# АНАЛИЗ ПАТТЕРНОВ В СТАТИКЕ И ДИНАМИКЕ, ЧАСТЬ 1: ОБЗОР ЛИТЕРАТУРЫ И УТОЧНЕНИЕ ПОНЯТИЯ[1]

***Ф.Т. Алескеров,***
*доктор технических наук, руководитель Департамента математики факультета экономики Национального исследовательского университета «Высшая школа экономики», заведующий лабораторией Института проблем управления имени В.А. Трапезникова РАН*

***В.Ю. Белоусова,***
*кандидат экономических наук, заведующий отделом методологии бюджетного планирования Института статистических исследований и экономики знаний, доцент кафедры банковского дела Департамента финансов факультета экономики, с.н.с. Банковского института Национального исследовательского университета «Высшая школа экономики»*

***Л.Г. Егорова,***
*преподаватель Департамента математики факультета экономики Национального исследовательского университета «Высшая школа экономики»*

***Б.Г. Миркин,***
*доктор технических наук, профессор кафедры анализа данных и искусственного интеллекта отделения прикладной математики и информатики факультета бизнес-информатики Национального исследовательского университета «Высшая школа экономики»*

*E-mail: alesk@hse.ru, vbelousova@hse.ru, legorova@hse.ru, bmirkin@hse.ru*
*Адрес: г. Москва, ул. Мясницкая, 20*

***Анализ паттернов – это новая область анализа данных, связанная с поиском взаимосвязей исследуемых объектов, построением их классификации и исследованием развития объектов во времени. В первой части статьи вводится понятие «паттерн» и приводится обзор литературы по методам кластерного анализа и анализа паттернов.***

**Ключевые слова**: паттерны данных, динамический анализ паттернов, кластерный анализ.

[1] Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту от 14.06.2012 г. № 07.514.11.4144 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».
Исследование осуществлено в рамках программы фундаментальных исследований НИУ ВШЭ в 2012 году. Работа авторов была также поддержана рядом лабораторий НИУ ВШЭ: Лаборатория анализа и выбора решений (Алескеров Ф.Т., Белоусова В.Ю., Егорова Л.Г., Миркин Б.Г.), Лаборатория алгоритмов и технологий анализа сетевых структур (Егорова Л.Г., Миркин Б.Г.). Мы благодарны за эту поддержку.

### 1. Введение

Понятие «паттерн» широко используется в задачах машинного обучения и обработки данных. Впервые (более 50 лет назад) оно упоминалось как составная часть словосочетания «Pattern Recognition», которое было переведено на русский язык и закрепилось в нем как «распознавание образов». При этом не очень чёткое английское слово «pattern» было вполне адекватно представлено в русскоязычной литературе в качестве столь же нечёткого «образа». В машинном обучении этот термин обозначает группу многомерных объектов, указанных «учителем» как в чём-то сходных. Одновременно с ним распространение получило столь же нечёткое на первый взгляд понятие «кластер», означающее группу объектов, просто похожих в признаковом пространстве без участия «учителя». Однако за прошедшее время слово «паттерн» проникло в русский язык в качестве рабочего термина ряда специальных дисциплин, таких как «технический анализ» динамики цен, психология или инженерия, где этот термин получил несколько отличающееся от первоначального значение и начал означать что-то типа «шаблона». Именно это более узкое значение понятия «паттерн» мы и хотим использовать для обозначения определённого класса образов или кластеров, постоянно возникающих в анализе данных и машинном обучении, но, кажется, до сих пор не закреплённого в специально выделенном термине.

Под «паттерном» в данной работе понимается такая комбинация определённых, с точностью до погрешности, значений некоторого подмножества признаков, что объекты с этими значениями достаточно сильно отличаются от других объектов. Это понятие можно считать эмпирическим аналогом концептуально-логического понятия «тип» в той же мере, в которой понятие кластера является эмпирическим аналогом концептуально-логического понятия «класс». Нас это понятие интересует, прежде всего, с точки зрения анализа динамики объектов – они естественным образом распадаются на группы, придерживающиеся единого паттерна на временном интервале, – так называемые устойчивые группы поведения и менее устойчивые, чередующие паттерны с течением времени.

Способность находить и использовать паттерны в данных – одна из движущих сил современной науки и промышленности. Интернет-магазины, такие как Amazon.com, рекомендуют своим покупателям продукты, основанные на закономерностях, обнаруженных в базе данных прошлых транзакций. Биологи могут обнаружить гены во многом таким же образом, путем автоматического сравнения последовательности генома со всеми известными последовательностями. Google может получать вебстраницы, которые имеют отношение к запросу с использованием аналогичных идей. Этот список можно продолжить.

Ниже приведен обзор работ, связанных с каждым из трёх упомянутых понятий – паттерн, кластер, анализ динамики многомерных объектов, прежде всего, в их отношении к рассматриваемому понятию. Во второй части будет представлено несколько приложений понятия «паттерн» к анализу социально-экономических явлений.

### 2. Использование понятия «паттерн» в литературе

Понятие паттерн широко используется в финансах, экономике, техническом анализе, прогнозной аналитике, медицине, криминалистике и некоторых других предметных областях. Что характерно, он определяется в этих областях знания по-разному:

- ✦ как сущность явления, имеющего повторяющиеся черты;
- ✦ как свойство повторяющихся компонентов, объединенных общей структурой;
- ✦ как процесс, фиксирующий модель взаимодействия изучаемых объектов, включающего повторения.

Во всех этих предметных областях паттерны данных могут использоваться в смысле выделения групп схожих объектов и изучения их ключевых характеристик с проведением кластерного анализа для разбиения всех объектов выборки на непересекающиеся кластеры для формирования их классификации. В качестве паттерна обозначается некая выявленная закономерность в данных или некая шаблонная структура данных. Например, в [6] термин «паттерн» употребляется для описания некой устойчивой структуры экономических показателей (см. также [1-3]).

Как упоминалось, в англоязычной литературе по искусственному интеллекту и машинному обучению широко используется термин «Pattern recognition» или «Pattern analysis», который переводится на русский язык как распознавание образов. Распознавание образов — это отнесение исходных данных к определенному, не обязательно заранее заданному, классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы данных. Задачами теории


*Рис. 1.* Паттерн технического анализа «треугольник» (рис. взят с обучающего портала по техническому анализу «Биржа и мы», http://exchangeandwe.ru/)

распознавания образов являются, например, распознавание лиц, речи и изображений, штрих-кодов и автомобильных номеров, классификация документов и проч. Алгоритмы распознавания образов зависят от конкретной задачи и типа исходных данных и могут включать в себя методы классификации и кластеризации, а также нейронные сети, марковские модели и байесовские сети [41].

В техническом анализе паттерном называются устойчивые повторяющиеся изменения сочетания цены, объёма или индикаторов рынка за определенный промежуток времени. Анализ паттернов здесь основывается на одной из аксиом технического анализа — «история повторяется» — считается, что повторная конфигурация данных в динамике приводит к аналогичному результату. В русскоязычной литературе паттерны иногда называют «шаблонами» или «фигурами» технического анализа. Также этот термин может упоминаться в словосочетании «trading pattern».

В этом случае паттерн иллюстрируется линией графика цены или индекса, соединяющей соседние цены (цены на момент закрытия торгов, их максимальное и минимальное значения) или значения индекса за определенный промежуток времени. Анали-

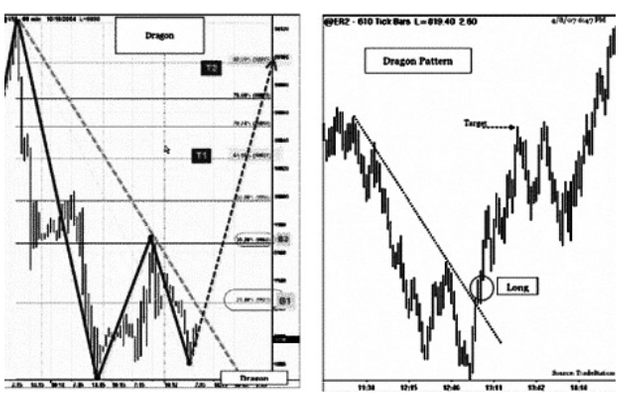


*Рис. 2.* Паттерн технического анализа «дракон» (рис. взят с обучающего портала по торговле на рынке ценных бумаг ForexTurbo, http://www.forexturbo.ru/)

тики рынка пытаются идентифицировать паттерны, чтобы попытаться предугадать ценовые движения рынка в будущем. В техническом анализе выделено много различных видов паттернов: треугольник (*рис. 1*), клин, флаг, фигура «голова и плечи» и т.д.

Иногда встречаются и совсем экзотичные фигуры, например, дракон (*рис. 2*).

Прогнозная аналитика использует методы статистики, интеллектуального анализа данных и теории игр для анализа текущих данных и для составления предсказаний о будущих событиях. В бизнесе прогнозные модели используют паттерны, найденные в исторических данных, чтобы идентифицировать риски и возможности. Модели фиксируют связи среди многих факторов, чтобы сделать возможной оценку рисков или потенциала, связанного с конкретным набором условий, при руководстве принятием решений о возможных сделках [36].

Прогнозная аналитика применяется в актуарных расчётах, финансах, страховании, телекоммуникациях, розничной торговле, туризме, здравоохранении, фармацевтике и других областях. Одним из наиболее известных приложений в финансах является кредитный скоринг, модели которого обрабатывают кредитную историю потенциального заемщика, информацию о его текущих займах и финансовом благополучии, потребительские данные и т.д. На основе этих результатов потенциальные заемщики упорядочиваются по вероятности качества обслуживания ими планируемых обязательств по кредиту, включающих выплаты по основному долгу и начисленным процентам за пользование заемными средствами банка в установленные кредитным договором сроки.

Другим прикладным примером является использование паттернов в розничной торговле, когда в результате анализа данных о покупках человека в розничной сети ему могут быть предложены рекламные акции или купоны на скидки по товарам, которые ему вероятнее всего потребуются. Например, если Вы купили купальник в апреле, то в мае Вам предложат крем для загара. Для реализации такого поведенческого «таргетинга» необходимо выявление подобных зависимостей в покупках, т.е. анализ паттернов, называемых в майнинге данных «ассоциациями».

В медицине термин «паттерн» встречается в анализе кардиограмм, энцефалограмм (ЭЭГ) и проч. [9], когда под ним понимают:

- ✦ или последовательность определенных форм колебаний биопотенциалов, повторяющуюся в

одном или нескольких отведениях (то есть парах электродов, с которыми производится регистрация биопотенциалов) при одинаковых состояниях и условиях [27];

- ✦ или же картину, отражающую особенности распределения различных компонентов ЭЭГ по всем отведениям в целом.

Иногда этот термин используется для обозначения последовательности нервных импульсов, имеющей определённое информационное значение [31], например, «паттерны боли при биомеханических нарушениях суставов краниовертебрального перехода и шейного отдела позвоночника» [34] или «паттерны двигательных и чувствительных расстройств при патологии нервных структур в дистальных отделах верхней конечности» [35].

Смысл термина «паттерн» зависит от области знаний, в которой он используется и иногда употребляется для обозначения явлений и процессов, никак не связанных с анализом статистических данных, описанным в разделе 1. К таким областям можно отнести паттерны в программировании, психологии, музыке, педагогике и т.д.

В разработке программного обеспечения шаблон проектирования или паттерн (англ. *design pattern*) представляет собой повторяемую архитектурную конструкцию решения определенной локальной проблемы проектирования в рамках некоторого часто возникающего контекста. Обычно шаблон не является законченным образцом, который может быть прямо преобразован в код; это лишь пример решения задачи, который можно использовать в различных ситуациях. Объектно-ориентированные паттерны показывают отношения и взаимодействия между классами или объектами без определения того, какие конечные классы или объекты приложения будут использоваться [26].

В психологии это слово чаще всего используется в контексте паттернов поведения или гипнотических паттернов [42]. Паттерн поведения – это набор стереотипных реакций или последовательность действий индивида. Каждый человек имеет типовые способы взаимодействия с окружающим миром – устойчивые модели поведения, которые он предпочитает использовать при общении с другими людьми. Кто-то чаще выбирает паттерны уверенного поведения, кто-то – саморазрушающие или манипулятивные паттерны. Чтобы с высокой точностью предсказать, как поведет себя человек в той или иной ситуации, иногда достаточно определить основные паттерны поведения, которых он придерживается в своей жизни в целом и в подобных ситуациях в частности. Гипнотический паттерн – это текст, который использует гипнолог для наведения транса и последующей работы с индивидом, находящимся в трансе. Некоторые гипнологи могут оказывать подобное влияние без введения в глубокий транс, когда человек искренне считает, что находится в сознании, но его выбор в какой-то мере заведомо определен используемым гипнотическим паттерном.

В трекерной, т.е. создаваемой на компьютере, музыке паттерн — это таблица, определяющая порядок и режим воспроизведения семплов на нескольких каналах за некоторый промежуток времени или группа одновременно воспроизводимых каналов, представляющая полноценную часть музыкальной композиции. Внутри композиции паттерны могут повторяться; это делает возможным относительно быстрое заполнение общей структуры произведения.

В зарубежной литературе по педагогике используют педагогические паттерны (Pedagogical Patterns) [16], которые формируются для передачи передового опыта в конкретной области. С их помощью пытаются осуществить передачу экспертных знаний о практике преподавания и обучения. Цель состоит в обучении каждого нового преподавателя стандартному набору педагогических знаний и навыков. Паттерн включает, например, мотивацию студентов, выбор материалов и последовательность их донесения, правила оценки студентов и тому подобное.

Используется этот термин также в криминалистике, в основном, как синоним слов «шаблон» и «структура», например, «паттерн образца ДНК подозреваемого», «паттерн роста опухоли», «паттерн кровоснабжения папиллярной карциномы» или «паттерн сосудистой системы сетчатки», а словосочетание «papillary patterns» означает папиллярные узоры в дактилоскопии.

Как видно, понятие «паттерн» широко используется в самых разных областях науки и практики. Несмотря на кажущуюся специфичность, все приведенные примеры его использования в общем и целом соответствуют тому пониманию, которое было нами сформулировано во введении.

### 3. Методы кластер-анализа в литературе

В соответствии с [33], методы кластер-анализа разделим по типу формируемых кластерных структур:

***1) Методы, выделяющие отдельные кластеры:***

- Кластеры по Апресяну.
- Метод ФОРЭЛЬ.

- Аппроксимационные кластеры.
- Монотонные кластеры.
- Логические таксоны.

***2) Методы построения разбиений:***

- Метод кластеризации K-средних (K-means).
- Имитирующие природу алгоритмы: генетические, эволюционные алгоритмы и алгоритм роя частиц.
- Методы нечеткой кластеризации.
- Метод аномальных кластеров.
- Модернизированные методы K-средних: использование весов для переменных, вероятностная формулировка и EM-алгоритм, самоорганизующиеся карты Кохонена.
- Методы построения разбиений по матрицам связи, включая так называемые иерархические, спектральные, локальные и графо-теоретические подходы.

***3) Методы построения иерархий:***

- агломеративные методы,
- дивизимные методы.

***4) Бикластерный анализ.***

Рассмотрим перечисленные подходы подробнее.

### 3.1. Методы, выделяющие отдельные кластеры

#### 3.1.1. Кластеры по Апресяну [4]

Подмножество объектов выборки $S$ называется А-кластером, если для любых различающихся объектов $x_i$, $x_j$, $x_k \in X$ таких, что объекты $x_i$, $x_j$ принадлежат А-кластеру $S$, а объект $x_k$ не принадлежит $S$, расстояние между объектами кластера $x_i$ и $x_j$ меньше расстояния между первым объектом $x_i$ и объектом $x_k$, не принадлежащим А-кластеру ($d_{ij} \le d_{ik}$, где $D=(d_{ij})$ – это заданная матрица коэффициентов различия или расстояния между объектами). Нетрудно видеть, что множество всех А-кластеров для данной матрицы $D$ образует иерархию, т.е. если два А-кластера пересекаются, то один из них является частью второго.

#### 3.1.2. Метод Форэль [28]

Этот метод кластеризации основан на идее последовательного объединения элементов в кластер в областях их наибольшего сгущения. Идея метода заключается в следующем: на первом шаге выбирается «центральный» объект $g$ из выборки, далее кластером объявляется множество всех объектов, находящихся от выбранного объекта $g$ на расстоянии меньше заданного заранее расстояния $R$ (то есть все объекты, находящиеся внутри шара радиуса $R$ с центром в точке, соответствующей выбранному объекту $g$), и находится центр тяжести $S$ этого кластера. Найденный центр тяжести объявляется новым центром и кластер переопределяется уже вокруг нового центра. Так продолжается до тех пор, пока центр не стабилизируется. При необходимости можно продолжить процедуру кластеризации, предварительно удалив уже найденный кластер.

На реальных данных этот алгоритм производит несколько крупных кластеров и множество мелких, что в контексте «частичного» кластер-анализа можно отнести к числу преимуществ.

#### 3.1.3. Аппроксимационные кластеры [20]

В этом методе искомый кластер $S$ представляется бинарным вектором принадлежности $s = (s_i)$, где $s_i = 1$, если объект $x_i \in S$, и $s_i = 0$ – в противном случае. Затем вводится положительный уровень интенсивности $\lambda$, так чтобы суммарная разница квадратов $L(S) = \Sigma_{i,j \in X} (b_{ij} - \lambda s_i s_j)^2$ была минимальной, где $B = (b_{ij})$ заданная матрица связи между объектами. Кластер $S$ формируется последовательным присоединением «оптимальных» объектов до тех пор, пока критерий $L(S)$ не перестает уменьшаться.

К преимуществам метода относятся:

- ✧ математически доказанное свойство «тесноты» получаемого кластера $S$: для всякого объекта из кластера $S$ средняя связь с $S$ превышает $\lambda/2$, а для всякого объекта $x_i \notin S$ средняя связь с $S$ меньше, чем $\lambda/2$;
- ✧ декомпозиция разброса связей (т.е. суммы их квадратов) на части, объясненные и не объясненные кластерной структурой;
- ✧ возможность получения пересекающихся решений.

#### 3.1.4. Монотонные кластеры [29]

Рассмотрим монотонную функцию сходства $f(i,S)$ между всеми объектами $x_i \in X$ и подмножествами $S \subset X$. Монотонность подразумевает, что либо $f(i,S) \ge f(i,S \cup T)$ для всех $S$ и $T$ (сходство между объектом и кластером $S$ больше, чем между объектом и объединенным кластером $S \cup T$), либо $f(i,S) \le f(i,S \cup T)$ для всех $S$ и $T$. Например, при заданной неотрицательной матрице связи $A = (a_{ij})$ можно определить $f(i,S)$ как $min_{i \in S} a_{ij}$ или $\Sigma_{i \in S} a_{ij}$, так чтобы $f$ была монотонна по $S$ в нужную сторону. Для определенности рассмотрим монотонно возрастающую $f(i, S)$ и определим функцию множеств $F(S) = min_{i \in S} f(i, S)$, характеризующую «наислабейшее звено» в $S$. Такие функции $F(S)$ можно максимизировать, подбирая

на каждом шаге наилучшего кандидата, что ведет к «оптимальной» последовательности объектов, которая определяет не только «ядро» – кластер $S$, максимизирующий $F(S)$, как фрагмент этой последовательности, но и совокупность его «оболочек» – объемлющих фрагментов. Монотонные кластеры использовались для анализа организационных систем.

#### 3.1.5. Логические таксоны [30]

Любой предикат (математическое высказывание, в котором есть хотя бы одна переменная), сформированный из признаков, например, "$y_1$=3 и $y_2/y_3$ >5", где $y_i$-один из признаков, описывающих объекты выборки, определяет множество объектов $S$, удовлетворяющих ему. Доля $f$ множества $S$ в $I$ – это наблюдаемая частота истинности предиката. С другой стороны, каждая составляющая предиката, связанная с отдельным признаком (в нашем примере – "$y_1$=3" и "$y_2/y_3$ >5"), имеет свою частоту: $f_1$ и $f_2$. Эти индивидуальные частоты легко скомбинировать так, чтобы определить ожидаемую частоту предиката $ef$ по правилам вероятности для логических операций – например, $ef=f_1*f_2$. Оптимизационный критерий – это разница между $f$ и $ef$. Чем больше эта разница, тем лучше. Вероятно, этот подход можно применить и для непосредственного построения паттернов как логических таксонов путем формирования предикатов, имеющих вид конъюнкций значений одних и тех же признаков.

### 3.2. Методы построения разбиений

#### 3.2.1. Метод кластеризации K-средних (k-means)

Это наиболее популярный метод кластеризации, разбивающий все объекты выборки на заранее заданное число кластеров $k$, представленных их центральными точками (центроидами) и списками попавших в них объектов. В качестве критерия эффективности выбирается суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Начальные центры кластеров выбираются случайным образом, затем осуществляется последовательность итераций, состоящая из двух шагов. На первом шаге происходит обновление кластера: каждый объект приписывается одному из центров по правилу минимального расстояния, на втором – обновление центра кластера: для каждого кластера вычисляется центр тяжести, который объявляется его новым центром. Все повторяется до тех пор, пока на какой-то итерации центры масс не останутся прежними. Среди недостатков этого метода самыми серьезным является то, что он обычно сходится к локальному, а не глобальному минимуму. Это приводит к зависимости результата от выбора начальных центров кластеров, поэтому существует множество модернизированных методов $K$-средних; основные рассмотрены ниже.

#### 3.2.2. Имитирующие природу алгоритмы

Имитирующие природу алгоритмы: генетические [17], эволюционные алгоритмы [5], алгоритм роя частиц, муравьиной колонии и пр.

Генетическими и эволюционными называются алгоритмы кластерного анализа, использующие механизмы, напоминающие биологическую эволюцию. Они решают оптимизационные задачи с использованием моделей естественной эволюции, таких как наследование, мутации, отбор и кроссовер. Несколько другой биологический процесс, отыскания пищи роем пчел, имитируется в методе роя частиц. Рой движется в случайном направлении, при этом запоминая наилучшие места из уже посещенных.

Среди недостатков этих методов можно выделить невозможность оценки качества получаемых решений и, главное, плохую масштабируемость для задач с большой размерностью данных, поскольку методы, имитирующие природу, требуют довольно много вычислительного времени.

#### 3.2.3. Методы нечеткой кластеризации [25]

Нечеткий кластер задается функцией принадлежности $z=(z_i)$, $x_i \in X$, так что величины принадлежности $z_i$ $(0 \le z_i \le 1)$ интерпретируются как степени принадлежности объектов кластеру (для четких кластеров значения $z_i$ могут быть только 1 или 0).

Нечеткое разбиение объектов на кластеры – это разбиение по кластерам с центрами $c_k=(c_{k1}, ..., c_{kv}, ..., c_{kV})$ в пространстве признаков и $K$ векторов принадлежности $z_k=(z_{1k},..., z_{ik},..., z_{Nk})$, $0 \le z_{ik} \le 1$, таких что $\Sigma_k z_{ik} = 1$ для всех $x_i \in X$.

#### 3.2.4. Метод аномальных кластеров [19]

Этот метод заключается в применении стратегии последовательного исчерпания данных, в данном случае – по одному кластеру. Начальный шаг состоит в том, что точка, наиболее удаленная от центра тяжести всех точек, берется в качестве центра аномального кластера, затем формируется сам кластер, в который входят все точки, расстояние от которых до центра тяжести меньше, чем расстоя-

ние до центра аномального кластера. Далее центр аномального кластера заменяется на центр тяжести кластера и происходит следующая итерация.

### 3.2.5. Модернизированные методы *K*-средних

Модернизированные методы *K*-средних: введением весов для переменных [12], EM-алгоритм [7] и пр.

Модернизированный метод *K*-средних с введением весов для переменных похож на метод *K*-средних нечеткого кластер-анализа, только веса здесь отражают не в функции принадлежности объектов, а в признаках в кластере. *EM*-алгоритм – это метод кластерного анализа, требующий априорного знания модели порождения данных. Согласно модели, наблюдаемые данные – независимая случайная выборка из распределения с функцией плотности вида смеси распределений, параметры которой неизвестны. Каждая итерация алгоритма состоит из двух шагов: на первом (*E*-шаге) вычисляется ожидаемое значение функции правдоподобия, на втором (*M*-шаге) вычисляется оценка максимального правдоподобия; таким образом увеличивается величина ожидаемого правдоподобия, вычисляемая на *E*-шаге. Полученные значения параметров используются для *E*-шага на следующей итерации. Особенно удобно это осуществлять при нормальных распределениях, моделирующих отдельные кластеры. *EM*-алгоритм или его модификации используются практически во всех разработках, основанных на модели распределения. К его недостаткам относятся:

1. трудность инициализации,

2. большое число переменных, оценка которых возможна только при большом числе объектов в кластере;

3. сходимость к локальному оптимуму.

### 3.2.6. Методы построения разбиений по матрицам связи

При заданной матрице сходства $A=(a_{ij})$ между объектами множества $X$ попробуем разбить $X$ на две части – $S_1$ and $S_2$ – таким образом, чтобы сходство между $S_1$ и $S_2$ было минимально, тогда как сходство внутри – максимально. Методы получения оптимальных разбиений бывают: иерархические, спектральные, локальные и графо-теоретические.

### 3.2.7. Иерархические алгоритмы [15]

Такие методы работают либо «агломеративно», склеивая на каждом шаге два наиболее близких кластера, начиная с тривиального разбиения на одноэлементные кластеры, либо «дивизимно» – разбивая на каждом шаге какой-либо кластер, начиная с универсального кластера, состоящего из всех объектов.

### 3.2.8. Спектральный подход [18]

Этот метод решает задачу кластерного анализа в терминах так называемого Лапласиана $L(A)$ как задачу минимизации отношения Рэлея, известного в теории собственных чисел и векторов, при котором кластеры получаются путем огрубления собственных векторов Лапласовой матрицы.

Принцип спектрального кластер-анализа в данной версии: найти собственный вектор, соответствующий минимальному ненулевому собственному значению, после чего определить кластеры в соответствии со знаком компонент: индексы положительных элементов – в один класс, а индексы отрицательных элементов – в другой. Иными словами, этот собственный вектор аппроксимируется вектором из 1 и -1, так что положительные компоненты заменяются на 1, а отрицательные – на -1, после чего определяется разбиение $\{S_1, S_2\}$ следующим образом: $S_1$ как множество объектов, соответствующих 1, а $S_2$ – соответствующих -1.

### 3.2.9. Локальные алгоритмы [20]

Такие методы на каждом шаге рассматривают некоторое разбиение $\{S\}$ и производят локальное его преобразование в сторону улучшения значения критерия. Из всех возможных локальных преобразований чаще всего рассматривается только перенос одного объекта из класса в класс. Для этого отыскивается объект, на котором приращение критерия максимизируется, и если оно положительно – перенос производится. Если нет – разбиение объявляется окончательным результатом.

### 3.2.10. Графо-теоретический подход [20]

Это алгоритмы кластерного анализа данных, визуализация которых обеспечивается с помощью графов. Объекты представляются как вершины, или узлы графа, а связи между объектами — как дуги или рёбра. Для разных методов виды графов могут различаться направленностью, ограничениями на количество связей и дополнительными данными о вершинах или рёбрах.

Как и большинство визуальных способов представления зависимостей, графы быстро теряют наглядность при увеличении числа объектов.

### 3.3. Построение иерархий

Здесь существует два вида методов – агломеративные и дивизимные методы.

#### 3.3.1. Агломеративные методы [15,24]

Агломеративными называются методы, в которых вычисления начинаются с тривиальных – одиночных – кластеров, и продолжаются итерациями, каждая из которых состоит в объединении двух наиближайших кластеров с последующим определением расстояния между вновь построенным кластером и остальными. В качестве примеров агломеративных методов можно назвать метод ближнего (дальнего) соседа и метод Уорда.

Агломеративные методы обычно используют все попарные расстояния для отыскания минимума, что может существенно осложнить вычисления на больших данных.

#### 3.3.2. Дивизимные методы [20]

Эти методы создают кластеры путем разбиения больших кластеров на меньшие части, начиная со всего множества. Удобство таких методов состоит в том, что процесс деления можно в любой момент остановить. Три наиболее популярных дивизимных метода – бисекция *к*-средних, бисекция главной компоненты и концептуальный кластер-анализ. Каждый шаг концептуального кластер-анализа состоит в разбиении какого-либо кластера на две части по одному из имеющихся признаков так, что в одну часть включаются объекты, на которых значение этого признака меньше, чем некоторое оптимально подбираемое значение $a$, а в другую – больше $a$. Поэтому каждый из финальных кластеров характеризует некоторый паттерн – тот, который определяется конъюнкцией предикатов, ведущих к этому кластеру от корня построенной иерархии.

### 3.4. Бикластерный анализ [8, 32]

При заданной матрице признаковых описаний объектов $Y=(y_{iv})$, где $i \in I$ обозначает объект и $v \in V$ – признак, $I$ – множество строк, $V$ – множество столбцов, бикластером называется пара множеств $(S, T)$, где $S \subset I$ – подмножество строк, а $T \subset V$ – подмножество столбцов, такая что подматрица $Y(S,T)=(y_{iv})$, где $i \in S$ и $v \in T$, имеет какую-либо особенность, обычно – одинаковые строчки или даже одинаковые значения, свидетельствующие об определенной связи между $S$ и $T$. Однако иногда рассматривают и более сложные зависимости – например, рост значения, пропорциональный номеру столбца.

Из перечисленных методов кластер-анализа наиболее соответствуют задаче построения паттернов следующие три: (а) построение логического таксона, (б) метод *K*-средних и (в) концептуальный кластер-анализ. Центроиды кластеров по методу *K*-средних, как комбинации определённых значений всех признаков, играют роль паттернов. К сожалению, такие решения не всегда удовлетворительны, так как используют все имеющиеся признаки и недостаточно заботятся о том, чтобы паттерны были действительно различны. Последнее определяется требованием, чтобы кластеры образовывали разбиение, т.е. покрывали все объекты, какими бы неадекватными они не были. Напротив, каждый кластер, формируемый методом концептуального кластер-анализа, определяется комбинацией небольшого числа признаков. Однако эти комбинации имеют тот недостаток, что могут включать совершенно различные признаки, тем самым исключая возможность сравнения паттернов друг с другом. В этом плане наибольшие перспективы может иметь метод логического таксона, но применимость этого метода к реальным данным остается неясной, поскольку за 30 лет своего существования, метод никем, кроме авторов [30], не использовался.

## 4. Исследование многомерных динамических процессов

Существует большое количество исследований по анализу динамики развития социально-экономических объектов. Как правило, такие работы связаны с использованием методов кластерного анализа и распознавания сигналов и изображений. В качестве примеров таких работ, «близких по духу» динамическому анализу паттернов, в данном разделе приводятся краткие описания зарубежных научных теоретических и практических работ, в том числе несколько широкомасштабных проектов Европейской комиссии.

Исследования в области динамического анализа паттернов социально-экономических объектов, проведенные отечественными авторами, рассмотрены во второй части обзора.

В [14] проводится эмпирический анализ закономерностей в эволюции кривых процентных ставок, или кривых доходности – interest rate curves (IRC). Кривые IRC рассматриваются как объекты (кривые) в многомерном пространстве, и ставится проблема изучения сходств и различий между ними. Это типичная проблема кластеризации и классификации в машинном обучении. В качестве примера рассмотрены данные по ежедневной доходности швей-

царского франка (CHF). Информация о динамике доходности была доступна для разных интервалов времени (ежедневная, еженедельная, ежемесячная) с разными сроками. В этом исследовании анализировались данные по дневным ставкам лондонского межбанковского рынка (LIBOR) со сроком 1 неделя, 1, 2, 3, 6, 9 месяцев и ставки по свопам (SWAP) со сроком 1, 2, 3, 4, 5, 7 и 10 лет (*рис. 3*).

Было получено 4 основных кластера, паттерны которых представлены на *рис. 4*. Эти паттерны мало различаются – например, паттерн кластера 4 очень похож на паттерны кластеров 1 и 3.

Однако, на самом деле, соответствующие паттерны наблюдаются в разные моменты функционирования системы. Учет этого обстоятельства делает паттерны хорошо различимыми, что показано на *рис. 5*, где выделены временные периоды действия этих паттернов.

Интересной находкой является выявление нескольких типичных видов поведения кривых, отраженных кластером с низким уровнем ставок, кластером с высоким показателем уровня ставок и переходными между ними кластерами. Информация о динамике процентной ставки используется для экономических и финансовых решений и управления рисками. Такой анализ может помочь в прогнозировании кривых доходности, оценке стоимости финансовых инструментов и управлении финансовыми рисками.

В [23], как и во множестве других подобных работ, задача состоит в выявлении кластеров генов, которые демонстрируют одновременную экспрессию, то есть процесс, в ходе которого наследственная информация от гена преобразуется в функциональный продукт — РНК или белок. Модель, представленная в [23], позволяет учитывать предварительную инфор-

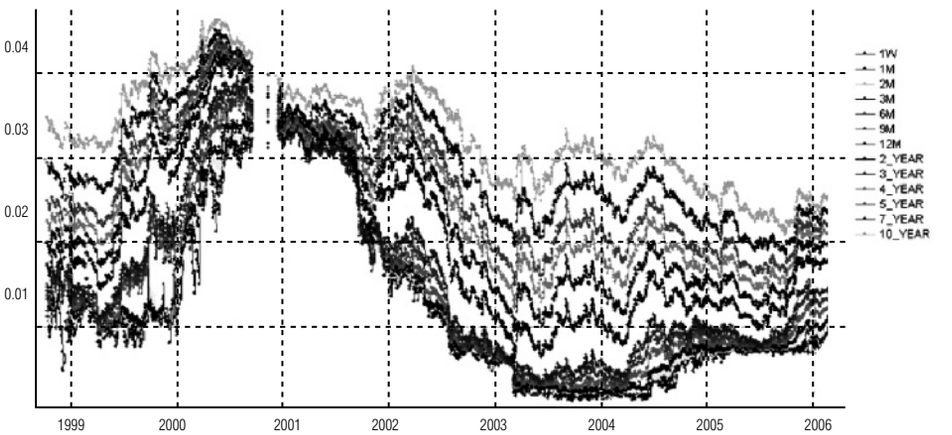


*Рис. 3.* Эволюция доходности швейцарского франка

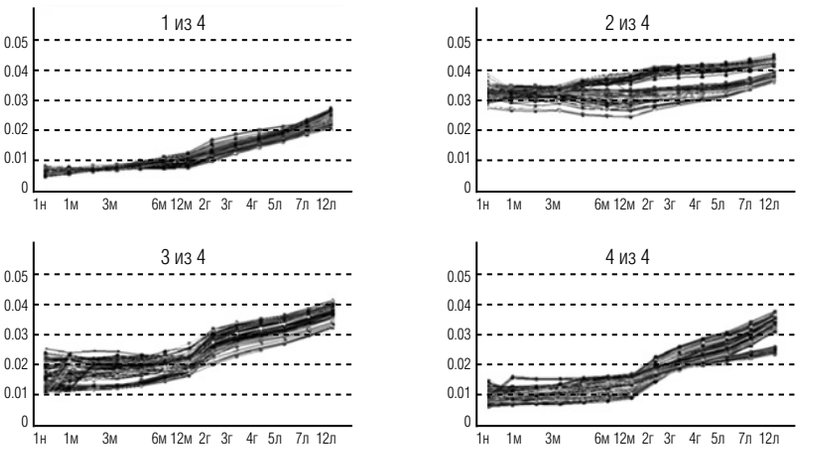


*Рис. 4.* Примеры IR кривых для каждого из 4 кластеров (на координатных осях – срочность и доходность)

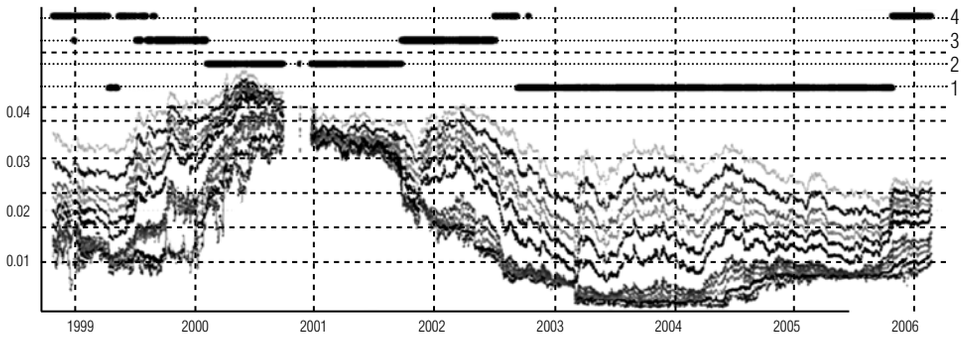


*Рис. 5.* Результаты классификации, соотнесенные с исходными данными. Верхние жирные линии и точки соответствуют найденным 4 кластерам (правая ось)

мацию и предполагает вероятностное распределение по кластерам. Эта модель учитывает то, что каждый ген следует по одной из нескольких основных траекторий, форма и число которых зависит от экспериментальных условий. Был использован начальный фильтр на основе факторного анализа для снижения размерности данных. Результаты иллюстрируются с помощью двух экспериментов, осуществленных на дрожжах. Показатель эффективности данной модели – это апостериорная вероятность совместного поведения генов в процессе экспрессии, которая соответствует вероятности того, что два или более гена попадут в одну группу.

Для результатов первого эксперимента по спорообразованию дрожжей были использованы данные шести временных интервалов. Были оценены уровни совместной экспрессии генов в этом процессе и получены 12 «характеристических кривых» генной экспрессии по всему временному промежутку. На рис. 6 отмечены все кластеры, сплошная темная линия соответствует усредненным значениям, выделенным для описания каждого кластера.

Было обнаружено, что число кластеров, а иногда и вероятность совместной экспрессии объектов, могут быть весьма чувствительны к их априорному распределению.

Существует ряд совместных проектов, выполняемых российскими учеными совместно с европейскими коллегами, финансируемых Европейской комиссией, точнее ее подразделением CORDIS (Community Research and Development Information Service).

Например, проект INTAS 2004-77-7067 «Medical image mining: Theoretical foundation and technological aspects» направлен на развитие теоретических основ специализированного математического аппарата [37]. В нем также предложена реализация различных технологий и методологий для автоматического распознавания и анализа медицинских изображений.

Объективный анализ биомедицинских (цитологических и гистологических) изображений были предметом исследований в течение многих лет. Анализ медицинских изображений необходим для извлечения информации о состоянии больного и обеспечения правильного диагноза его заболевания. Одной из самых сложных задач в области анализа биомедицинских изображений является автоматизированное распознавание объектов из изображений и последующая классификация. Критическая проблема для автоматизации анализа биомедицинских изображений заключается в создании и исследовании алгебраических структур для представления изображения в качестве объектов анализа и распознавания, моделей преобразования изображения в качестве инструмента для эффективного синтеза и реализации основных процедур обработки и анализа изображений.

Практическая цель исследования – разработка:

✧ инструментария для решения основных задач автоматизированного анализа цветных биомедицинских изображений с использованием новых методов алгебры изображения, эффективных алгоритмов соответствия изображения, сегментации, нейронных сетей и автоматического преобразования растровых в векторные изображения;

✧ новых информационных технологий для морфологического анализа лимфоидных клеточных ядер у больных с опухолями кровеносной системы на основе комбинированного использования методов распознавания паттернов и методов анализа изображений.

Для оценки и проверки разработанных подходов и методов использовались медицинские базы данных изображений, содержащие более 12000 изображений. Кроме того, в целях разработки правил принятия решений для диагностики заболеваний включены морфометрические оценки биологиче-

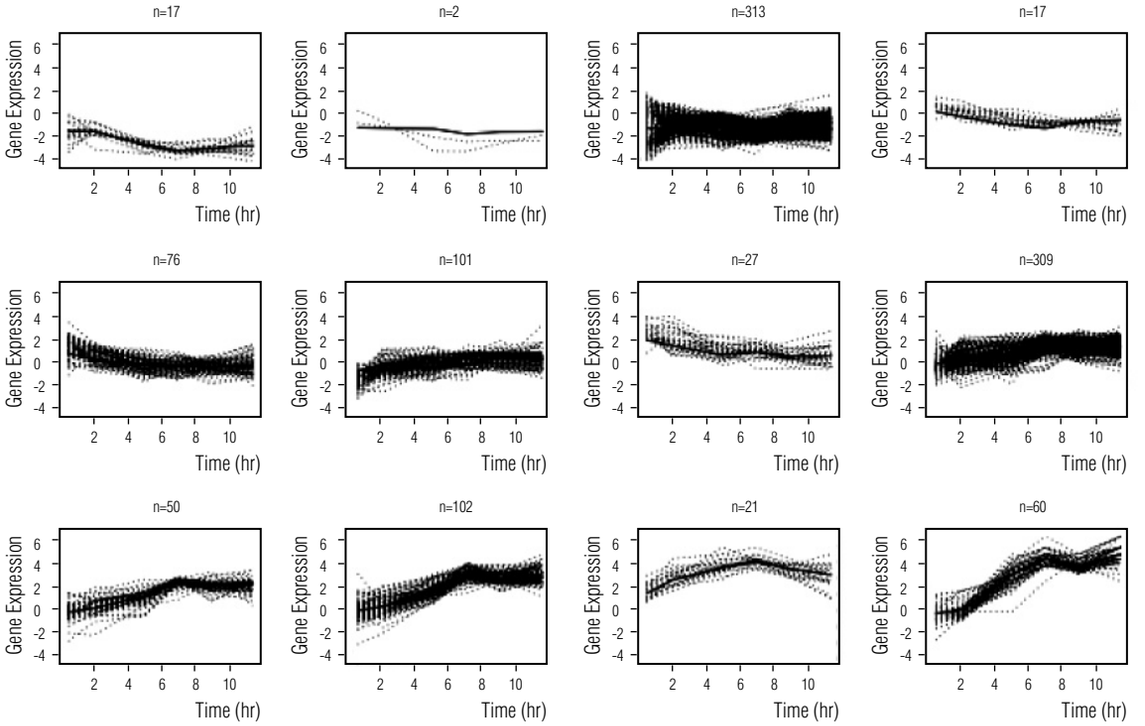


*Рис. 6.* Кластеры генной экспрессии

ских объектов в виде количественных параметров, характеризующих изменения клеточных ядер, волокон и тканей. В качестве иллюстрации эффективности полученных теоретических методов авторы проводили анализ изображений клеток крови и лимфатической ткани.

Еще один проект по смежной тематике – INTAS 2004-77-7347 «Principles of dissimilarity-based pattern recognition in signals, symbolic sequences and images (PRINCESS)» [38]. Научной целью проекта PRINCESS является разработка теоретической основы для алгоритмических технологий распознавания образов и паттернов данных. Особое внимание уделяется распознаванию объектов с сигналами, символьными последовательностями и изображениями. В нем применяется традиционный подход кластерного анализа, заключающийся в выделении такого рода информативных признаков в объектах, которые представляют каждый объект в векторном пространстве признаков, и использовании эвристических способов для оценки различия между сигналами, символьными последовательностями или изображениями, используя различные функции, которые обладают всеми свойствами метрики.

Задача проекта заключается в создании общей теоретической и методологической базы в следующих областях:

- ✦ нахождения эмпирических закономерностей и паттернов данных в наборах сигналов, символьных последовательностей и изображений,
- ✦ структурное выделение паттернов в данных на основе различных метрик для описания сходства/различия объектов,
- ✦ распознание образов с помощью классификаторов.

Практическая часть исследования состояла в применении рассмотренных методов к задаче распознавания человеческого лица на изображении, использования структурированной модели для 3D-реконструкции с изображения 2D и для задачи распознавания математических формул в печатных документах.

Два других проекта Европейской комиссии FP6-IST, а именно «Pattern analysis, statistical modelling and computational Learning» [39] и FP7-ICT «Pattern analysis, statistical modelling and computational Learning 2» [40], легли в основу создания полномасштабного общеевропейского распределенного института Паскаля (сайт института http://www.pascal-network.org), развивающего методы распознавания образов, статистического моделирования и машинного обучения.

Среди практических задач, которые рассматривают участники института Паскаля, можно выделить:

- ■ Компьютерное зрение
  - ✧ Распознавание человека/объекта
  - ✧ Слежение за человеком/объектом

■ Обработку и исследование текстов на естественных языках
- ✧ Машины поисковых запросов
- ✧ Переводы с разных языков
- ✧ Фильтрация, в том числе от спама
- ✧ Адаптивное управление веб-контентом

■ Машинное обучение
- ✧ Задачи классификации
- ✧ Задача кластерного анализа
- ✧ Задачи регрессионного анализа
- ✧ Задачи построения ранжирования

■ Анализ данных
- ✧ Проблема баз данных большой размерности
- ✧ Обнаружение закономерностей (Knowledge discovery)

■ Дополнительные задачи
- ✧ Распознавание речи
- ✧ Биоинформатика
- ✧ Портфельное инвестирование и риск-менеджмент
- ✧ Теория управления

Проект Паскаль активно развивается, в институте существуют программа поддержки и финансирования исследовательских программ «The Harvest Programme», проводятся соревнования (Challenges) по упомянутым темам (http://pascallin.ecs.soton.ac.uk/Challenges/), библиотека насчитывает свыше 6 тысяч публикаций и научных работ (http://eprints.pascal-network.org/view/year/).

### 5. Уточнение понятия паттерн и его использование для анализа динамики

Возможны различные уточнения понятия паттерн, связанные с вариативностью способов описания того, какие погрешности допускаются как в значениях признаков, так и количествах объектов, не покрываемых найденными паттернами. В частности, имеются три эквивалентных математических способа представления паттернов. Эти способы апеллируют к различным когнитивным подсистемам: один – к образной, другой – к логической, а третий – к геометрической.

Для примера рассмотрим одно из самых популярных множеств данных в машинном обучении и распознавании образов, так называемые Ирисы. Это множество задаётся таблицей 150×4 (см. http://archive.ics.uci.edu/ml/datasets/Iris), характеризующих 4 измерения 150 экземпляров цветка ирис, представленных в статье Р. Фишера по дискриминантному анализу [11], которая рассматривается как самая первая основополагающая статья по распознаванию образов.

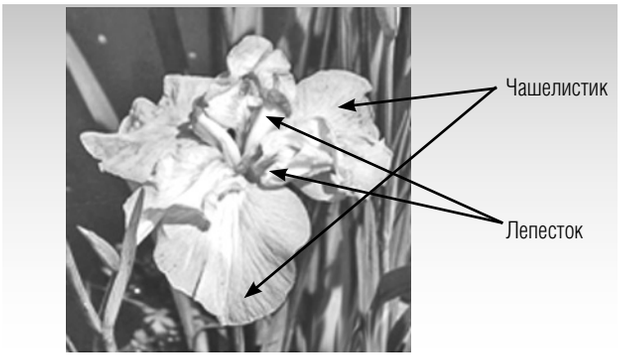


*Рис. 7.* Чашелистик (*sepal*) и лепесток (*petal*) в цветке ириса

Цветы ириса в этих данных относятся к трем видам (таксонам): I *Iris setosa* (диплоид), II *Iris versicolor* (тетраплоид) и III *Iris virginica* (гексаплоид). Два признака таблицы относятся к измерениям длины и ширины чашелистика (w1, w2), а два – к таким же измерениям лепестков (w3, w4). Эти элементы на *рис. 7* показаны стрелками.

Подсчитаем средние значения признаков на таксонах, нарисуем для каждого из них вертикальную ось, отложим средние на этих осях и соединим отрезками прямых средние, относящиеся к одному и тому же таксону (*рис. 8*). Такой способ представления многомерных объектов в литературе называют «параллельными координатами» (см., например, [10, 13]). Полученные ломаные в определённой степени характеризуют паттерны данных таксонов. Почему в определённой степени? Потому что здесь они проявляются на уровне средних, и не очень понятно, насколько эти паттерны характерны для всех объектов таксонов. Как видим, два паттерна более или менее похожи, тогда как паттерн, показанный сплошной линией (первый таксон), обнаруживает несколько иную структуру средних. Кроме того, совершенно очевидно, что по признаку w2 паттерны практически не отличаются, так что этот признак следует исключить, чтобы паттерны действительно отличались друг от друга по всем признакам, участвующим в паттернах.

*Рис. 9*, выполненный с помощью операции parallelcoords системы МатЛаб, представляет паттерны трёх данных таксонов в различных системах признаков. Это, прежде всего, средние внутривидовые паттерны в исходных признаках (график слева), а также модификации этих паттернов, полученные после удаления «неразделяющего» признака w2, — ширина чашелистика (второй слева график). Рассматривая этот график, можно заметить, что средние одинаково упорядочены по всем трем признакам, причем по признакам w3 и w4 разница между видами 1 и 2

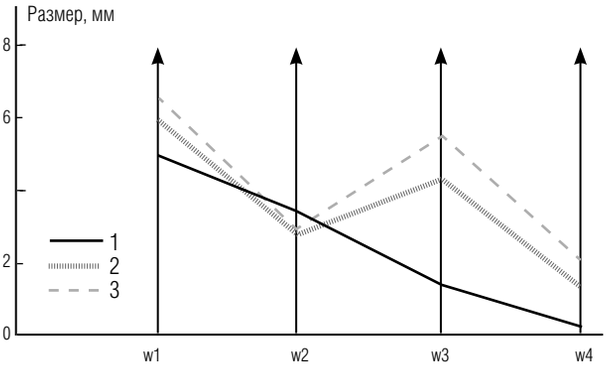


*Рис. 8.* Паттерны трёх таксонов
по данным об ирисах на уровне средних

больше, чем между видами 3 и 2. Это означает, что межкластерные разности станут ещё больше, если сформировать новый признак путём умножения w3 на w4. Действительно, паттерны в пространстве двух признаков, w3 и w4, сильно расходятся (третий слева график). Это позволяет предположить, что в системе этих двух признаков таксоны ирисов будут иметь достаточно убедительно различающиеся паттерны. Рисунок справа подтверждает оправданность этого предположения. На нём представлены «ломаные» (в данном случае просто отрезки), соответствующие всем 150 объектам таблицы данных.

Как видим, представление в параллельных координатах даёт удобную форму для визуального анализа и преобразования данных с целью построения различающихся паттернов для заранее заданных таксонов.

Обратимся теперь к другому способу представления паттернов – через конъюнктивные предикаты. Рассмотрим для примера паттерны таксонов ириса на параллельных координатах самого правого графика на *рис. 9*. Приблизительно оценим, в каких интервалах, в основном, сосредоточены точки первого таксона, отмеченного сплошной линией, на обеих координатных осях. Скорее всего, это будут, например, интервалы: [4.3, 5.9] на оси w1 и [0,1] на оси w3*w4. Каждый из этих интервалов определяет «интервальный» предикат, значения которого – «истина» для точек интервала и «ложь» для точек вне его. Естественно обозначить первый предикат через w1[4.3, 5.9], а второй – как w3*w4[0,1], Тогда предикат, соответствующий первому таксону будет не что иное, как конъюнкция этих двух, истинная на пересечении множеств истинности этих интервальных предикатов, P1= w1[4.3, 5.9] & w3*w4[0,1]. Это и есть паттерн первого таксона, представленный в виде предиката P1. Аналогично формируются конъюнкции интервальных предикатов, задающие паттерны для второго и третьего таксонов, P2= w1[5, 8] & w3*w4[4,8] и P3= w1[6,8] & w3*w4[7,16].

В принципе, можно рассматривать нечеткие предикаты, представляющие таксоны предикатами нечёткой логики, но здесь мы этого делать не будем. Заданные в форме предикатов паттерны образуют совокупность чётких классификаторов, которые для каждого конкретного объекта и заданного таксона определяют, принадлежит ли объект таксону или нет. При этом, как всегда, могут возникнуть ошибки в форме ложных плюсов и ложных минусов. Суммарная характеристика решений в данном примере представлена в *табл. 1.*

Как видим, только один предикат, а именно P1, в точности соответствует своему таксону. Два других предиката «покрывают» соответствующие таксоны на 41/50=82% и 43/50=86%, причём выдают сбои трёх типов – ложные плюсы, ложные минусы и отказы от решения. На самом деле подобные уровни

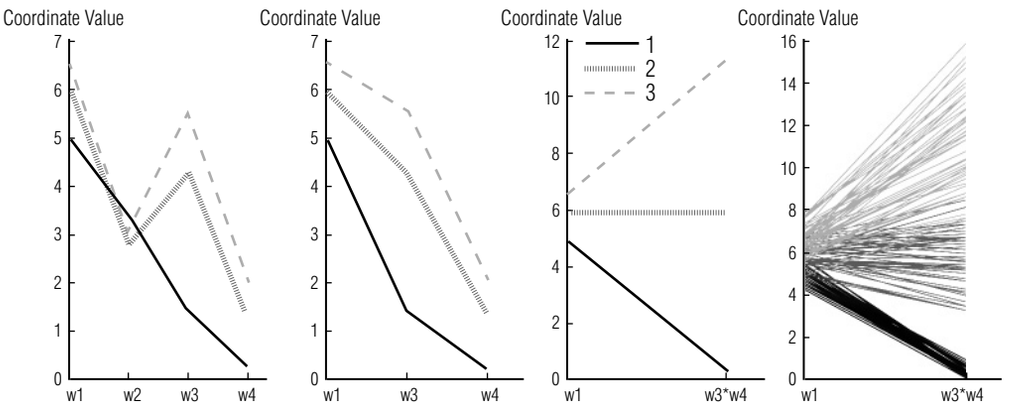


*Рис. 9.* Средние внутривидовые паттерны ирисов в разных системах признаков:
(а) все четыре признака ирисов; (б) три признака; ширина чашелистика w2 исключена;
(в) два признака, w1 и произведение измерений лепестка w3 и w4,
а также паттерны всех 150 объектов в разрезе этих двух признаков

*Таблица 1.*

**Соответствие предикатов и таксонов ириса. «Остаток» представляет собой объекты, не покрытые ни одним из предикатов**

| Таксоны / Предикаты | T1 | T2 | T3 | Итого |
|---|---|---|---|---|
| P1 | 50 | 0 | 0 | 50 |
| P2 | 0 | 41 | 3 | 44 |
| P3 | 0 | 7 | 43 | 50 |
| Остаток | 0 | 2 | 4 | 6 |
| Итого | 50 | 50 | 50 | |

покрытия вполне удовлетворительны, если рассматривать паттерн как закономерность. Многие объекты удовлетворяют паттернам неточно, характеризуя как бы «пограничные» состояния. Кроме того, уровни покрытия и точности можно отчасти контролировать, изменяя границы интервалов, входящих в формируемый паттерн.

Следует указать, что конъюнкция интервальных предикатов взаимно-однозначно отображается в предикат, истинный только на симплексе прямого (декартова) произведения соответствующих интервалов. Это потенциально приводит к геометрическому формализму, выражающему паттерны многомерными симплексами (*см. рис. 10*). Геометрически паттерн на *рис. 10(а)* есть не что иное как декартово произведение соответствующих интервалов, а именно [a1,a2]×[b1,b2].

Проблемы формирования паттернов для заданных подмножеств в терминах семейств однородных ломаных на системах параллельных координат трактуются в книгах основоположника метода А. Инсельберга [13]. Метод формирования конъюнктивных предикатов описан в [19], но в этом методе предикаты формируются для каждого кластера независимо и, следовательно, могут использовать

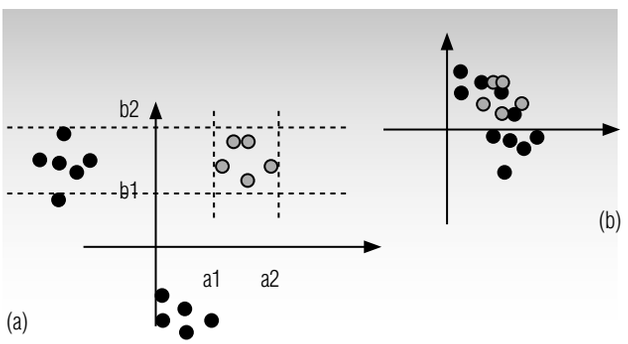


*Рис. 10.* Кластер, состоящий из светлых кружков, на графике (а) описывается паттерном, задаваемым предикатом P =a1<x<a2 & b1<y<b2, без ошибок. Напротив, аналогичный кластер на графике (b) зашумлён присутствием чёрных кружков и не может быть хорошо отделён от них конъюнктивным предикатом без использования преобразований координат

различающиеся системы признаков. При построении паттернов все множества должны рассматриваться в одной и той же системе признаков.

Ситуация усложняется тем, что на самом деле задача состоит вовсе не в том, чтобы описывать имеющиеся кластеры, а в том, чтобы определять кластеры, описанные паттернами. В этой более сложной задаче, к сожалению, пока похвастаться нечем. Формализм параллельных координат приводит к очень сложным проблемам подбора близких по форме ломаных как совокупностей отрезков прямых на плоскости, на которой изображены параллельные координаты [13], и для практического использования остаётся только разработка как можно более гибких систем, обеспечивающих возможность интерактивной работы исследователя с многообразными данными, визуализируемыми на экране компьютера. Подобных систем, в том числе бесплатных, в Интернете имеется довольно много; одна из последних разработок такого рода [22]. N.Mishra, D. Ron and R. Swaminathan утверждают, что их статья [21] посвящена проблеме автоматического формирования кластеров в виде конъюнктивных предикатов. Но это не так: название статьи не отражает существа дела – речь идет всего навсего о поиске бикластеров в бинарной матрице. Получается, что разработок по формированию кластеров в виде конъюнкций интервальных предикатов и, тем более, многомерных симплексов, т.е. декартовых произведений интервалов, вообще не делалось.

Это не оставляет нам иного выбора как использовать следующий двухэтапный метод для автоматизации формирования паттернов. На первом этапе происходит формирование кластеров с использованием обычного кластер-анализа, на втором – ищутся паттерны, достаточно полно представляющие полученные кластеры [1-3].

Когда данные характеризуют динамику функционирования социально-экономических объектов, к этому добавляется третий этап: оценка устойчивости их поведения с течением времени. В действительности смена паттерна во времени означает изменение стратегии деятельности субъекта или динамики развития объекта. В условиях относительного постоянства факторов внешней среды частота смены паттернов объектами отражает существенные внутренние параметры их структуры и системы управления. Таким образом, устойчивые по моделям поведения, т.е. не меняющие паттерна, объекты представляют особый интерес, поскольку могут представлять те объекты, которые относительно приспособились к среде и

определились со стратегией своего развития – в рамках присущего им паттерна. Напротив, социально-экономические объекты, часто меняющие паттерн, по-видимому, не могут найти свое место в относительно стабильных рыночных условиях и представляют специальный интерес с точки зрения изучения их поведения, поскольку могут выступать в качестве элементов повышенной волатильности в системе и обладать признаками банкротства организаций. Во второй части статьи мы попытаемся применить эту логику к анализу некоторых конкретных ситуаций.

### 6. Заключение

В работе предложено использовать термин «паттерн» для обозначения комбинаций значений значимых признаков, характерных для определённых групп объектов, соответствующих отличающимся от других типам поведения. Проведен обзор литературы в каждом из трёх основных аспектов этого понятия. Среди них были выделены следующие: использование понятия «паттерн» в науке и технике, методы кластер-анализа, динамика многомерных объектов. Указаны математические формализмы, позволяющие адекватно представлять понятие паттерна в соответствии с каждой из трёх когнитивных подсистем: образной, логической и геометрической. Для данных о динамике функционирования социально-экономических объектов предложено использовать типологию функционирования объектов по частоте смены паттерна. Этот подход будет применён к анализу реальных данных во второй части статьи. ■

**Литература**


1. Aleskerov F., Ersel H., Yolalan R. Personnel allocation among bank branches using a two-stage multicriterial approach // European Journal of Operational Research. – 2003. – Vol. 148/1. – P. 116-125.
2. 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.
3. Aleskerov F., Erser H., Yolalan R. Clustering Turkish commercial banks according to structural similarities // Yapi Kredi Discussion Paper Series. – 1997. – No. 97-02. – Istanbul, Turkey.
4. Apresian Y.D. An algorithm for finding clusters by a distance matrix // Computer. Translation and Applied Linguistics. – 1966. – Vol. 9. – P. 72-79 (in Russian).
5. Bandyopadhyay S., Maulik U. An evolutionary technique based on K-means algorithm for optimal clustering in RN // Information Sciences. – 2002. – 146. – P. 221-237.
6. Bassanini A., Duval R. Employment patterns in OECD countries: Reassessing the role of policies and institutions // OECD Economics Department Working Papers, No. 486, 2006. – OECD Publishing. Источник в Интернет: http://dx.doi.org/10.1787/846627332717.
7. Biernacki C., Celeux G., Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1990. – 22 (7). – P. 719-725.
8. Cheng Y., Church G.M. Biclustering of expression data // Proceedings of 8th International Conference Intelligent Systems for Molecular Biology. – 2000. – P. 93-103.
9. Ciaccio E.J., Dunn S.M., Akay M. Biosignal pattern recognition and interpretation systems. Part 4 of 4: Review of applications // IEEE Engineering in Medicine and Biology Magazine. – 1994. – Vol. 13, 2006, Issue 2. – P. 269-273.
10. Few S. Multivariate analysis using parallel coordinates, Perceptual Edge. Источник в Интернет: http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf.
11. Fisher R.A. The use of multiple measurements in taxonomic problems // Annals of Eugenics. – 1936. – 7. – P. 179-188.
12. Huang Z., Ng M.K., Rong H., Li Z. Automated variable weighting in k-means type clustering // IEEE Transactions on Pattern Analysis and Machine Learning. – 2005. – 27 (5). – P. 657-668.
13. Inselberg A. Parallel coordinates: Lecture notes, 2004. Источник в Интернет: http://astrostatistics.psu.edu/su06/inselberg061006.pdf.
14. Kanevski M., Timonin V., Pozdnoukhov A., Maignan M. Evolution of interest rate curve: Empirical analysis of patterns using nonlinear clustering tools // ESTSP 2008 Proceedings. Источник в Интернет: http://www.mafy.lut.fi/timeseries/ESTSP/PDF/26.pdf
15. Lance G.N., Williams W.T. A general theory of classificatory sorting strategies: 1. Hierarchical Systems // Comp. Journal. – 1967. – 9. – P. 373-380.
16. Laurillard D. Teaching as a design science: Building pedagogical patterns for learning and technology. – Routlegde, 2012.
17. Lu Y., Lu S., Fotouhi F., Deng Y., Brown S. Incremental genetic algorithm and its application in gene expression data analysis // BMC Bioinformatics. – 2004. – 5. – P. 172.

18. von Luxburg U. A tutorial on spectral clustering // Statistics and computing. – 2007. – 17 (4). – P. 395-416.
19. Mirkin B. Clustering for data mining: A data recovery approach. – Chapman and Hall/CRC, Francis and Taylor, Roca Baton, Fl., 2005.
20. Mirkin B. Mathematical classification and clustering. – Dordrecht-Boston-London: Kluwer Academic Publishers, 1996.
21. Mishra N., Ron D., Swaminathan R. A new conceptual clustering framework // Machine Learning. – 2004. – 56. – P. 115-151.
22. Steinberger M., Waldner M., Streit M., Lex A., Schmalstieg D. Context-preserving visual links // IEEE Transactions on Visualization and Computer Graphics. – December 2011. – Vol. 17, No. 12. – P. 2249-2258.
23. Wakefield J.C., Zhou C., Self S.G. Modelling gene expression data over time: Curve clustering with informative prior distributions. Bayesian Statistics 7 / J.M.Bernardo, M.J.Bayarri, J.O.Berger (eds). – Oxford University Press, 2003.
24. Ward J.H., Jr. Hierarchical grouping to optimize an objective function // Journal of American Statist. Assoc. – 1963. – 58. – P. 236-244.
25. Бауман Е.В. Методы размытой классификации (вариационный подход) // Автоматика и телемеханика. – 1988. – №12. – С. 143-156.
26. Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного программирования. Паттерны проектирования (Design Patterns: Elements of reusable Object-Oriented Software). – СПб: Питер, 2007.
27. Гапонова О.В. Электроэнцефалографические паттерны синдрома Веста // Медицинский совет. – 2008. – № 1-2.
28. Елкина В.Н., Загоруйко Н.Г. Об одном алфавите распознавания // Вычислительные системы. – 1966. – №12. – Новосибирск: Институт математики СО АН СССР.
29. Кузнецов Е.Н., Мучник И.Б. Монотонные системы для анализа организационных структур // Методы анализа многомерной экономической информации. – Новосибирск: Наука, Сиб. отд., 1981. – С. 71-83.
30. Лбов Г.С., Пестунова Т.М. Группировка объектов в пространстве разнотипных переменных // Анализ нечисловой информации в социологических исследованиях. – М.: Наука, 1985. – С. 141-149.
31. Малая медицинская энциклопедия. – М.: Медицинская энциклопедия. 1991-96 гг.
32. Миркин Б.Г. Группировки в социально-экономических исследованиях. – М.: Финансы и статистика, 1985.
33. Миркин Б.Г. Методы кластер-анализа для поддержки принятия решений: обзор. Препринт WP7/2011/03. – М.: Изд. дом НИУ ВШЭ, 2011.
34. Небожин А.И., Ситель А.Б. Паттерны боли при биомеханических нарушениях шейного отдела позвоночника // Мануальная терапия. – 2007. – № 1 (25). – С. 2-8.
35. Паттерны двигательных и чувствительных расстройств при патологии нервных структур в дистальных отделах верхней конечности // Медицинский портал для врачей и студентов doctorspb.ru. 2010. Источник в Интернет: http://doctorspb.ru/articles.php?article_id=1477
36. Поляков К. Прогноз – не роскошь, а инструмент управления // Директор информационной службы. – 2012. – № 2. – С. 40-44.
37. Проект INTAS 2004-77-7067 «Medical image mining: Theoretical foundation and technological aspects». 2005-2007. Источник в Интернет: http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=9945206
38. Проект INTAS 2004-77-7347 «Principles of dissimilarity-based pattern recognition in signals, symbolic sequences and images (PRINCESS)», 2008. Источник в Интернет: http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=9947067
39. Проект FP6-IST «Pattern analysis, statistical modelling and computational Learning». 2003-2008. Источник в Интернет: http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=6533866
40. Проект FP7-ICT «Pattern analysis, statistical modelling and computational Learning 2», 2008-2013. Источник в Интернет: http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=9905200
41. Сайт 19ой международной конференции по распознаванию образов http://www.icpr2008.org/index.html
42. Смит Н. Современные системы психологии. История, постулаты, практика (Current systems in psychology. History, Theory, Research, and Applications) / Пер. с англ., под общ. ред. А.А.Алексеева – М.: ОЛМА-ПРЕСС, 2003.