ОПРЕДЕЛЕНИЕ ЦЕНТРОИДОВ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ ПОРЯДКОВО-ИНВАРИАНТНОЙ ПАТТЕРН-КЛАСТЕРИЗАЦИИ[1]

Мячин А. Л.[2]

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

Работа продолжает исследования, направленные на создание методов анализа паттернов в системе параллельных координат с независимыми от последовательности входных данных результатов. Описаны основные операции над объектами порядково-инвариантных паттерн-кластеров. Доказано утверждение о принадлежности центроида порядково-инвариантного паттерн-кластера исходному кластеру, что позволяет проводить оценку внутрикластерных расстояний «объект – центроид» в многомерном пространстве признаков. Приведены примеры выявления структурной схожести объектов в системе параллельных координат. Отмечены основные отличия методов анализа паттернов и кластерного анализа. Описана методология выявления центроида порядково-инвариантного паттерн-кластера. Предложен алгоритм объединения групп объектов на базе их структурной схожести – с одной стороны, и минимизации внутрикластерных расстояний – с другой, что позволяет повысить точность конечных результатов и частично решить проблему поиска качественно близких объектов при наличии погрешности в исходных данных. Предложенный алгоритм использует понятие внутрикластерных расстояний «объект – центроид» и удовлетворяет следующим условиям: эндогенное определение как количества, так и состава искомых групп изучаемых объектов; невысокая (относительно) вычислительная сложность; независимость исходного разбиения от изначальной последовательности входных данных. Продемонстрирована работа предложенного алгоритма на классических наборах данных. Приведены результаты тестирования и отмечено повышение точности кластеризации.

Ключевые слова: паттерн, анализ паттернов, кластерный анализ.

---

[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] *Алексей Леонидович Мячин, к.т.н.* (*amyachin@hse.ru*).

1. Введение

Анализ паттернов является современной и активно развивающейся областью анализа данных. Выявление структурной схожести изучаемых объектов среди большого количества разнородных данных является непростой, но весьма важной задачей, решение которой существенно усложняется при наличии погрешностей в исходных данных. Подобные погрешности возникают на различных этапах сбора данных, переноса на цифровые носители, а также при округлении и приближенных вычислениях. Как указано в [4], «В большинстве случаев вычисления производятся с приближенными числами и притом приближенно. Поэтому даже для точного метода решения задачи на каждом этапе вычислений возникают погрешность действий и погрешность округлений. При неблагоприятных обстоятельствах суммарная погрешность может быть столь велика, что полученный результат будет иметь лишь иллюзорное значение».

Обнаружение выбросов возможно при использовании стандартных программ, к примеру, Microsoft Excel. Однако обнаружение погрешностей при помощи стандартных методов математической статистики либо последовательной ручной проверке является непростой, а часто и нерешаемой в условиях ограниченного времени задачей. Одним из возможных и часто используемых подходов является изначальное округление исходных данных до определенного значения, диктуемого исходной постановкой задачи и ожидаемыми результатами. Однако подобное решение не всегда применимо для различных методов анализа, к примеру, основанных на парном сравнении показателей, поскольку может значительным образом сказываться на конечных результатах. В связи с этим в работе предлагается другой подход, позволяющий повысить точность конечного разбиения за счет двухэтапного подхода к решению задачи. Изначально использование порядково-инвариантной паттерн-кластеризации [6] позволяет получить базовое разбиение, не зависящее от исходной последовательности показателей. Далее поиск центроида каждой группы и оценка близости исследуемых объектов до каждого из них позволяет скорректировать полученные на первом этапе результаты.

В развиваемой в работе методологии предъявляются определенные требования к входным данным. Во-первых, предполагается использование только количественных шкал измерения данных либо, при наличии бинарных и номинальных шкал, имеется возможность их перевода к количественному виду. Во-вторых, отсутствуют пропуски в данных либо имеется возможность их заполнения. Учтем также рекомендации стохастического анализа, согласно которым не рекомендуется включать в анализ высоко коррелированные показатели [8].

Статья содержит краткое описание методов анализа паттернов, в том числе порядково-инвариантной паттерн-кластеризации, являющейся основой для построения предложенного в работе алгоритма объединения объектов в единые группы, методологию выявления центроидов порядково-инвариантных паттерн-кластеров и доказательство утверждения о принадлежности данного центроида к определенному порядково-инвариантному паттерн-кластеру, а также демонстрацию использования описываемого подхода на классических тестовых данных.

2. Выявление структурной схожести объектов

2.1. КЛАСТЕРНЫЙ АНАЛИЗ И АНАЛИЗ ПАТТЕРНОВ

Выявление отдельных закономерностей в данных является широко востребованной на сегодняшний день задачей. Одним из наиболее популярных подходов является использование методов кластерного анализа. Согласно [5] «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 элементов».

Обзору методов кластерного анализа посвящено множество работ [5, 12]. Одним из наиболее известных методов является *k*-means [15], различные модификации которого [14] используются в настоящее время при решении прикладных задач в различных областях, а эквивалентные переформулировки его кри-

терия представлены в [16]. Методы кластерного анализа также используются и для обнаружения выбросов в данных [13].

В классическом понимании термин «паттерн» не является однозначным и на практике используются различные подходы к его определению в различных областях знаний. В частности, в [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»; в [18] под паттерном предлагается понимать «любые отношения, закономерности или структуру, присущую некоторому набору данных». В данной работе мы будем использовать обобщение данных определений, и для однозначности под паттерном будем понимать комбинацию качественно похожих признаков.

Анализу паттернов посвящено множество работ [2, 3, 18], в которых затронуты как теоретических аспекты, так и применение методов анализа паттернов к конкретным прикладным данным.

Кратко приведем формальное описание. Имеется некоторая исследуемая выборка из n объектов x_i, описанных векторами $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$, где через x_{ij} будем обозначать j-й показатель i-го объекта. Используя определенную меру близости, требуется разбить все объекты на схожие группы. При этом объекты одной группы должны быть максимально схожими (согласно выбранной мере близости), а объекты разных групп – максимально отличаться. Для визуализации традиционно используется система параллельных координат [11], в которой параллельно распределенные оси характеризуют исследуемые показатели. Важным отличием классических методов кластерного анализа от методов анализа паттернов является объединение не только близких по абсолютным значениям, но и качественно схожих объектов у последних. Отметим также, что анализ паттернов относится к классу алгоритмов, для которых число формируемых групп объектов заранее не задается и определяется в процессе использования соответствующих методов.

Отдельным и весьма важным вопросом является качество конечного разбиения. Для ответа приведем уже ставшим клас-

сическим пример из [9], наглядно описанный и продемонстрированный в [12].

На рис. 1 приведено сравнение результатов использования различных методов на одинаковом наборе данных. Из [12]: «Нет лучшего алгоритма кластеризации. Каждый алгоритм, в явном виде или нет, предполагает определенную структурированность данных; если совпадение «хорошее», алгоритм – успешный». Таким образом, использование того или иного метода всегда должно быть обосновано исходными данными и ожидаемыми результатами. Данное утверждение верно и для методов анализа паттернов. Однако в связи с использованием системы параллельных координат здесь появляется дополнительное условие: независимость конечных результатов от выбора последовательности входных данных.

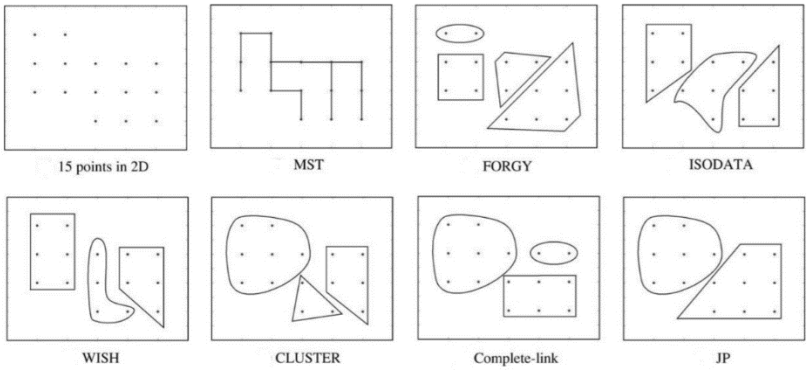

Рис. 1. Сравнение различных методов кластерного анализа. Рисунок взят из [*12*] (*исходные данные* [9])

2.2. ВЫЯВЛЕНИЕ ПАТТЕРНОВ С НЕЗАВИСИМЫМИ ОТ ВЫБОРА ИСХОДНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ПОКАЗАТЕЛЕЙ КОНЕЧНЫХ РЕЗУЛЬТАТОВ

Рассмотрим методы анализа паттернов, результат которых не изменится при выборе альтернативной последовательности исходной выборки данных. Подробно данные методы описаны в [6, 7]. Приведем краткую алгоритмическую реализацию, согласно [6].

Как описано ранее, в исходной постановке задачи исследуется некоторая выборка из n объектов x_i, каждый из которых описан вектором $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$. Для каждого объекта x_i составим матрицу парных сравнений

$$(1) \quad A_i = \begin{pmatrix} a_{11}^i & a_{12}^i & \ldots & a_{1m}^i \\ a_{21}^i & a_{22}^i & \ldots & a_{2m}^i \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1}^i & a_{m2}^i & \ldots & a_{mm}^i \end{pmatrix},$$

где

$$(2) \quad a_{wj}^i = \begin{cases} 1, & \text{если\_} x_{iw} < x_{ij}, \\ 0, & \text{если\_} x_{iw} = x_{ij}, \\ 2, & \text{если\_} x_{iw} > x_{ij}. \end{cases}$$

Поскольку любые парные сравнения, согласно предъявленным к исходным данным требованиям, определены однозначно ($x_{iw} > x_{ij} \Rightarrow x_{ij} < x_{iw}$), для дальнейшего сравнения могут быть использованы только элементы матрицы A_i, стоящие выше главной диагонали:

$$(3) \quad A_i^* = \begin{pmatrix} \ldots & a_{12}^i & a_{13}^i & \ldots & a_{1m}^i \\ & \ldots & a_{23}^i & \ldots & a_{2m}^i \\ & & \ldots & \ldots & \ldots \\ & & & \ldots & a_{m-1,m}^i \\ & & & & \ldots \end{pmatrix}.$$

В [6] показано, что объекты могут быть отнесены к одному порядково-инвариантному паттерн-кластеру, если у них совпадают результаты парных сравнений показателей, определяемых выражением (2). Таким образом, согласно пояснению выше, объекты могут быть отнесены к одному порядково-инвариантному паттерн-кластеру, если их матрицы вида A_i^* совпадают.

Таким образом, для получения конечного разбиения используется следующий критерий: при $A_i^* = A_z^*$ объекты x_i и x_z объединяются в единую группу, в противном случае – разделяются.

2.3. ВЫЯВЛЕНИЕ СРЕДНИХ ОБЪЕКТОВ ПОРЯДКОВО-ИНВАРИАНТНЫХ ПАТТЕРН-КЛАСТЕРОВ

Для решения ряда задач важным является наличие возможности объединения относительно небольших кластеров в более крупные. Часто такая необходимость обуславливается наличием погрешности в исходных данных, что в определенной степени влияет на конечные результаты. В рассматриваемом выше методе соотнесение объектов к той или иной группе определяется результатом парных сравнений типа «больше», «меньше» и «равно», согласно формуле (2). При округлении данных возможно (при наличии небольших, трудных для обнаружения и автоматического исправления погрешностей в характеризующих исходные объекты данных) разбиение объектов, хотя и качественно близких, но имеющих различные кодировки. При работе с небольшими массивами данных очевидным решением является выявление подобных погрешностей вручную (что несложно, учитывая схожесть в визуальном представлении в системе параллельных координат данных объектов) и их исправление. Однако при работе с большими массивами, содержащими разнородные данные, важным является создание алгоритмов, позволяющих автоматически решить эту проблему. Одним из возможных подходов является определение центроида каждой группы для их сопоставления.

С этой целью дадим определение центроида порядково-инвариантного паттерн-кластера.

Определение 1. Под центроидом порядково-инвариантного паттерн-кластера v будем понимать объект вида

$$x^{v}_{average} = \frac{1}{|v|}\sum_{i=1}^{|v|} x_i, \quad (4)$$

где $|v|$ – количество объектов, входящих в соответствующий кластер v.

Важным является вопрос принадлежности $x^{v}_{average}$ исходному порядково-инвариантный паттерн-кластеру. В связи с этим приведем следующее утверждение.

Утверждение 1. Центроид $x^{v}_{average}$ всегда принадлежит тому же порядково-инвариантному паттерн-кластеру, что и объекты, на основе значений показателей которых он образован.

Доказательство. Рассмотрим два произвольных объекта $x_b = (x_{b1}, x_{b2}, \ldots, x_{bm})$ и $x_c = (x_{c1}, x_{c2}, \ldots, x_{cm})$, входящих в кластер v. Принадлежность объектов x_b и x_c одному порядково-инвариантному паттерн-кластеру означает равенство матриц A_b^* и A_c^*, равно как и результатов парных сравнений показателей каждого объекта (2). Тогда справедливы утверждения:

$$\begin{cases} x_{bj} > x_{bs}, \\ x_{cj} > x_{cs}; \end{cases} \Rightarrow (x_{bj} + x_{cj}) > (x_{bs} + x_{cs}).$$

$$\begin{cases} x_{bj} = x_{bs}, \\ x_{cj} = x_{cs}; \end{cases} \Rightarrow (x_{bj} + x_{cj}) = (x_{bs} + x_{cs}),$$

$$\begin{cases} x_{bj} < x_{bs}, \\ x_{cj} < x_{cs}; \end{cases} \Rightarrow (x_{bj} + x_{cj}) < (x_{bs} + x_{cs}).$$

Данные утверждения верны и для суммы всех n объектов, входящих в кластер v. Далее, возьмем целое число $\lambda > 1$. Верно, что если $x_{bj} > x_{bs}$, то $x_{bj}/\lambda > x_{bs}/\lambda$. Аналогичное соотношение верно и для случаев $x_{bj} < x_{bs}$ и $x_{bj} = x_{bs}$.

Соответственно, для всех объектов кластера v справедливы утверждения:

$$\begin{cases} x_{bj} > x_{bs}, \\ x_{cj} > x_{cs}, \\ \quad \ldots \\ x_{gj} > x_{gs}; \end{cases} \Rightarrow \frac{1}{|v|}\sum_i x_{ij} > \frac{1}{|v|}\sum_i x_{is},$$

$$\begin{cases} x_{bj} = x_{bs}, \\ x_{cj} = x_{cs}, \\ \quad \ldots \\ x_{gj} = x_{gs}; \end{cases} \Rightarrow \frac{1}{|v|}\sum_i x_{ij} = \frac{1}{|v|}\sum_i x_{is},$$

$$\begin{cases} x_{bj} < x_{bs}, \\ x_{cj} < x_{cs}, \\ \quad \ldots \\ x_{gj} < x_{gs}; \end{cases} \Rightarrow \frac{1}{|v|}\sum_i x_{ij} < \frac{1}{|v|}\sum_i x_{is}.$$

Из определения 1 следует, что

$$x^{v}_{average} = \left(\frac{1}{|v|}\sum_{i=1}^{|v|} x_{i1}, \frac{1}{|v|}\sum_{i=1}^{|v|} x_{i2}, ..., \frac{1}{|v|}\sum_{i=1}^{|v|} x_{ij}, ..., \frac{1}{|v|}\sum_{i=1}^{|v|} x_{im} \right).$$

Таким образом, поскольку результаты парных сравнений показателей объекта $x^{v}_{average}$ будут аналогичны результатам парных сравнений соответствующих показателей любого объекта кластера v, то и матрица центроида $A^{*}_{average}$ будет совпадать с матрицами объектов порядково-инвариантного паттерн-кластера. Следовательно, объект $x^{v}_{average}$ также будет принадлежать данному кластеру.

Утверждение доказано.

Далее, поскольку возможно однозначное соотнесение центроидов к определенному порядково-инвариантному паттерн-кластеру, приведем методологию последовательного объединения объектов.

2.3. ОБЪЕДИНЕНИЕ ОБЪЕКТОВ НА ОСНОВЕ ОЦЕНКИ ИХ РАССТОЯНИЯ ДО ЦЕНТРОИДОВ ПОРЯДКОВО-ИНВАРИАНТНЫХ ПАТТЕРН-КЛАСТЕРОВ

Приведем алгоритмическую реализацию предложенной методологии. Пусть имеется v_{inv} порядково-инвариантных паттерн-кластеров. Для каждого из них рассчитывается центроид $x^{v}_{average}$. Для всех объектов x_i вычисляется евклидово расстояние $d^{v}_{E}(x_i, x^{v}_{average})$ согласно формуле

$$(5) \quad d^{v}_{E}(x_i, x^{v}_{average}) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x^{v}_{average})^2}.$$

Таким образом, для каждого объекта рассчитывается v_{inv} расстояний d^{v}_{E} (соответствующее количеству порядково-инвариантных паттерн-кластеров). Каждый объект относим к группе, расстояние до центроида которой минимально. Другими словами, для всех x_i рассчитывается дополнительный параметр p_i объединения объектов согласно формуле

$$(6) \quad p_i = \min\left(d^{1}_{E}(x_i, x^{1}_{average}), ..., d^{1}_{E}(x_i, x^{v_{inv}}_{average})\right).$$

При $p_i = d^{v}_{E}$ объект x_i относим к группе v. Таким образом происходит корректировка результатов порядково-

инвариантной паттерн-кластеризации, повышающая плотность группировки объектов (за счет минимизации расстояния) вокруг найденных центроидов. При этом процедура корректировки не предполагает перерасчет центроидов ранее выделенных кластеров.

В случае если минимум в формуле (6) достигается для нескольких центроидов, то на основании визуального анализа предпочтение отдается тому из них, форма графика которого наиболее близка форме графика объекта.

Замечание 1. Целесообразность применения подобной модели определяется постановкой конкретной задачи и ожидаемыми результатами разбиения. Перегруппировка объектов после применения порядково-инваринатной паттерн-кластеризации в ряде случаев может объединить качественно схожие объекты, абсолютные значения показателей которых существенным образом разнятся.

Необходимость такого разбиения требует понимания изучаемой структуры объектов, а также практической интерпретации конечных результатов.

3. Тестирование на классических тестовых данных

В разделе исследована предложенная методология применительно к классическим тестовым данным Wine Data [19] и Iris Data [10].

3.1. ТОЧНОСТЬ РАЗБИЕНИЯ ПРИ ОПРЕДЕЛЕНИИ ЦЕНТРОИДОВ

Прежде всего продемонстрируем целесообразность использования понятия «центроид» для получения конечного разбиения при заранее известных результатах с целью определения точности разбиения и демонстрации возможности улучшения результатов порядково-инвариантной паттерн-кластеризации. Для этой цели используем набор данных Wine Data [19], описывающий по 13 показателям различные марки вин, изготавливаемых на трех винодельнях. Конкретные показатели представлены в таблице 1.

Таблица 1. Показатели химических и физических свойств

№	**Оригинальное название**	**Краткое описание**
1.	Alcohol	Алкоголь – процентное содержание алкоголя в вине (% по объему)
2.	Malic acid	Малеиновая кислота – одна из основных органических кислот, встречающихся в винограде (г/л)
3.	Ash	Зольность – содержание золы является одним из важных показателей в определении качества вина (мС/см)
4.	Alcalinity of ash	Алкалин – несколько различных типов кислот, найденных в вине, влияют на кислотный вкус вина (PH)
5.	Magnesium	Магний – содержание магния в винах (гм на 1 кг)
6.	Total phenols	Общее содержание фенолов – флаваноиды, которые способствуют созданию различных танинов и способствуют восприятию горечи в вине (Мг/л)
7.	Lavanoids	Флаваноиды – самые распространенные полифенолы в вине (Мг/л)
8.	Nonflavanoid phenols	Нефлаваноидные фенолы – фенольные соединения, вносящие специфический вкус и аромат и возникающие в результате сложных взаимодействий, происходящих в вине во время ферментации и виноделия (Мг/л)
9.	Proanthocyanins	Проантоцианины – класс фенола (Мг/л)
10.	Color intensity	Интенсивность цвета – простая мера того, насколько темное вино
11.	Hue	Оттенок – цветовой оттенок вина
12.	OD280/OD315 of diluted wines	OD280 / OD315 разбавленных вин
13.	Proline	Пролин – аминокислота (Мг/л)

Таким образом, имеется множество вин V: $|V| = 178$. Вина $v_i \in V$ описываются векторами $v_i = (v_{i1}, v_{i2}, \ldots, v_{i13})$, где v_{ij} – значение j-го показателя i-го вина. На рис. 2 представлены ломаные, характеризующие вина всех трех виноделен

в 13-мерной системе параллельных координат, нормированные согласно формуле

$$(4)\quad \breve{v}_{ij} = \frac{v_{ij} - v_{\min j}}{v_{\max j} - v_{\min j}},$$

где $\breve{v}_{ij}$ – нормированное значение *j*-го показателя *i*-го вина; $v_{\max j}, v_{\min j}$ – соответственно максимальное и минимальное значение *j*-го показателя среди всех вин.

Поскольку конечный результат известен заранее, построим кусочно-линейные функции вин каждой винодельни (рис. 3).

Как видно из рис. 3, группы остаются достаточно зашумленными и разнотипными, что особенно проявляется на данных второй винодельни. Поэтому методы анализа паттернов, основанные на парном сравнении показателей, не могут дать 100% верное разбиение, поскольку матрицы A_i^* (см. (3)) будут отличаться. В связи с этим найдем центроид каждой группы и вычислим расстояние каждого объекта до центроидов.

Согласно формуле (6), исследуемые объекты (в данном случае вина) относим к той группе (винодельне), расстояние до центроида которой минимально. Результаты приведены в таблице 2.

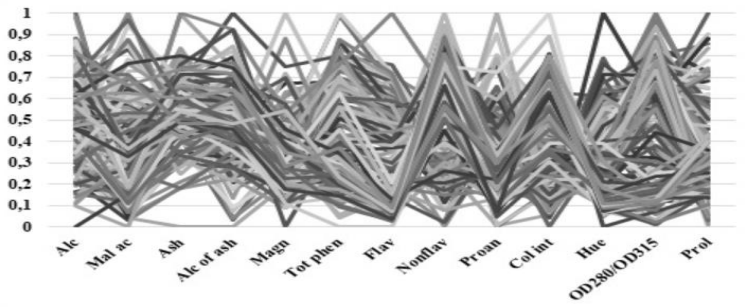

Рис. 2. Ломаные нормированных показателей 178 вин трех виноделен в 13-мерной системе параллельных координат

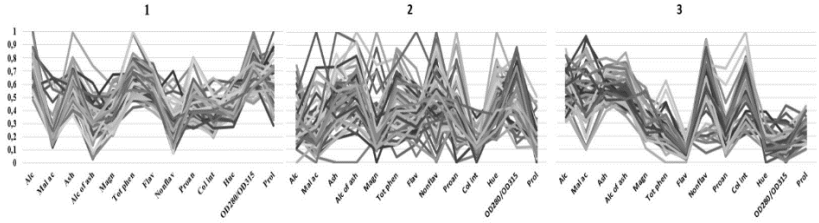

Рис. 3. Корректное распределение вин по винодельням

Таблица 2. Результаты, полученные на базе оценки расстояний до центроидов

Группа	**Ошибки**
1	0
2	5
3	0

Таблица 2 демонстрирует, что процент ошибок относительно невысок и составляет 2,8% от общего числа объектов. Таким образом, использование понятия центроида группы и оценки расстояния до него для формирования конечного разбиения является достаточно эффективным. Далее, продемонстрируем целесообразность использования данной методологии для повышения точности порядково-инвариантной паттерн-кластеризации на примере набора данных Iris Data.

3.2. ПОВЫШЕНИЕ ТОЧНОСТИ ПОРЯДКОВО-ИНВАРИАНТНОЙ ПАТТЕРН-КЛАСТЕРИЗАЦИИ

Результаты применения порядково-инвариантной паттерн-кластеризации к набору данных Iris Data подробно представлены в [17]. Этот набор данных, описанный в [10], является одним из наиболее популярных в машинном обучении и кластерном анализе (к примеру, [10] цитируется 15412 раз по результатам интернет-источника Google Academy на 28.01.2019 г.). Задачей является разбить множество, состоящее из 150 цветов, по трем видам: Iris Setosa, Iris Verginica и Iris Versicolor. При этом каждый цветок характеризуется набором из 4 показателей: длина чашелистика, длина лепестка, ширина чашелистика, ширина лепестка. В конечном итоге мы должны получить разбиение

150 цветов по трем видам, каждый из которых представлен 50 экземплярами. Приведем их представление в 4-мерной системе параллельных координат (рис. 4).

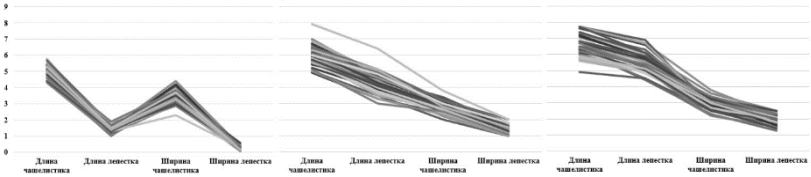

Рис. 4. Слева направо: Iris Setosa, Iris Versicolor, Iris Verginica

Как видно из рис. 4, ход ломаных для цветков Iris Setosa существенно отличается от цветков Iris Verginica и Iris Versicolor, для которых ход ломаных весьма близок. Поэтому ряд авторов используют дополнительный показатель, дающий некоторую оценку площади как произведения длины и ширины лепестка [1]. Для сохранения линейности используемых характеристик нами использован квадратный корень их произведения. Полученный результат представлен в таблице 3.

Таблица 3. Результаты порядково-инвариантной паттерн-кластеризации

Группа	**Количество**	**Ошибка**
Iris Setosa	50	0
Iris Virginica	49	4
Iris Versicolor	51	5

Далее рассчитываются центроиды каждой группы и, основываясь на формуле (6), проводится перегруппировка исследуемых объектов. Полученный результат представлен в таблице 4.

Таблица 4. Результаты корректировки порядково-инвариантной паттерн-кластеризации при использовании центроидов

Группа	**Количество**	**Ошибка**
Iris Setosa	50	0
Iris Virginica	49	3
Iris Versicolor	51	2

Таким образом, рассмотренный пример также демонстрирует повышение точности конечных результатов при использовании центроидов.

4. Заключение

В работе предложена методика повышения точности конечных результатов применения порядково-инвариантной паттерн-кластеризации к различным наборам данных, в том числе при наличии в них погрешностей. Предложено понятие центроида порядково-инвариантного паттерн-кластера, а также доказано утверждение о его принадлежности к тому же порядково-иваринтному паттерн-кластеру, что и объекты, на основе значений показателей которых он образован. Приведены численные расчеты на базе классических наборов данных Wine Data и Iris Data.

Литература

1. АЛЕСКЕРОВ Ф.Т., БЕЛОУСОВА В.Ю., ЕГОРОВА Л.Г., МИРКИН Б.Г. *Анализ паттернов в статике и динамике, часть 1: Обзор литературы и уточнение понятия* // Бизнес-информатика. – 2013. – №3(25). – С. 3–18.
2. АЛЕСКЕРОВ Ф.Т., БЕЛОУСОВА В.Ю., ЕГОРОВА Л.Г., МИРКИН Б.Г. *Анализ паттернов в статике и динамике, часть 2: Примеры применения к анализу социально-экономических процессов* // Бизнес-информатика. – 2013. – №4(26). – С. 3–20.
3. 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.
4. ДЕМИДОВИЧ Б.П., МАРОН И.А. *Основы вычислительной математики*. – М.: Физматгиз, 1963. – 660 с.
5. МИРКИН Б.Г. *Методы кластер-анализа для поддержки принятия решений: обзор* // Высшая школа экономики. Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике». – 2011. – №03. –88 с.

6. МЯЧИН А.Л. *Анализ паттернов: порядково-инвариантная паттерн-кластеризация* // Управление большими системами. – 2016. – №61. – С. 41–59.
7. МЯЧИН А.Л. *Анализ паттернов: диффузионно-инвариантная паттерн-кластеризация* // Проблемы управления. – 2016. – №4. – С. 2–9.
8. САВИЦКАЯ Г.В. *Экономический анализ*. – М.: ИНФРА-М, 2017.
9. DUBES R., JAIN A.K. *Clustering techniques: the user's dilemma* // Pattern Recognition. – 1976. – Vol. 8, No. 4. – P. 247–260.
10. FISHER R.A. *The use of multiple measurements in taxonomic problems* // Annals of Eugenics. – 1936. – No. 7. – P. 179–188.
11. INSELBERG A. *The plane with parallel coordinates* // The visual computer. – 1985. – Vol. 1, No. 2. – P. 69–91.
12. JAIN A.K. *Data clustering: 50 years beyond K-means* // Pattern recognition letters. – 2010. – Vol. 31, №8. – P. 651–666.
13. JIANG M.F., NSENG S.S., SU C.M. *Two-phase clustering process for outliers detection* // Pattern recognition letters. – 2001. – Vol. 22, No. 6-7. – P. 691–700.
14. LIKAS A., VLASSIS N., VERBEEK J.J. *The global k-means clustering algorithm* // Pattern recognition. – 2003. – Vol. 36, No. 2. – P. 451–461.
15. MACQUEEN J. *Some methods for classification and analysis of multivariate observations* // Proc. of the fifth Berkeley symposium on mathematical statistics and probability. – 1967. – Vol. 1, No. 14. – P. 281–297.
16. MIRKIN B.G. *Core Concepts in Data Analysis: Correlation, Summarization, Visualization*. – London: Springer, 2011.
17. MYACHIN A.L. *New methods of pattern analysis in the study of Iris Anderson-Fisher Data* // 6th Int. Conf. on Computers Communications and Control (ICCCC) – 2016. – Oradea: Agora University. – 2016. – P. 97–102.
18. SHAWE-TAYLOR J., CRISTIANINI N. *Kernel methods for pattern analysis*. – Cambridge university press, 2004.
19. https://archive.ics.uci.edu/ml/datasets/wine (дата обращения: 28.01.2019).

DETERMINATION OF CENTROIDS TO INCREASE THE ACCURACY OF ORDINAL-INVARIANT PATTERN CLUSTERING

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

Abstract: The work continues the research of constructing methods for analyzing patterns in parallel coordinates independent of the sequence of input data of the results. The basic operations on objects of ordinal-invariant pattern clusters are described. The assertion that the centroid of an ordinal-invariant pattern cluster belongs to the original cluster is proved, which allows one to estimate the intracluster object - centroid distances in the multidimensional feature space. Examples of revealing the structural similarity of objects in parallel coordinates are given. The main differences between the methods of analysis of patterns and cluster analysis are noted. The methodology of the centroid detection of the ordinal-invariant pattern-cluster is described. An algorithm for combining groups of objects based on their structural similarity, on the one hand, and minimizing intracluster distances, on the other, is proposed, which makes it possible to improve the accuracy of the final results and partially solve the problem of finding similar objects in the presence of error in the original data. The proposed algorithm uses the concept of intracluster distances "object - centroid" and satisfies the following conditions: endogenous determination of the number and composition of the desired groups of objects under study; low (relatively) computational complexity; independence of the original partition from the initial sequence of input data. The work of the proposed algorithm on classical data sets is demonstrated. The results of testing are presented and the clustering accuracy is increased.

Keywords: pattern, pattern analysis, cluster analysis.

УДК 51-74
ББК 32.973.26-018.2
DOI: 10.25728/ubs.2019.78.1

Статья представлена к публикации членом редакционной коллегии Ф.Т. Алескеровым.

Поступила в редакцию 18.06.2018.
Опубликована 31.03.2019.