейных функций и, как следствие, формированию различных порядково-фиксированных паттерн-кластеров. Поэтому для реализации порядково-инвариантной кластеризации необходимо выполнить порядково-фиксированную паттерн-кластеризацию для всех возможных перестановок исходных показателей и затем объединить те объекты, которые остаются в едином кластере вне зависимости от рассматриваемой последовательности показателей.

Число различных перестановок P_m по m показателям определяется формулой

(8) $P_m = m!$,

что свидетельствует о высокой вычислительной сложности непосредственной реализации порядково-инвариантной паттерн-кластеризации при последовательном переборе всех возможных последовательностей исследуемых показателей. Однако в действительности не все комбинации нуждаются в рассмотрении. Так, например, использование прямой и обратной последовательности наборов значений показателей $(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{i(m-1)}, x_{im})$ и $(x_{im}, x_{i(m-1)}, \ldots, x_{i3}, x_{i2}, x_{i1})$ дает одинаковый результат кластеризации (что графически определяется рас-

смотрением кусочно-линейных функций «слева-направо» и «справа-налево»). Поэтому возможно рассмотрение не всех, а лишь половины возможных перестановок:

(9) $P'_m = \frac{m!}{2}$.

Еще более существенное снижение числа необходимых к рассмотрению перестановок достигается путем применения доказанной ниже теоремы, которая позволяет не только снизить вычислительную сложность метода, но и является основой рассматриваемого алгоритма порядково-инвариантной паттерн-кластеризации.

Выполним следующее построение. Для каждого объекта построим полный, взвешенный орграф, вершины которого соответствуют значениям показателей, а веса ребер – значениям парных отношений (определяемых выражениями 1–3). На данном этапе будем полагать, что каждые две вершины орграфа соединяются двумя ребрами, имеющими противоположные направления. В этом случае произвольная последовательность анализируемых показателей объекта определяет конкретный гамильтонов путь в рассматриваемом орграфе. Множество всех возможных перестановок показателей соответствует множеству всех возможных гамильтоновых путей орграфа.

Теорема. Два объекта x_1 и x_2, описанные векторами $x_1 = (x_{11}, x_{12}, \ldots, x_{1m})$ и $x_2 = (x_{21}, x_{22}, \ldots, x_{2m})$ 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 (определяемых выражениями 3–5), соединяющих их соответственные вершины.

Доказательство. Пусть два объекта x_1 и x_2 принадлежат одному порядково-инвариантному паттерн-кластеру. Рассмотрим произвольную последовательность анализируемых показателей. Этой последовательности соответствует определенный гамильтонов путь в орграфах G_1 и G_2.

Согласно условию, объекты x_1 и x_2 принадлежат одному порядково-инвариантному паттерн-кластеру. Это означает, что

отношения между значениями соседних показателей, определяемых формулами (1)–(3), одинаковы для обоих объектов. Повторяя это рассуждение для всех возможных гамильтоновых путей, охватывающих все ребра графов G_1 и G_2, приходим к утверждению теоремы о идентичности весовых значений их ребер, соединяющих соответственные вершины.

Обратно, пусть два объекта $x_1 = (x_{11}, x_{12}, \ldots, x_{1m})$ и $x_2 = (x_{21}, x_{22}, \ldots, x_{2m})$ представлены полными взвешенными орграфами G_1 и G_2, а вес ребер, соединяющие соответственные вершины в обоих графах, имеет одинаковые значения. Рассмотрим произвольную последовательность исследуемых показателей. Этой последовательности соответствует одинаковый гамильтонов путь в обоих графах. Поскольку, в силу условия, ребра гамильтоновых путей, соединяющих соответственные вершины, имеют одинаковые значения, то и отношения значений показателей соответственных вершин идентичны, что означает принадлежность объектов к одному порядково-фиксированному паттерн-кластеру. Повторяя данное рассуждение для всех возможных последовательностей исследуемых показателей, приходим к выводу о принадлежности обоих объектов одному порядково-фиксированному паттерн-кластеру вне зависимости от самой последовательности, что и определяет их принадлежность одному порядково-инвариантному паттерн-кластеру.

Теорема доказана.

Доказанная выше теорема о взаимном соответствии объектов, принадлежащих порядково-инвариантному паттерн-кластеру, и полного взвешенного орграфа позволяет использовать его в качестве удобного инструментария графического представления и исследования свойств порядково-инвариантных паттерн-кластеров.

Выше предполагалось, что две вершины V_{if} и V_{ig} полного взвешенного орграфа соединяются двумя ребрами $V_{if}V_{ig}$ и $V_{ig}V_{if}$, имеющими противоположные направления и различные веса. Это предположение является удобным для построения произвольных гамильтоновых путей орграфа, однако является избыточным для описания порядково-инвариантных паттерн-

кластеров, поскольку вес ребра $V_{if}V_{ig}$, определяемый выражениями (1)–(3), однозначно определяет вес ребра обратного направления $V_{ig}V_{if}$.

Следствием этого является возможность выбора направлений ребер орграфа, удобного с точки зрения анализа данных и программной реализации порядково-инвариантной паттерн-кластеризации.

Опишем используемый алгоритм порядково-инвариантной паттерн-кластеризации, основанный на доказанной выше теореме.

Для каждого объекта формируется характеризующий его орграф. Для удобства программной реализации использована процедура последовательного обхода вершин графа с формированием ребер по направлению от заданной вершины ко всем последующим, как это показано на рис. 7 (направление ребер не указано для простоты рисунка).

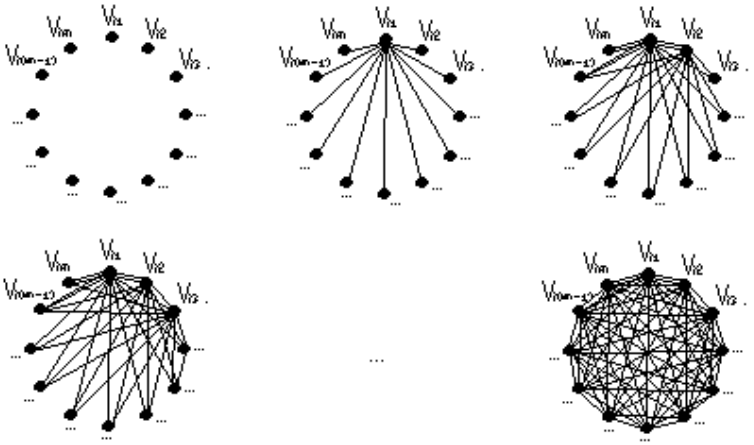

Рис. 7. Формирование орграфа порядково-инвариантного паттерн-кластера

Далее, используя выражения (1)–(3), определяются весовые значения каждого ребра. Поскольку каждая пара вершин фор-

мируемого орграфа соединена одним ребром, то их общее число u_{inv} определяется известным выражением:

(10) $$u_{inv} = \frac{m(m-1)}{2},$$

что и определяет количество необходимых парных сравнений для формирования весовых значений.

Далее реализуется процедура кластеризации, которая (согласно теореме) помещает в единый кластер объекты, характеризующиеся сформированными орграфами с идентичными весовыми показателями соответственных ребер.

Выше качественно (на основании графиков) продемонстрирована идея порядково-инвариантной паттерн-кластеризации, используя данные примера 1, представленные в таблице 1. Вновь воспользуемся этими данными, чтобы наглядно продемонстрировать работу алгоритма порядково-инвариантной паттерн-кластеризации.

Отметим, что таблица 1 содержит данные объектов (банков), работа которых характеризуется тремя показателями: A, B и C. Каждому из этих объектов поставим в соответствие полный взвешенный орграф, число вершин которого $m = 3$, а их значения соответствуют значениям показателей A, B и C соответствующего объекта (рис. 8). Число ребер, вес которых определяется характером отношений соединяемых вершин (см. формулы 1–3), также равно трем.

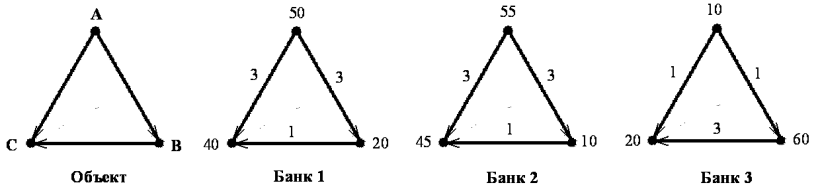

Рис. 8. Представление объектов таблицы 1 в форме полных взвешенных орграфов

Максимальное число возможных последовательностей расположения показателей A, B и C определяется формулой (8):

$P_3 = 3! = 6$. Однако, как отмечалось выше, для реализации порядково-инвариантной паттерн-кластеризации нет необходимости рассматривать их все. Достаточно ограничиться числом ребер формируемого орграфа, определяемого формулой (10):

$$u_{inv} = \frac{m(m-1)}{2} = 3.$$

Поэтому ограничимся тремя последовательностями показателей, выбрав для наглядности ABC, ACB и BAC (данные последовательности выбраны в качестве примера для наглядного сопоставления с рис. 5–7). Возможно также использование как обратных последовательностей (CBA, BCA и CAB), так и их комбинаций. Запишем соответствующие им позиционные коды, определяемые (с учетом направления) весовыми значениями ребер графа, согласно формулам (1)–(3).

Таблица 3. Кодировки гипотетических банков

Объект	Значение формируемого кода для рассматриваемой последовательности показателей		
	ABC	ACB	BAC
Банк 1	21	22	12
Банк 2	21	22	12
Банк 3	12	11	21

Как видно из приведенной таблицы, значение формируемых кодов Банков 1 и 2 совпадает для всех анализируемых последовательностей показателей, что предопределяет нулевое значение расстояния Хемминга между кодовыми последовательностями. Поэтому, Банк 1 и Банк 2 помещают в один порядково-инвариантный паттерн-кластер. Банк 3, соответственно, размещается в другом порядково-инвариантном паттерн-кластере. Отметим, что полученный результат полностью согласуется с результатом, полученным ранее графически.

5. Заключение

Описан алгоритм выделения паттернов на основе предложенного метода «порядково-инвариантной паттерн-кластеризации». Приведены примеры, поясняющие суть метода и работу соответствующей процедуры кластеризации.

Сформулирована и доказана теорема, описывающая представление порядково-инвариантных паттерн-кластеров в форме полных взвешенных орграфов. Возможность установления взаимно однозначного соответствия паттерна порядково-инвариантного паттерн-кластера и соответствующего ему полного взвешенного орграфа позволяет использовать методы теории графов для исследования свойств таких кластеров.

В связи с простотой использования и невысокой вычислительной сложностью предложенный метод позволяют не только обрабатывать большие массивы данных, но и существенно снижать временные затраты.

Литература

1. АЛЕСКЕРОВ Ф.Т., БЕЛОУСОВА В.Ю., ЕГОРОВА Л.Г. и др. *Анализ паттернов в статистике и динамике. Часть 1: Обзор литературы и уточнение понятия* // Бизнес-информатика. – 2013. – Т. 3. – С. 3–18.
2. 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. – С. 48–62.
3. БЛЕЙХУТ Р. *Теория и практика кодов, контролирующих ошибки.* – М.: Мир, 1986. – 576 с.
4. ALESKEROV F., ALPER C.E. *A clustering approach to some monetary facts: a long-run analysis of cross-country data* // The Japanese Economic Review. – 2000. – Vol. 51, No. 4. – P. 555–567.
5. ALESKEROV F., ERSEL H., YOLALAN R. *Multicriterial Ranking Approach for Evaluating Bank Branch Performance* //

International Journal of Information Technology and Decision Making. – 2004. – Vol. 3, No. 2. – P. 321–335.

6. 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. – Vol. 48. – P. 1385–1395.
7. FEW S. *Multivariate Analysis Using Parallel Coordinates*. – 2006. – [Электронный ресурс] – URL: http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf (дата обращения: 11.01.2016).
8. FISHER R.A. *The use of multiple measurements in taxonomic problems* // Annals of Eugenics. – Vol. 7. – 1936. – P. 179–188.
9. MIRKIN B. *Summary and semi-average similarity criteria for individual clusters, in: Models, Algorithms, and Technologies for Network Analysis* / Ed. by B.I. Goldengorin, V.A. Kalyagin, P.M. Pardalos. – Vol. 59. – NY: Springer, 2013. – P. 101–126.
10. MIRKIN B. *Clustering for Data Mining: A Data Recovery Approach. Boca Raton: Chapman-Hall/CRC.* – Taylor and Francis Group, 2005. – 350 p.

PATTERN ANALYSIS: ORDINAL-INVARIANT PATTERN-CLUSTERING

Alexey Myachin National Research University Higher School of Economics, Moscow, Institute of Control Sciences of RAS, Moscow, (amyachin @hse.ru).

Abstract: *New algorithms of patterns analysis based on methods of ordinal-fixed and ordinal-invariant pattern clustering are developed. The definition of the proposed methods as well as the evaluation of the computational complexity is given. We provide some examples that demonstrate features of these clustering procedures and explain their operation. We also formulate and prove the theorem on the interconnection of clusters obtained by the use of ordinal-invariant pattern-clustering with complete weighted digraphs. These results allow to apply graph theory for the study of properties of obtained clusters.*

Keywords: pattern analysis; ordinal-fixed pattern clustering; ordinal-invariant pattern clustering; cluster analysis.

Статья представлена к публикации
членом редакционной коллегии Н.И. Базенковым

Поступила в редакцию 13.01.2016.
Опубликована 31.05.2016.