С.А. Судаков

КЛАСТЕРНЫЙ АНАЛИЗ В ПСИХИАТРИИ И КЛИНИЧЕСКОЙ ПСИХОЛОГИИ

Руководство для научных сотрудников, студентов и аспирантов

Под общей редакцией
академика РАМН А.С. Тиганова

Медицинское информационное агентство
МОСКВА
2010

УДК 519.2:616.89
ББК 22.172:56.14
С89

Судаков С.А.

С89 Кластерный анализ в психиатрии и клинической психологии: Руководство / Под общ. ред. А.С. Тиганова. – М.: ООО «Медицинское информационное агентство», 2010. – 160 с.: ил.

ISBN 978-5-8948-1844-3

Автором книги предложен алгоритм, позволяющий обрабатывать данные о группах объектов (обследуемых, больных)/признаков (симптомов) в системе иерархически организованных кластеров. Алгоритм показал свою универсальность и эффективность на расширенном множестве задач в сочетании с разнотипностью шкал используемых признаков.

К руководству прилагается CD-диск с программой Clust.

Для научных сотрудников, студентов и аспирантов.

УДК 519.2:616.89
ББК 22.172:56.14

Учебное пособие

Судаков Станислав Арсеньевич

КЛАСТЕРНЫЙ АНАЛИЗ В ПСИХИАТРИИ И КЛИНИЧЕСКОЙ ПСИХОЛОГИИ

Под общей редакцией академика РАМН А.С. Тиганова

Санитарно-эпидемиологическое заключение № 77.99.60.953.Д.008014.07.09 от 08.07.2009 г.
Подписано в печать 09.07.2010. Формат 60×90/16. Бумага офсетная. Гарнитура «NewtonC».
Печать офсетная. Объем 10,0 печ. л. Тираж 1000 экз. Заказ №

ООО «Медицинское информационное агентство»
119048, Москва, ул. Усачева, д. 62, стр. 1, оф. 6. Тел.: (499) 245-45-55
E-mail: miapubl@mail.ru; http://www.medagency.ru. Интернет-магазин: www.medkniga.ru

Отпечатано в ООО «Типография ПОЛИМАГ», 127247, Москва, Дмитровское ш., 107

ISBN 978-5-8948-1844-3

Оглавление

ПРЕДИСЛОВИЕ

Вы держите в руках необычную книгу. Ее автор, кандидат технических наук Станислав Арсеньевич Судаков, не приобрел широкой известности в научных кругах лишь в силу скромности и самодостаточности. Будучи интеллигентным человеком, он не придавал значения карьере, но испытывал глубокий творческий интерес к окружающему миру. Станислава Арсеньевича отличала широкая эрудиция и активная вовлеченность в самые разнообразные явления жизни. Любой посетитель его кабинета мог оценить не только широту его интересов, но и его пристрастное, с симпатиями и антипатиями, отношение к событиям, касающимся науки, образования, искусства и общественных событий. Азарт Станислава Арсеньевича завораживал как сверстников, так и людей много моложе его. Они видели перед собой глубокую, цельную личность с ярким внутренним миром и собственной позицией. В научной деятельности это проявлялось жесткой ориентацией на свои, не зависящие от моды или иных сторонних влияний приоритеты, основанные на огромном опыте ученого, сочетающего в себе честность и добросовестную кропотливость с творческой свободой и готовностью к риску.

Действительно, профессиональный путь Станислава Арсеньевича был богат и разнообразен. Достаточно сказать, что значительная его часть проходила в рамках деятельности Комитета государственной безопасности и Научного центра психического здоровья РАМН, где научные задачи нередко отличались большим количеством данных, необходимостью поиска тонких, сложных, порой неожиданных связей и закономерностей, либо, напротив, дефицитом надежной информации. Описанные условия определяли необходимость разработки универсального, гибкого и на-

дежного метода обработки информации. Таким методом стал разработанный Станиславом Арсеньевичем алгоритм кластерного анализа CLUST. CLUST является многофункциональным инструментом, адекватным для решения широкого круга задач, связанных со структурированием и иерархизацией разнотипных данных в рамках разных дисциплин, и особенно удобен при мультидисциплинарном подходе.

Интересно отметить, что в описываемый период «авторитет» кластерного анализа падал, предпочтение отдавалось факторному анализу, как дающему с формальной точки зрения более понятный, легкий для трактовки результат. В связи с этим разработка и совершенствование метода CLUST не всегда легко воспринимались окружением Станислава Арсеньевича. Но результаты практического применения метода оказались более чем убедительны. CLUST лег в основу докторской диссертации, которая успешно прошла апробацию на Ученом совете НЦПЗ РАМН. Незадолго до защиты Станислав Арсеньевич скоропостижно скончался.

На своем рабочем месте С.А. Судаков не привлекал к себе внимания окружающих, но его уход остро переживается сотрудниками. Становится все очевиднее влияние Станислава Арсеньевича на профессиональную ментальность коллектива. По сути дела, он постоянно, кропотливо, систематически, но деликатно и незаметно воспитывал у сотрудников научное мышление в полном смысле этого слова: помогал грамотно планировать эксперимент и строить гипотезы, ставить задачи. Имея дело преимущественно с описательным, нередко аморфным материалом психологов и психиатров, он приучал специалистов к доказательному подходу, четкости формулировок, систематизировал их научное мышление. Результатом подобного влияния было то, что за очень короткий период работы С.А. Судакова в НЦПЗ РАМН в сотрудничестве с ним были защищены 2 кандидатские диссертации по психиатрии, 4 кандидатских диссертации по клинической психологии и 6 дипломных работ, выполненных студентами МГУ им. М.В. Ломоносова, опубликованы статьи по различным вопросам. Творческий потенциал и возможности Станислава Арсеньевича Судакова, позволившие проникнуть в самые, казалось бы, далекие от математики направления научных исследований, способствовали осуществлению нескольких серьез-

ных совместных исследовательских проектов НЦПЗ РАМН и МГУ им. М.В. Ломоносова. С факультетом психологии МГУ, в частности с кафедрой нейро- и патопсихологии, Станислав Арсеньевич постоянно безвозмездно сотрудничал.

Важной особенностью стиля работы Станислава Арсеньевича, нетипичной для большинства математиков, являлось глубокое проникновение в суть исследуемой проблемы. Он тщательно вникал в содержание запроса, предпочитал лично знакомиться с основными библиографическими источниками, формируя свое представление о проблеме. После этого, как правило, начинался развернутый диалог, в котором и рождалось решение о выборе статистической процедуры для анализа и интерпретации полученных результатов. На разных этапах обработки информации Станислав Арсеньевич умел заметить возможность разных подходов к проблеме. Приобретая в процессе работы достаточную компетентность по исследуемому вопросу, Станислав Арсеньевич мог предложить различные способы организации баз данных, математической обработки материала и его трактовки, что нередко открывало исследователю новые неожиданные аспекты изучаемой проблемы.

Настоящее издание представляет собой подробное, поэтапное описание автором метода математической обработки данных CLUST с рассмотрением алгоритмов решения с его помощью научных задач. К книге прилагается компакт-диск с компьютерной программой, позволяющей читателю самостоятельно применять данный метод. Указанная программа разработана под руководством Станислава Арсеньевича его учеником и помощником С.Н. Мясоедовым и успешно использовалась при научных исследованиях последних лет.

Коллектив научных сотрудников разных специальностей, авторов диссертационных и иных исследовательских работ, успешно проведенных с применением метода CLUST, считает необходимым сохранить разработанный С.А. Судаковым метод математической обработки данных и сделать возможным его самостоятельное применение желающими.

Желающих использовать предлагаемый метод математической обработки данных, имеющих вопросы и предложения, просим обращаться к Сергею Николаевичу Мясоедову (адрес: smyasoe@mail.ru).

* * *

С благодарностью и теплой памятью о деликатности, бесконечной терпеливой доброжелательности и глубокой порядочности Станислава Арсеньевича Судакова.

Академик РАМН, профессор А.С. Тиганов;
д.м.н., профессор В.С. Ястребов;
к.п.н., профессор С.Н. Ениколопов;
д.м.н. А.В. Немцов;
д.б.н. И.С. Бокша;
д.м.н. И.С. Лебедева;
д.м.н. О.В. Божко;
к.м.н. О.А. Гонжал;
к.п.н. Н.В. Зверева;
к.п.н. Н.К. Корсакова;
к.м.н. И.И. Михайлова;
к.м.н. Е.В. Андриенко
и многие другие

От автора

Автор обратился к задачам, связанным с психиатрией и клинической психопатологией, в середине 1990-х гг. Главной трудностью работы с клиницистами, помимо отсутствия у него понятийных основ психиатрии, оказалась необходимость в междисциплинарном языке-посреднике, на котором формулировалась бы требуемая задача.

Существовавшая многие годы у автора уверенность, что математика является таким универсальным языком, была сильно поколеблена. Оказалось, что многому нужно учиться заново, не пытаясь всерьез обучать медиков математике, а приспосабливать свои профессиональные навыки к выработке для внутренней потребности некоторого языка, на котором, в конечном счете, и оказывается сформулированной задача, нужная клиницисту.

Опыт работы привел к необходимости реализации схемы: «медицинская» постановка задачи → осмысление пути решения → оценка его «медиком» → собственно решение → его окончательная оценка «медиком».

Если решение удовлетворяет «медика», то цель достигнута (в данном конкретном случае). Если нет, то схему нужно пройти заново, особо тщательно согласовывая общие понятия.

В технической среде существует многолетняя практика разработки технического задания, включающего как постановку задачи, так и сроки выполнения этапов ее решения. Однако эта простейшая договорная схема либо вообще не воспринимается медиками, либо игнорируется ее исполнительская дисциплина. Поэтому было решено разработать процедуру, обобщающую типичные задачи и действующую как помощник клинициста.

Математической основой предлагаемой процедуры явился кластерный анализ, последнее время постепенно проникающий в психиатрию. Выбор кластер-анализа как метода решения задач психиатрии – не единственный (см., например, [77, 78]), он объясняется лишь жизненным опытом автора.

Ведущие европейские психиатры, особенно в Германии, России, Франции, в XIX–XX вв. выработали систему описаний основных психических заболеваний на клиническом языке. К настоящему времени она уже может быть поддержана адекватным математическим аппаратом. Это соответствует и возникшим в конце XX в. требованиям доказательности используемой методики в медицине в целом и психиатрии в частности.

Кластерный анализ позволяет выделять объективно однородную группу больных/симптомов, «не используя интуицию клинициста, не поддающуюся формализации» [48, 74]. Статистически значимый рост числа статей с применением математического аппарата наблюдается в англоязычных журналах по психиатрии, а в отечественных такие статьи практически отсутствуют [48]. Тем не менее, «именно эти методы наиболее близки структуре психиатрических знаний, для которых характерна иерархическая организация, многозначность признаков, их сильная коррелированность и связанная с этим избыточность, неопределенность весовых значений определенных признаков, а также размытость границ нозологических форм» [47].

Автора подвиг на разработку вышеупомянутой процедуры опыт работы в 1960-х гг. над распознаванием образов [1–6, 9] и автоматизация ввода машинописных знаков в ЭВМ [7, 8, 10–26].

Одновременно им была решена задача машинной классификации отрезков некоторых функций действительного переменного, трактуемой как распознавание образов.

Так как тогда было особенно важно экономить оперативную память ЭВМ, то автором был предложен алгоритм, ведущий к последовательной редукции матрицы попарных расстояний заданной системы векторов, на основе которой велся поиск их скоплений – кластеров. Центрами найденных скоплений назначались их выборочные средние. Было обработано ~ 10^3 векторов размерности ~ 10^1 и выделено ~ 10^2 центров скоплений. Итогом этого стала защита в 1969 г. диссертации на соискание ученой степени кандидата технических наук.

После определенного периода вживания в новую для него тематику, отраженную в публикациях [52–55], автор решил применить найденный способ обнаружения скоплений при разработке комплекса алгоритмов, ориентированных на задачи в новой области.

Автор выражает огромную признательность С.Н. Мясоедову, взявшему на себя труд по разработке программного обеспечения.

Список сокращений и условных обозначений

$\|\|x_{ij}\|\|$, $1 \le i \le N$, $1 \le j \le M$	– матрица базы данных
$x_i = (x_{i1}, \ldots x_{iM})$	– вектор признаков, описывающий объект i
$\rho(x, y)$	– мера близости (расстояние) для векторов x и y
K_ℓ	– кластер объектов/признаков в системе m кластеров, $1 \le \ell \le m$
N_s	– абсолютная частота признака с кодом s
Н-шкала	– шкала наименований
П-шкала	– шкала порядка
О-шкала	– шкала отношений
к	– каппа, критерий оценки качества кластеризации выборки объектов/признаков
КТ	– компьютерная томография
МРТ	– магнитно-резонансная томография
ПЭТ	– позитронно-эмиссионная томография

Глава 1

Предмет и метод кластер-анализа, состав и строение алгоритма обработки данных о больных

1.1. Краткий обзор методов кластер-анализа

Представляется разумным использовать достаточно богатую монографию [28] в качестве материала, ориентированного на описание процедуры. Для этого она будет определенным образом сжата в целях выделения общих черт, набранных жирным шрифтом. Кластер-анализ начали активно применять полвека тому назад. Однако следы его использования, пусть в зачаточной форме, прослеживаются гораздо раньше. Так, еще в 1913 г. отмечена работа К. Чекановского, содержащая основные черты кластерного подхода к исследуемым данным [27]. В настоящее время разработаны многие десятки процедур кластер-анализа и число исследований продолжает увеличиваться.

1.1.1. Предмет кластер-анализа и концепции однородности

Точной постановки кластер-анализа как действия нет, и поэтому традиционная для статистики проблема выделения однородных групп уже полвека тому назад стала трактоваться как проблема распознавания образов без учителя (самообучения). При исходной неопределенности данных цель классификации/кластеризации не может быть четко сформулирована.

Большое число и разнообразие существующих алгоритмов отражает не только разнообразие вычислительных приемов, но и стоящих за ними концепций. **Если же имеются в виду кластеры с заранее заданными свойствами, то такой подход порождает алгоритмы прямой классификации:**

1) Вероятностно-статистический подход трактует однородные группы как реализации некоторых случайных величин.
2) **Структурный подход (собственно кластер-анализ) трактует однородные группы как некоторые «сгущения» в пространстве признаков, удаленные друг от друга.**
3) Вариативный подход предлагает разделение на группы по признакам; при последовательном переборе всех заданных признаков многомерного пространства – монотетический подход, при одновременном использовании всех признаков – политетический [29].

1.1.2. Типы процедур прямой классификации

Характер результирующего отношения:
а) **разбиения (классы не пересекаются);**
б) классы пересекаются или размыты;
в) **иерархическое дерево.**
Участие человека:
а) человеко-машинная классификация;
б) **машинная классификация.**
Вид задаваемых параметров:
а) **свободный от параметров;**
б) **задается число классов;**
в) вводятся пороги любого вида;
г) задается число классов и пороги.
Особенности работы алгоритма:
а) **не зависит от порядка просмотра точек выборки;**
б) зависит от порядка просмотра.

В работе [28] сказано, что кластерный анализ представляет собой специфическую методологию классификации неоднородных статистических совокупностей. Удобно выбрать универсальный язык помимо языка математической статистики и использовать универсальную терминологию бинарных отношений.

А.И. Уемов [30] показал, что в терминах «вещи, свойства, отношения» (или на языке статистики – «объекты, признаки, отношения»), практически описываются любые задачи.

1.1.3. Отношения

Вот некоторые важные отношения:

1) **эквивалентность** – разбивает все множество объектов на непересекающиеся классы;
2) **квазипорядок – отношение «быть не меньше»;**
3) **толерантность – отношения «похожести».**

1.1.4. Определение кластеров в алгоритмах прямой классификации

В работе [28] дается несколько типов определений, описывающих основные способы выделения кластеров.

Некоторые обозначения для большей компактности описаний типов кластеров:

A – описываемый кластер;

$a, a' \in A$, $b \notin A$;

ρ – расстояние;

$\bar{\rho}$ – среднее расстояние.

1) **Кластер типа сгущения** [31]**, он же компактная группа** [33]**, он же ядро** [32].

Все расстояния между объектами внутри кластера меньше любого расстояния между объектами кластера и остальной частью множества (выборки) (рис. 1, (1));

2) **Кластер с центром** [32].

Существует $\tau > 0$ и некоторая точка c, такие, что если $a \in A$, то $\rho(a,c) \le \tau$; если $a \notin A$, то $\rho(a, c) > \tau$: $\rho(a, c) < \tau$, $\rho(c, b) > \tau$ (рис. 1, (2));

3) **Кластер типа слабого сгущения** [31] **или лента** [33].

Существует $\tau > 0$, такое, что для любого $a \in A$ найдется такой объект $a' \in A$, что $\rho(a, a') \le \tau$, а для любого $b \notin A$ – $\rho(a, b) > \tau$: $\rho(a, a') < \tau$, $\rho(a, b) \ge \tau$ (рис. 1, (3));

4) **Кластер типа сгущения в среднем** [31].

Среднее расстояние внутри кластера меньше среднего расстояния объектов кластера до всех остальных объектов: $\bar{\rho}(a, a') < \bar{\rho}(a, b)$ (рис. 1, (4));

5) **Сильный кластер** [40].

Среднее внутреннее расстояние не менее чем в $\lambda > 1$ раз меньше, чем среднее расстояние от любого объекта, не принадлежащего кластеру, до всех объектов кластера: $\bar{\rho}(a, a') < \lambda^{-1} \bar{\rho}(a, b)$ (рис. 1, (5)).

6) **Кластер типа изолированного облака.**

Существует $\tau > 0$, такое, что для всех $\boldsymbol{a} \in A$, $\boldsymbol{b} \notin A$, $\rho(\boldsymbol{a}, \boldsymbol{b}) > \tau$ (рис. 1, (6)).

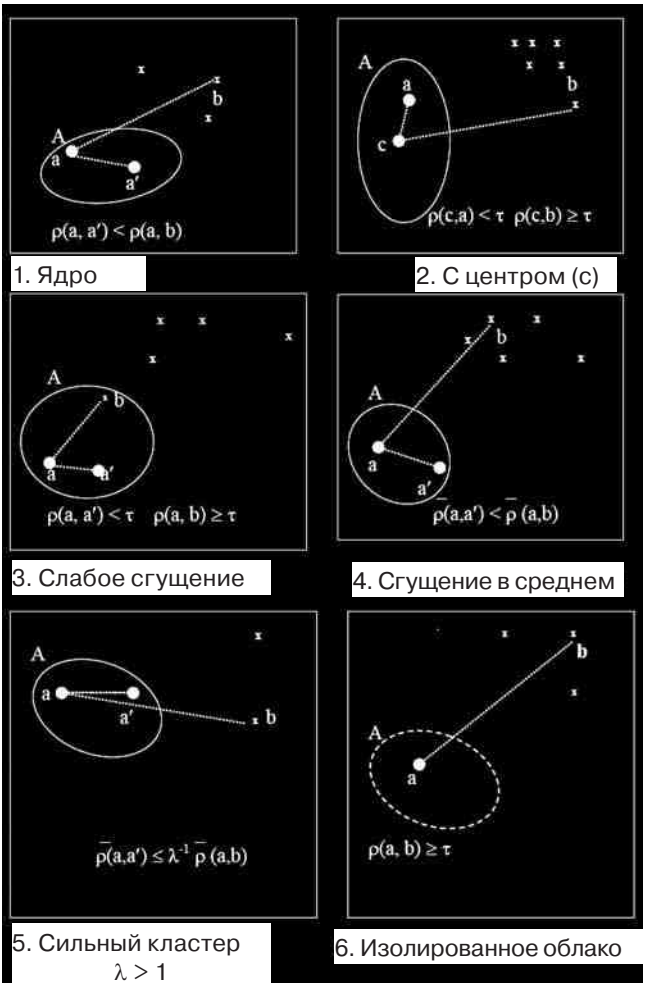

Рис. 1. Типы кластеров

Перечень определений понятия кластера [28] показывает сильную зависимость результатов обработки выборки прямым алгоритмом от установок исследователя на содер-

жательном уровне, что является особенностью кластер-анализа.

Группы алгоритмов прямой классификации

На базе 68 конкретных алгоритмов со ссылками на литературу предлагается пять групп алгоритмов:

1. **Иерархические.**
2. Типа диагонализации матрицы.
3. Эталонные процедуры.
4. Типа разрезания графа.
5. Прочие и комбинированные алгоритмы.

1.1.5. Функционалы качества классификации

На базе 46 случаев указывается перечень богатого разнообразия таких функционалов. Из них идейно близкими по замыслу можно выделить следующие три [28]:

1) **Отклонения от центров кластеров (при заданном их числе)** (F_4) [35].

2) **Сумма квадратов расстояний до центров кластеров или сумма расстояний** (F_{10}) [36, 37].

3) **Расстояние до экстремальной точки кластера** (F_{35}) [38].

Некоторые пояснения относительно функционалов качества. Пусть $\boldsymbol{m}$ – число кластеров, $\{к_\ell\}$, $1 \le \ell \le \boldsymbol{m}$ – система этих кластеров. Если $\boldsymbol{x}_i$ принадлежит шкале отношений, то

$$F_4 = \sum_{\ell=1}^{m} \sum_{i \in к_\ell} (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_\ell)^2 ,$$

т.е. суммарное отклонение от центров всех кластеров системы;

$$F_{10} = \sum_{\ell=1}^{m} \sum_{i \in к_\ell} \rho^2 (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_\ell) ,$$

что тесно связано с внутриклассовой дисперсией;

$$F_{35} = \sum_{\ell=1}^{m} \sum_{i \in к_\ell} \rho (\boldsymbol{x}_i - \boldsymbol{u}^*_\ell) ,$$

т.е. сумма расстояний до экстремальных точек $\boldsymbol{u}^*_\ell$ системы кластеров $\{к_\ell\}$.

В этом случае $\boldsymbol{x}_i$ не обязательно принадлежат шкале отношений, а экстремальные точки $\boldsymbol{u}^*_\ell$ являются обобщением центров тяжести $\bar{\boldsymbol{x}}_\ell$, реализующими

$$\min \sum_{i \in K_\ell} \rho(\boldsymbol{x}_i, \boldsymbol{u}^*_\ell).$$

1.2. Комплекс алгоритмов кластеризации данных

1.2.1. Некоторые соображения относительно конструкции алгоритма и его ориентированности на данные

Хотя любая классификация всегда является воплощением наших модельных представлений [28], при отсутствии пока такой дисциплины, как теоретическая психиатрия, в основании алгоритмов классификации/кластеризации могут быть положены лишь весьма общие подобные представления:

1) Алгоритм должен работать с достаточно богатым описанием объекта (больного), т.е. с признаками, принадлежащими нескольким шкалам измерения данных, в частности наименований, порядка, отношений.

2) В алгоритме исходно должна быть задана мера близости/сходства между объектами в четко фиксированном пространстве признаков.

3) Результатом работы алгоритма должна быть структура системы выделенных кластеров вместе со средством оценки ее качества.

4) Алгоритм должен работать с исходными признаками, имеющими содержательный практический (клинический) смысл.

5) Число кластеров не предполагается заранее известным.

6) Алгоритм должен работать с множеством признаков ~10^2.

7) Наличие в работе алгоритма иерархичности «снизу – вверх» должно сочетаться с возможностью выбора пользователем некоторого «наилучшего» уровня кластеризации.

8) Для объектов (больных) должно быть предусмотрено указание их «типичных» представителей, роль которых играют центры Чебышева соответствующих кластеров.

9) Должна быть предусмотрена возможность вставки/исключения признаков и/или объектов.

10) Должно быть предусмотрено одинаковое упорядочение элементов, относящихся к шкале порядка.

11) Алгоритм должен быть опробован на объектах другой природы, нежели больные, с целью оценки его эффективности.

12) Так как процедура является алгоритмом прямой классификации, то:

а) им должна обеспечиваться простота и содержательная ясность результата;

б) допускается контролируемое вмешательство путем изменения параметров;

в) осуществляется визуализация данных;

г) обеспечивается невысокая трудоемкость.

13) Clust не проверялся на искусственных массивах данных с известной структурой, генерируемой на ЭВМ и в режиме статистического моделирования.

14) В работе [28] говорится по поводу оценки качества кластеризации, что критерий качества удобно сформулировать на языке, близком к экономическому – какие потери будут наблюдаться при отклонении от экстремума. Однако это вряд ли можно сделать для задач, связанных с психиатрией из-за отсутствия точного представления о структуре данных.

15) Наконец, окончательный критерий качества кластер-анализа — практическая полезность результата [39].

Представляется важным отметить [28], что у кластер-анализа нельзя отнимать способность отыскивать такие «скопления», которые впоследствии могут не просто объясняться некоторой теорией, но и давать изначальный толчок для ее создания. Эвристическая роль методов кластеризации в современных условиях интенсивного «наступления на многомерность» во всех областях науки очень значительна [28].

Проблема кластеризации трактуется как один из подходов к изучению плохо организованных (диффузных) систем [51]. Там же эта задача считается «давно и хорошо известной задачей предварительной классификации наблюдений, но до сих пор она решалась на интуитивном уровне... кластер-анализом лучше называть новые приемы, где известные методы статистики малоэффективны или бессильны».

За прошедшие четыре десятилетия появилось много результатов, поддающихся теоретическому осмыслению, и эффективных вычислительных процедур [28]. Описанный в данной монографии метод является представителем таких процедур.

Предложенная ниже процедура получила название Clust и представляет собой семейство трех алгоритмов Clust1, Clust2, Clust3.

Clust1 – наиболее развитый алгоритм, по которому получена основная масса результатов на конкретных задачах. В нем отсутствует информация о числе ожидаемых кластеров.

Clust2 выбирает последовательно заданное число наиболее удаленных объектов/признаков выборки и строит кластеры как выборочное содержимое в их областях Дирихле–Вороного.

Clust3 строит последовательное покрытие заданной выборки сферами фиксированного радиуса с центрами в каждом объекте/признаке и позволяет оператору программы выбрать их оптимальное число – такое, при котором минимизируется число сфер покрытия.

1.2.2. Алгоритм Clust1

Обобщающей конструкцией, позволяющей формулировать задачи кластеризации данных «больной×симптом», служит матрица «объект×признак» $\boldsymbol{X} = \|\boldsymbol{x}_{ij}\|$, $1 \le \boldsymbol{i} \le \mathrm{N}$, $1 \le \boldsymbol{j} \le \mathrm{M}$ [40] объема N×M, где N – число объектов, а M – число признаков в рассматриваемой задаче.

Объекты обладают однородностью, отличаясь друг от друга лишь именами. Для признаков существенна принадлежность к шкале измерения данных [40]. В описываемой ниже конструкции выбраны три шкалы измерения данных – наименований (Н), порядка (П), отношений (О), точнее интервалов и отношений [40]. Главное отличие признаков друг от друга заключается в природе шкал. Лишь О-шкала представляет «настоящие» числа, с которыми можно проделывать все алгебраические операции. Для нее пространство столбцов матрицы $\boldsymbol{X}$ является полноценным векторным пространством. Такой набор шкал напоминает введенное Стивенсом [43] представление измерений четырьмя основными шкалами:

1. Номинальной (произвольная нумерация).
2. Ранговой (школьные отметки).
3. Интервальной (градусы Цельсия или Фаренгейта).
4. Относительной (температура по Кельвину).

В [28] введено понятие шкалы разностей, позволяющее оперировать оценочными баллами, точнее говорить об их разности, однако в системе Стивенса оно включено в ранговую. В описании Clust это также подразумевается, причем здесь объединены в одну шкалу интервальная и отношений.

Естественно возникают два вида кластеризации – объектов и признаков, тесно взаимосвязанные. В первом случае ищется разбиение множества объектов на близкие группы при заданном множестве признаков. Во втором – разбиение множества признаков на близкие группы при заданном множестве объектов. При этом требуются две матрицы мер близости – «объект×объект» порядка N×N и «признак×признак» порядка M×M.

1.2.3. Меры близости для объектов

Описание объекта как M-мерного вектора в общем случае имеет вид

$$X = (x^н, x^п, x^о),$$

где $x^н, x^п, x^о$ – подвекторы соответствующих шкал.

Меры близости объектов x_i, x_j определяются следующим образом.

Для Н-шкалы

$$\rho^{Н}_{ij} = \begin{cases} \dfrac{N(i) + N(j)}{N_Н}, & \text{если } N(i) \neq N(j), \\ 0, & \text{если } N(i) = N(j), \end{cases}$$

где N(i) и N(j) – абсолютные частоты имен в Н-шкале для i-го и j-го объектов в выборке N элементов, где $N_Н$ – число объектов с Н-шкалой.

В общем случае в выборке объема N с $N_н$ параметрами описаний объектов, относящихся к Н-шкале, мера близости для пары объектов i, j имеет вид:

$$\rho^{H}_{ij} = \sum_{\alpha \in \mathbf{H}} \rho^{\alpha}_{ij}.$$

Для П-шкалы

$$\rho^{П}_{ij} = \left| r_i - r_j \right|,$$

где r_i и r_j — нормированные ранги элементов П-шкалы для i-го и j-го объектов выборки. Нормированным рангом элемента $\alpha \in$ П-шкале называется число, равное числу всех предшествующих α членов возрастающего вариационного ряда плюс половина длины той серии, к которой относится сам α:

$$\underbrace{\ldots\ldots\ldots\ldots}_{p}\underbrace{\alpha\alpha\ldots\ldots\alpha}_{\ell} \qquad r(\alpha) = p + \frac{\ell+1}{2} = \frac{1}{\ell}\sum_{\alpha=1}^{\ell}\alpha + \rho.$$

Как и для Н-шкалы, наличие нескольких параметров описания объектов, принадлежащих П-шкале, означает суммирование мер сходства по ним. В общем случае выборки объема N с $N_п$ параметрами из П-шкалы мера близости для пары объектов i, j имеет вид

$$\rho^{П}_{ij} = \sum_{\alpha \in П} \rho^{\alpha}_{ij}.$$

Для О-шкалы

$$\rho^{O}_{ij} = \frac{\left| x_i - x_j \right|}{\hat{x} - \check{x}},$$

где x_i и x_j – значения элементов О-шкалы, соответствующих i-му и j-му объектам выборки, а $\hat{\mathbf{x}}$ и $\check{\mathbf{x}}$ – соответственно наибольший и наименьший из ее элементов.

При $\hat{\mathbf{x}} = \check{\mathbf{x}}$ выражение $\frac{\mathbf{0}}{\mathbf{0}}$ принимается за 0.

Мера близости в О-шкале для объектов i, j имеет вид

$$\rho^{O}_{ij} = \sum_{\alpha \in O} \rho^{\alpha}_{ij}.$$

В [40] общая мера близости для объектов x_i и x_j определяется как среднее арифметическое по всем имеющимся в базе данных шкалам, например:

$$\rho(x_i, x_j) = \begin{cases} \frac{1}{3}\left(\rho_{ij}^{Н} + \rho_{ij}^{П} + \rho_{ij}^{О}\right) - \text{для всех трех шкал Н, П, О,} \\ \frac{1}{2}(\rho_{ij}^{Н} + \rho_{ij}^{П}) - \text{для двух шкал, например Н, П,} \\ \rho_{ij}^{Н} - \text{для одной шкалы, например Н.} \end{cases}$$

В [40] говорится о возможности выбора такой меры в зависимости от априорной информативности признака, но при разнотипных шкалах предлагается задавать весовые параметры пропорционально потенциальной информативности шкалы.

Всю конструкцию можно обобщить, рассматривая выпуклые комбинации весовых параметров $\lambda_Н$, $\lambda_П$, $\lambda_О$, что удобно бывает на практике как средство подбора «удачной» кластеризации.

1.2.4. Меры близости для признаков

Как и для мер близости между объектами x_i, x_j, для признаков используется конструкция, описанная в [40].

1. Признаки однотипные.

1.1. При сопоставлении двух признаков X_i, X_j, принадлежащих О-шкале, мера определяется как

$$\rho\,(X_i^{О}, X_j^{О}) = 1 - |r(X_i, X_j)|,$$

где r – классический коэффициент корреляции.

1.2. Если X_i и X_j принадлежат шкале порядка, то мера близости определяется как мера Кендалла–Кемени [40–42].

Она определяется как средняя мера несогласованности признаков X_i, X_j на всех $C_м^2$ парах признаков выборки:

$$\rho^{П}(X_i, X_j) = 1/C_M^2 \sum_{a,b} \rho^{П}(a, b),$$

где

$$\rho^{П}\ (a, b) = \begin{cases} 0, & \text{если порядковые отношения одинаковы;} \\ 0{,}5, & \text{если по } X_i - \text{в одну сторону } (>, <), \\ & \text{а по } X_j - \text{одинаковы } (=); \\ 1, & \text{если по } X_i - \text{в одну сторону } (>, <), \\ & \text{а по } X_j - \text{в другую } (<, >). \end{cases}$$

1.3. Если X_i, X_j принадлежат шкале наименований, то

$$\rho^{н}(X_i, X_j) = 1/C_M^2 \sum_{a,b} \rho^{н}(a, b),$$

где

$$\rho^{н}(a, b) = \begin{cases} 0, & \text{если } X_i(a) = X_i(b), \\ & X_j(a) = X_j(b), \\ & \text{или } X_i(a) \neq X_i(b), \\ & X_j(a) \neq X_j(b), \\ 1, & \text{если } X_i(a) = X_i(b), \\ & X_j(a) \neq X_j(b), \\ & \text{или } X_i(a) \neq X_i(b), \\ & X_j(a) = X_j(b). \end{cases}$$

2. Признаки разнотипные [40].

1) «Ослабление» признака $X^{о}$ до $X^{п}$, т.е. преобразование $x^{о} \to \tilde{x}^{п}$ достигается учетом на элементах $X^{о}$ только отношения П, после чего определяется

$$\tilde{\rho}^{пп} = 1/C_M^2 \sum_{a,b} \rho(a, b) \cdot$$

2) «Усиление» признака $X^{п}$ до $X^{о}$, т.е. преобразование $X^{п} \to$ $\to X^{о}$, делается так, чтобы, во-первых, значение порядка объектов по признаку $\tilde{x}^{о}$ совпадало с тем, что указано признаком $X^{п}$, а, во-вторых, числовые значения признака $\tilde{x}^{о}$ были максимально коррелированы со значениями признака $X^{о}$.

Наконец, определяется $\tilde{\rho}^{оо} = 1 - |r|$, где r – коэффициент корреляции для $\tilde{x}^{о}$. Окончательно для случая шкал О и П определяется средняя мера близости вида

$$\rho^{оп} = \frac{\alpha\tilde{\rho}^{пп} + \beta\tilde{\rho}^{оо}}{\alpha + \beta},$$

где α и β – веса, отражающие доверие к величинам $\tilde{\rho}^{пп}$ и $\tilde{\rho}^{оо}$, $\alpha+\beta = 1$.

Для шкал О и Н:

1) «Ослабление» признака $X^{о}$ до $X^{н}$, т.е. преобразование $x^{о} \to \tilde{x}^{н}$, достигается приписыванием всем различным значениям $X^{о}$ разных имен, после чего определяется:

$$\tilde{\rho}^{нн} = 1/C_M^2 \sum_{a,b} \rho(a, b).$$

2) «Усиление» признака $X^{Н}$ до $X^{О}$, т.е. преобразование $X^{Н} \to \tilde{X}^{О}$, достигается подобно описанному в п. 2 для шкал О и П, после чего вычисляется $\tilde{\rho}^{ОО}$ с использованием коэффициента корреляции r для $\tilde{X}^{О}$.

Окончательно для случая шкал О и Н определяется средняя мера близости вида:

$$\rho^{ОН} = \frac{\alpha\tilde{\rho}^{НН} + \beta\tilde{\rho}^{ОО}}{\alpha + \beta}.$$

Для шкал П и Н:

1) «Ослабление» признака $X^{П}$ до $X^{Н}$, т.е. преобразование $X^{П} \to \tilde{X}^{Н}$ достигается приписыванием разных имен элементам $X^{П}$, имеющим разные порядковые номера, после чего определяется $\tilde{\rho}^{НН}$.

2) «Усиление» $X^{Н}$ до $X^{П}$, т.е. преобразование вида $x^{н} \to \tilde{x}^{п}$, происходит по аналогии со случаем 2 для шкал О и П, после чего находится $\tilde{\rho}^{ПП}$.

Окончательно для случая шкал П и Н определяется средняя мера близости вида:

$$\rho^{ПН} = \frac{\alpha\tilde{\rho}^{НН} + \beta\tilde{\rho}^{ПП}}{\alpha + \beta}.$$

Во всех трех случаях при выборе $\alpha = \beta$ значения $\rho^{ОП}$, $\rho^{ОН}$, $\rho^{ПН}$ равны средним арифметическим соответствующих $\tilde{\rho}$.

Детали преобразований вида $x \to \tilde{x}$ для всех трех случаев указаны в [40].

1.2.5. Работа алгоритма Clust1

После определения всех необходимых мер близости, а именно $\rho^{Н}$, $\rho^{П}$, $\rho^{О}$ для матрицы мер близости объектов порядка N×N и $\rho^{НН}$, $\rho^{ПП}$, $\rho^{ОО}$, $\rho^{НП}$, $\rho^{НО}$, $\rho^{ПО}$ для матрицы мер близости признаков порядка M×M, начинает действовать алгоритм Clust1.

Ниже для краткости будем называть «объект×признак» — «элемент», X – исходная выборка.

1. Все исходные элементы объявляются кластерами нулевого уровня.

2. По матрице мер близости элементов находится любая ближайшая пара, т.е. $i_1, i_2 = \arg\min_{i \neq j} \rho(x_i, x_j)$.

Начало подобной процедуры описано в [45] для случая цифровых признаков. В данном же случае это ограничение снимается.

3. Для каждого элемента x проверяется – ближе ли он к $\{x_{i1}, x_{i2}\}$ или к остальной части выборки:

$$1)\ \rho(x,\{x_{i1}, x_{i2}\}) \le \rho(x, X \setminus \{x_{i1}, x_{i2}\})$$

или

$$2)\ \rho(x,\{x_{i1}, x_{i2}\}) > \rho(x, X \setminus \{x_{i1}, x_{i2}\})$$

В случае 1 пара $\{x_{i1}, x_{i2}\}$ пополняется элементом x, а в случае 2 пара $\{x_{i1}, x_{i2}\}$ выделяется как отдельный кластер и в дальнейшем устраняется из рассмотрения. Такая процедура повторяется шаг за шагом, сопровождаясь соответствующей редукцией матрицы мер близости.

Процесс, реализуемый Clust1, носит уровневый характер, при котором выстраивается система кластеров «снизу вверх», как показано на рис. 2.

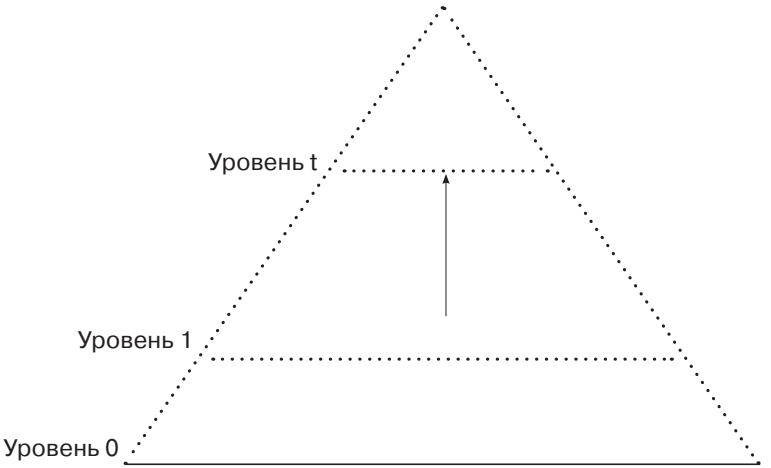

Рис. 2. Схема уровневой организации кластерной системы объектов/признаков

Предлагаемая система напоминает рост клеточной ткани при заживлении раны либо образование кристаллов в насыщенном растворе. Исходные данные считаются кластерами

нулевого уровня. Первичные кластеры – элементы первого уровня, а элементами второго уровня являются кластеры кластеров, полученных на первом уровне, и т.д. до тех пор, пока не останутся два кластера, слияние которых на последнем уровне и завершает процесс, приводя к исходной выборке.

Практика показала, что уровневая система кластеров обладает в целом неким оптимальным свойством аккумуляции скрытых свойств, распределенных в кластерах-носителях.

При этом, как отмечается ниже, в различных задачах выявляются кластеры, содержащие объекты, обладающие некоторыми свойствами «маргинальности». С другой стороны, на некоторых уровнях проявляются такие свойства коллективного поведения кластеризуемых объектов, как арьергардные и авангардные группы, называемые эффектом пелотона. Для решения задачи представляется полезным дать весь спектр кластеризуемого материала однотипным способом, пройдя в представленных данных «объект×признак» весь путь «снизу–вверх».

Вводятся два вида межкластерной меры близости – «жесткая» (h) и «мягкая» (s). Так, для пары кластеров K_1, K_2 имеем по определению:

$$\rho_h(K_1, K_2) = \min_{\substack{a \in K_1 \\ b \in K_2}} \rho(a, b),$$

$$\rho_S(K_1, K_2) = \frac{1}{N_1 N_2} \sum_{\substack{a \in K_1 \\ b \in K_2}} \rho(a, b),$$

где N_1 и N_2 – объемы кластеров K_1, K_2 (числа кластеров уровня 0).

В Clust1 используются обе меры близости, и на каждом уровне выше первого может быть выбрана любая. На первом же уровне всегда используется h-мера. На практике оказалось удобным повторять одну и ту же, выстраивая последовательности вида ss...s или hh...h. Решение об остановке оставляется на усмотрение пользователя/исследователя.

Можно предложить процедуру наилучшего выбора меры межкластерной близости. На шаге t+1, $t \geq 1$ выбирается та мера, которая реализует max{к(h), к(s)}, т.е. наилучшее «качество», определенное в разделе 2.7.

Мера близости «h» имитирует некий процесс, распространяющийся от исходного локуса близости до отдаленных элементов его окрестности. Мера близости «s» скорее ориентирована на выделение обособленных островков локальной плотности. Поэтому удобно использовать обе эти меры.

В целях уменьшения громоздкости излагаемого ниже материала во всех задачах не приводятся исходные базы данных — матрицы $\|x_{ij}\|$ порядка N×M. Изложение начинается с изучения близости кластеров первого уровня.

1.2.6. Центр Чебышева как «типичный представитель» кластера

Для каждого кластера находится точка

$$\operatorname{argmin}_x \max_y \rho(x, y),$$

называемая центром Чебышева подвыборки, содержащейся в кластере.

Эта точка расположена наиболее близко ко всем точкам кластера и поэтому может служить своего рода «типичным его представителем».

Центр Чебышева системы точек был введен в обиход еще в 1970-е гг. [44] в теории распознавания образов, где он был назван «обобщенным портретом». В описываемом выше случае речь идет лишь о выборочной оценке центра Чебышева по содержимому кластера.

1.2.7. Качество кластеризации объектов/признаков

Несколько замечаний об оценке качества процедур кластер-анализа. Задачей кластеризации называется поиск закономерностей в таблицах объект×признак [40]. Так как любой реальный объект обладает бесконечным числом признаков, то выделение их конечного числа даже в узко профессиональном смысле вносит элемент субъективности. Аналогично стоит вопрос и с мерами близости. Для снятия этой проблемы нужна цель, т.е. постановка задачи, для которой формируются кластеры.

Качество же кластеров определяет степень достижимости этой цели. Важно отметить, что выбор цели субъективен, но проверка ее соответствия — объективна. Далее излагается ряд соображений о качестве системы возникающих кластеров.

Так, если она «нехорошая», то может оказаться, что дело не столько в «плохом» алгоритме построения кластеров, сколько в постановке задачи (формулировке цели). «Плохая» кластеризация может говорить о гомогенности выборки, обработанной неудачно выбранной системой признаков. Обычно пользователь не может задать строго «целевую» систему признаков и пользуется лишь косвенными характеристиками, которые могут оказаться неинформативными. Это обстоятельство можно использовать для проверки информативности конкретной системы признаков.

В [40] описан ряд программ кластер-анализа возрастающей сложности, начиная со сферических кластеров простой формы – FOREL, FOREL-2, SKAT, GRUPPA, KOLAPS. Используются количественные описания признаков и евклидова метрика для мер близости. Программа FOREL-2 близка по смыслу к алгоритму Clust3.

Возвращаясь к проблеме оценки качества процедур кластер-анализа в комплексе Clust, перейдем к уровневой системе кластеров.

На каждом уровне итог кластеризации предлагается описывать так называемой структурной матрицей системы кластеров [61].

Пусть получено m кластеров для N точек выборки. Определим s_{ij} как число тех ее точек (кластеров нулевого уровня), которые, принадлежа кластеру K_i , находятся ближе к центру Чебышева кластера K_j. Ясно, что при отсутствии пустых кластеров $s_{ii} \geq 1$, а величина $\sum_i s_{ij}$ дает объем кластера K_j.

Представим себе двух экспертов, которые независимо друг от друга относят точки выборки к m категориям. Под «категорией» понимается принадлежность наблюдаемой точки к одному из кластеров системы. Тем самым эксперт 1 выбирает ее близость к центру Чебышева, а эксперт 2 – ее принадлежность к составу кластера. Тогда статистика каппа

$$\kappa = \frac{1/N \sum_{j=1}^{m} s_{ij} - 1/m}{1 - 1/m} \quad ,$$

при этом κ находится в пределах $0 < \kappa \leq 1$,

где числитель – разность наблюдаемой доли их согласованных оценок и ожидаемой доли согласованных случайно, а знаменатель – наибольшая величина такой разности, служит мерой «обособленности» m кластеров системы. Чем κ больше, тем естественнее считать выше «качество» кластеризации.

Всюду далее система m кластеров любого уровня описывается матрицей $S = \|s_{ij}\|$ порядка m×m:

	Кластеры принадлежности					
Кластеры близости		1	2	…	…	m
	1	s_{11}	s_{12}	…	…	s_{1m}
	2	s_{21}	s_{22}	…	…	s_{2m}
	…	…	…	…	…	…
	…	…	…	…	…	…
	m	s_{m1}	s_{m2}			s_{mm}
		$\sum_i s_{i1}$	$\sum_i s_{i2}$			$\sum_i s_{im}$

Эта матрица названа структурной матрицей системы m кластеров.

Введена следующая градация согласованности оценок экспертов

$$\kappa - \begin{cases} < 0,40 & \text{слабая,} \\ \text{от } 0,40 \text{ до } 0,75 & \text{хорошая,} \\ \geq 0,75 & \text{высокая.} \end{cases}$$

Предложенная оценка κ может быть использована для любого алгоритма, необязательно Clust1.

Кластер-анализ Clust1 ориентирован на выявление «тесных» скоплений в наблюдаемых данных. С этим связана доля элементов, близких к центрам Чебышева своих кластеров:

$$q = 1/N \sum_{i=1}^{m} s_{ii},$$

которую естественно назвать «кучностью»

$$q = \kappa\,(1 - 1/m) + 1/m.$$

Помимо обычной процедуры построения системы кластеров по матрице данных $\|x_{ij}\|$, оказывается полезным огрублять результат, работая с осреднением либо объектов, либо признаков. В этом случае выявляется крупномасштабный вклад одного компонента базы данных на фоне другого. Особенно это наглядно, когда они оба описываются одной шкалой измерения. Пусть, например, признаки принадлежат П-шкале с кодами 0, 1,... k, тогда

$$\begin{Vmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \dots & \vdots \\ x_{N1} & \dots & x_{NM} \end{Vmatrix} \rightarrow \begin{Vmatrix} n_{10} & \dots & n_{1K} \\ \vdots & \dots & \vdots \\ n_{N0} & \dots & n_{NK} \end{Vmatrix}$$

и происходит редукция матрицы N×M к N×1 или N×(K+1).

Если объекты принадлежат П-шкале с кодами 0, 1, ... m, то

$$\begin{Vmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \dots & \vdots \\ x_{N1} & \dots & x_{NM} \end{Vmatrix} \rightarrow \begin{Vmatrix} n_{01} & \dots & n_{0M} \\ \vdots & \dots & \vdots \\ n_{m1} & \dots & n_{mM} \end{Vmatrix},$$

происходит редукция матрицы N×M к 1×M или (m+1)×M.

Здесь всюду n_{ij} – соответствующие компоненты частотных описаний строк/столбцов. При описании конкретных задач, в которых был применен Clust1, указано как наличие подобного преобразования, так и его цель.

Оценка качества кластеризации, в отличие от указанных выше функционалов, осуществленная в терминах к-критерия, представляется полезной на материале рассмотренных ниже задач. Можно считать, что к не опирается на аппарат проверки статистических гипотез. За ним остается лишь индикативная функция, показывающая три степени некоего соответствия, что особенно наглядно для пользователя. Такая огрубленная картина более понятна клиницисту-исследователю.

Система кластеров, реализованная в Clust1, образует своего рода пирамиду с основанием N или M и завершающуюся одним кластером – исходной выборкой (см. рис. 2). Ее высота – число уровней – определяется как ближайшее целое к lnN, если описывать развитие уровней функцией

$$N_t = N\ \exp(-t),\ 0 \le\ t \le \ln N.$$

Можно сказать, что чем ниже уровень, тем выше концентрация некого прослеживаемого свойства, однако при этом «мельче» его носители. Поэтому оптимальные в этом смысле кластеры должны находиться где-то «не слишком низко, не слишком высоко».

Пусть N_t – число кластеров уровня t, а ℓ_t – их характерный размер, определяемый как полусумма наименьшего и наибольшего внутрикластерного расстояний:

$$\ell_t = \frac{1}{2}(\check{\ell}_t + \hat{\ell}_t).$$

Если нормировать (для объектов) величины N_t на N, то оптимальный уровень можно определить как точку минимума функций

$$1)\ f_1(t) = N_t / N + \ell_t \quad t \ge 1,$$
$$2)\ f_2(t) = \max\{N_t / N,\ \ell_t\}.$$

Если описывать поведение величин ℓ_t функцией $\ell_t = \exp(t)\ - 1$,

$$0 \le t \le \ln(L+1), \text{где } L = \max_t \ell_t,$$

то минимум функции $f\ell(t) = N \exp(-t) + \exp(t) - 1$ достигается в точке $t^* = 1/2\ln N$, что хорошо согласуется с данными рассмотренных задач. Наконец, можно связать оба параметра N и L, описывающих «пирамиду» на рис. 2, введя функцию $f_1(t) = N\ \exp(-t) + 1/L \exp(t) - 1$, минимум которой достигается при $t^* = 1/2\ln(NL)$. Все рассмотренные ниже случаи взяты для меры межкластерной близости ρ_h.

Характеристики «качества» уровневой системы показаны на рис. 3,

где

T — число уровней «снизу доверху» от 1 до T;

N_t – число кластеров на уровне t;

$\ell_t = \frac{1}{2}(\check{\ell}_t + \hat{\ell}_t)$ – характерный размер кластеров уровня t;

к — качество;

$$t^* = \underset{t}{\operatorname{argmin}} \begin{cases} \dfrac{N_t}{N'} + \ell_t \\ \max\left(\dfrac{N_t}{N'}, \ell_t\right) \end{cases}$$ — оптимальный уровень.

Широко распространенная форма представления результата кластер-анализа в виде дендрограммы [28] для уровневой системы получается громоздкой. Вместо нее используется аппарат структурных матриц [61].

$N \times M$ ↓	база данных						
$N \times M'$ или $N' \times M$ ↓	частотное описание данных по объектам/признакам						
$N' \times N'$ или $M' \times M'$ ↓	уровень – 0	N_t / N	ℓ_t	$N_t / N + \ell_t$	$\max(N_t / N, \ell_t)$	κ	
$N_1 \times N_1$ ↓	1						
$N_2 \times N_2$ ↓ ⋮ ↓	2						
2×2 ↓	$T - 1$						
1×1	T			t^*		t^*	

Рис. 3. Система уровневых кластеров в задаче...

1.3. Описание алгоритма Clust2

Если в Clust1 действие алгоритма начиналось с пары наиболее близких точек с последующим расширением локуса их плотности, то в данном случае в заданной выборке выделяется система наиболее далеко разнесенных точек, принадлежащих исходной выборке. Поэтому можно говорить о построении максимального каркаса, т.е. ребер, соединяющих эти точки. В Clust2 не идет речь об уровневом развитии системы кластеров «снизу–вверх» в том понимании, которое присутствует в Clust1. На первом этапе находится пара точек, для которой реализуется диаметр выборки, т.е. пара ее точек

$$\{i_1, i_2\} = \underset{i \neq j}{\operatorname{argmax}} \rho(x_i, x_j) \cdot$$

На следующем этапе ищется точка x_{i3}, наиболее удаленная от точек x_{i1}, x_{i2} в смысле реализации

$$i_3 = \arg\max_k \min\{\rho(x_k, x_{i1}), \rho(x_k, x_{i2})\}$$

и так далее до тех пор, пока не будет найдена последняя точка x_{im}, где m –заранее задано. Для описанной процедуры требуется задание матрицы попарных мер близости $\|\rho(x_i, x_j)\|$ порядка $N \times N$.

Представление выборки точками x_{i1}, x_{i2}, ... x_{im} позволяет, введя их области близости (области Дирихле), т.е. множества вида

$$D(x; x_k) = \left\{x : \rho(x, x_k) < \rho_{l \neq k}(x, x_l)\right\}$$

разбить пространство на ячейки, порожденные точками m-максимального каркаса. После этого каждую из m областей Дирихле можно представить ее центром Чебышева той части выборки, которая попала в эту D-область.

Алгоритм Clust2 служит средством представления исходной выборки существенно меньшим числом ее точек, тем не менее, подчеркивающим ее конфигурацию. Впервые идея Clust2 была реализована в конце 1960-х гг. В.И. Зайцевым–Зотовым при выборе эталонов читающего устройства для m = 10 [7]. Задача разумного выбора m оставляется на долю пользователя/исследователя.

1.4. Описание алгоритма Clust3

Clust3 ориентирован на покрытие выборки данных окрестностями сфер фиксированного радиуса с последующим уменьшением порядка матрицы попарных расстояний в ходе итерационного процесса, аналогичного примененному в Clust1.

Пусть выборка $X = \{x_i\}, 1 \le i \le N$ содержит N объектов и есть основания предполагать действие на ее элементы некоторого изотропного шума.

Сфера радиуса ρ с центром в точке $x_i (S(x_i; \rho) = \{x : \rho(x, x_i) \le \rho\})$ содержит $n_i(\rho)$ точек выборки X. Первым шагом алгоритма Clust3 служит выбор такого x_i, который реализует $\max_i n_1(\rho)$.

Иными словами, происходит выбор точки i_1 из условия $i_1 = \arg\max n_i(\rho)$ и переход к выборке X_1, являющейся X без всех $x_j \in S(x_i, \rho)$. Далее процесс идет итеративно, находят i_2, и т.д. до полного исчерпания исходной выборки.

Результатом применения Clust3 служит последовательность множеств, попадающих в сферы $S(x_{i1}, \rho)$, $S(x_{i2}, \rho)$... , которые объявляются кластерами выборки X, обработанной алгоритмом Clust3 сферами заданного радиуса ρ.

Выбор $\rho = 0$ оставляет исходную выборку неизменной, а выбор $\hat{\rho} = \max_{ij} \rho(x_i, x_j)$ осуществляет ее покрытие одной сферой максимального радиуса. Естественно, величина ρ должна зависеть от шкалы измерения данных, к которой относятся центры x_i. Если принять величину $\hat{\rho}$ за 1, то, изменяя свободный параметр ρ в сторону уменьшения, можно получить изменяемое качество кластеризации. Естественно, каждый отделяемый на некотором шаге кластер (не играющий роль уровня, введенного для процедуры Clust1) может быть охарактеризован своим центром Чебышева.

Схема процесса покрытия выборки такими сферами состоит в последовательном выделении подвыборок убывающих объемов, не оптимизирующих явным образом некую функцию качества. Однако такая конструкция может оказаться полезной исследователю, если будет определяться самой постановкой задачи.

1.5. Математические аспекты корректности и достаточности анализируемой выборки в задаче кластерного анализа

Остановимся на корректности полученных результатов. Оценим влияние неточности векторного представления для одного из данных. Можно предложить следующий вариант устойчивости решения. Пусть из R данных r-й имеет абсолютную погрешность δr. На первом шаге для двух ближайших векторов вводится координата их центра тяжести. Так перебираются все пары в R. Затем эта процедура повторяется снова. Конечный результат назовем первым шагом. В метрике L_1 после первого шага погрешность уменьшится до δr/4. При общем числе шагов k влияние погрешности уменьшится до значения $(\delta r/4)^k$.

Перейдем к вопросу влияния объема анализируемой выборки на результаты анализа. Он является наиболее острым для всех прикладных задач, использующих априорно неизвестные оценки плотностей распределения вероятностей анализируемых данных. Нетрудно показать, что в этом случае качество разделения гипотез будет улучшаться с объемом выборки как $R^{-1/3}$. Перейдем к постановке задачи.

Пусть имеется B гипотез $b \in 1,2,...B$. Всего проанализировано R реализаций $r \in 1,2,...R$. Рассмотрим b_i-ю гипотезу. Пусть с такой гипотезой имеется R_{bi} реализаций, тогда остальных реализаций $R - R_{bi}$. Введем следующие обозначения:

R_{bi} – количество реализаций с гипотезой b_i;

R_{0i} – количество реализаций с иными гипотезами.

Тогда:

$$R = R_{bi} \bigcup R_{0i}. \tag{1}$$

Все анализируемые реализации представим в виде n-мерного вектора $X_n(r)$. Тогда мы имеем следующие наборы векторов

$$X_n(r|bi);\ \ b=0,\ b_i. \tag{2}$$

Можно построить B отношений правдоподобий $\Lambda(r|bi)$ для гипотез (2). Выбор гипотезы будем осуществлять по максимальному значению отношения правдоподобия:

$$b_m \rightarrow Sup\Lambda(r|b_i). \tag{3}$$

Будем предполагать, что проекции вектора Xn для всех гипотез статистически независимы.

Рассмотрим способ построения вероятностной гистограммы $P_0(X_n)$. Рассмотрим проекцию X_1. Разобьем распределение $P_0(X_1)$ на K областей ξ:

$$\begin{aligned}
&\xi_1 \in \eta(C_1) - \eta(C_1 - \alpha_1),\\
&\xi_2 \in \eta(C_1 - \alpha_1) - \eta(C_1 - \alpha_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\xi_k \in \eta(C_1 - \alpha_{k-1}) - \eta(C_1 - \alpha_k),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\xi_K \in \eta(C_1 - \alpha_{k-1}) - \eta(C_1 - \alpha_K),
\end{aligned} \tag{4}$$

где: $\eta(...)$ – функция Хевисайда; $\alpha_1,...\alpha_k,...\alpha_K$ – точки внутри области определения финитной проекции X_1.

Координаты точек α выбираем таким образом, чтобы вероятность для проекции X_1 попасть в область ξ_k была равна:

$$P_0(\xi_k) = 1/K. \quad (5)$$

Таким образом, переход от переменной X_1 к переменной ξ позволил создать равномерное распределение вероятностей:

$$P_0(\xi_1) = 1/K\ [\eta(\xi_1) - \eta\ (\xi_1 - K)]. \quad (6)$$

Выражение (6) следует рассматривать как асимптотическое при $K \to \infty$. Аналогичным образом можно получить и одномерные распределения $P_0(\xi_2)$ и т.д.

Следует отметить, что плотности распределения вероятностей $P_b(\xi_1)$,... для любой гипотезы b не являются равномерными, хотя такая малая вероятность для хотя бы одной из переменных ξ имеется.

Полученные свойства распределений (6) и предположения об их статистической независимости позволяют получить следующее правило для оптимального обнаружения b-й гипотезы:

$$L_b = \sum_{k=1}^{k=K} \ln\left[P_b(\zeta_k)\right], \quad (7)$$

где L_b – статистика испытаний (логарифм отношения правдоподобия).

Выражение (7) получено с точностью до постоянного множителя.

Предлагаемый метод различия b-й гипотезы (4, 6, 5, 7) требует уточнения еще одного параметра, а именно величины K. Указанный параметр можно выбрать из следующих соображений. Пусть известны одномерные распределения для проекций векторов ξ_k. Эти распределения построены для выборки из R испытуемых. С увеличением величины K растет точность представления исходной плотности распределения вероятностей ε_1 в переменной ξ, которая равна:

$$\varepsilon_1 = 1/ K.$$

С другой стороны, с увеличением K падает среднеквадратичная погрешность ε_2 в каждом интервале ξ_k, которая равна:

$$\varepsilon_2 = [K/R]^{1/2}.$$

Минимум точности достигается при следующем оптимальном значении N:

$$\varepsilon = [\varepsilon_1^2 + \varepsilon_2^2] \Rightarrow Kопт = [2 R]^{1/3} . \quad (8)$$

Погрешность представления базисов векторов составит

$$\varepsilon_{min} \approx 2 R^{-1/3}. \quad (9)$$

Если B гипотез априорно должны хорошо различаться, то погрешность различения гипотез будет равна

$$\varepsilon \approx 2R^{-\frac{1}{3}}B^{-\frac{1}{2}}. \quad (10)$$

Указанный алгоритм проверялся на выборке R из 700 реализаций для исходного вектора размерностью $Z \in 1, ...12$, количества различаемых гипотез $B = 4$ и числа градаций разбиения исходных плотностей распределения вероятностей на $K = 5$ областей (с целью получения равномерных распределения для гипотезы H_0). Экспериментальная проверка показала, что вероятности пропуска и ложной тревоги не превышали значения 0,05, что хорошо согласуется с оценкой (10).

Для решения задач, не использующих данные о плотностях распределения вероятностей анализируемых событий, вряд ли стоит ожидать более быстрого критерия сходимости, чем $R^{-1/3}$. Поэтому для доказательства улучшения качества решения примерно на 30% автору необходимо было бы увеличить объем анализируемой выборки примерно на порядок и более. Для задач, рассматриваемых в данной книге, это достаточно сложно технически или вообще невозможно, например, в случае социально-экономических задач. Поэтому **в качестве критерия качест-**

ва следует принять совпадение результатов машинных расчетов с оценками специалистов-экспертов[1].

Результаты теоретических и практических исследований можно предложить как инструментальные средства для решения следующих задач:

- медицинской диагностики;
- анализа экономических данных;
- анализа данных социологических опросов;
- выявления центров отказов или сбоев в сложных технических системах;
- упорядочения баз данных по заданным критериям отбора[2].

[1] «Сделать разумный выбор всегда трудно... Выбор – это задача, которую нужно решить. Задачки из физики или математики – те, точные способы решения которых были найдены ранее, – не представляют трудности. Другое дело – задачи из гуманитарных сфер. Спокон веку правители, экономисты, педагоги находили их приближенные решения интуитивно, по наитию. И лишь сегодня наука – точная наука! – вплотную приблизилась к тому, чтобы искать и находить решения этих задач (у ученых есть замечательное выражение: «С точностью, достаточной для практического применения» – изящная формулировка!), пользуясь новейшим математическим аппаратом и методиками компьютерного моделирования» // Романов Ю. — Компьютерра. — № 01–02 от 13 января 2009 г.

[2] Аспекты корректности и достаточности анализируемой выборки в задаче кластерного анализа обсуждались С.А. Судаковым совместно с д-ром техн. наук Л.С. Чудновским, которым они и были окончательно сформулированы уже после кончины Станислава Арсеньевича.

Глава 2

Алгоритм Clust в медицинских задачах, связанных с психиатрией и клинической психологией

2.1. Феномен стигматизации психически больных

[Как известно, феномен стигмы («клейма») сопутствует остракизму. Стигма представляет собой комплекс стереотипных суждений, которые обосновывают особое положение изгоя. Ее носитель теряет для окружающих свои индивидуальные черты и становится воплощением описанных стигмой свойств. Отношение к нему в обществе становится предвзятым, нередко дискриминирующим. Длительное пребывание в подобной ситуации губительно действует на психику стигматизируемого человека, нередко доводя его до психической дезадаптации и самоубийства.

Очевидно, влияние стигмы в психиатрии является еще более ярким и пагубным. Его преодоление признано международным психиатрическим сообществом одним из приоритетных направлений охраны психического здоровья.

Как сказано выше, стигмой называется комплекс обывательских суждений о лицах с психическими расстройствами. Приписывание окружающими конкретному обладателю психиатрического диагноза своих представлений о психически больных в целом определяется как стигматизация. За самостигматизацию принята вся совокупность реакций пациента на свое заболевание и статус психически больного.

Изучение связанных со стигмой феноменов представляет большую сложность, поскольку они формируются при взаи-

модействии культурных, экономических, медицинских, психологических и многих иных факторов. Необходимость их одновременного учета при исследовании стигматизации и самостигматизации психически больных и определила применение С.А. Судаковым разработанных им методов Clust1 и Clust2. Речь идет об образовании качественно новых, вторичных и так далее признаков на основе исходных, заложенных в базу данных. Полученные признаки могут быть интерпретированы как объекты целостной системы. Применение кластеризации с жесткими (h) и мягкими (s) мерами близости, с заданным и произвольным числом кластеров, а также метода реперных решеток выявило разнообразные скрытые связи между исследуемыми параметрами. Это сделало возможным описать качественно новые функциональные компоненты изучаемых феноменов и характер их взаимодействия, что, в конечном итоге, привело к их пониманию как биопсихосоциальных систем и определению мишеней воздействия на них. — *Изд.*]

Исследование стигматизации в психиатрии было связано с выявлением стереотипов общественного сознания, касающихся психически больных и психиатрии. Оно производилось путем анкетирования респондентов с помощью опросника. Указанный опросник состоял из 140 утверждений, отражающих разнообразные суждения о психически больных, психических болезнях, психиатрах и психиатрии, как отрасли медицины. Опросник предполагал два варианта ответа: «верно» и «неверно», которым были присвоены коды 1 и 0 соответственно [58].

Первоначально проводился опрос студентов московских вузов – 90 мужчин и 58 женщин в возрасте 18–22 года – всего 148 человек.

Исходная матрица данных имела вид $\|x_{ij}\|$, $1 \le i \le 148$, $1 \le j \le 140$, где координаты 140-мерных векторов относятся к шкале наименований. Описанное выше при переходе к частотной форме сжатие матрицы данных порядка 148×140 до 2×140 переводит элементы из шкалы наименований в шкалу отношений (n_0, n_1), где n_0, n_1 — абсолютные частоты кодов 0 и 1.

Кластеризация 140 двумерных векторов (n_0, n_1) дает различные типы кластеров, описывающих суждения респондентов. Применение Clust1 к матрице 2×140 дает для s-меры на первом

уровне 44 кластера, на втором – 14, а на третьем – 5 (рис. 4). В последнем случае структурная матрица имеет вид

	1	2	3	4	5	
1	30					
2		34				
3			25			$\kappa \approx 0,97$
4				32		
5				3	16	
Σ	30	34	25	35	16	

Таким образом, было получено 5 кластеров, то есть групп суждений. Анализ их содержания позволил определить их как следующие социально-психологические компоненты стигматизации:

K_1 – доброжелательность и осторожность;
K_2 – отрицание медицинской модели болезни;
K_3 – излишняя категоричность и жестокость;
K_4 – гуманистическое отношение к больным;
K_5 – крайне негативное отношение к больным.

148×148 ↓ 2×140 ↓ ↓ 140×140 ↓ 44×44 ↓ 14×14 ↓ 5×5 ↓ 3×3 ↓ 2×2	база данных частотное описание данных по объектам уровень	$N_t/140$	ℓ_t	$N_t/140+\ell_t$	$\max N_t/140, \ell_t$	κ
	0					
	1	0,329	0,021	0,350	0,329	0,98
	2	0,100	0,056	0,156	0,100	0,98
	3	0,036	0,147	0,183	0,147	0,97
	4	0,021	0,308	0,329	0,308	0,96
	5	0,014	0,479	0,439	0,479	0,96
				$t^*=2$	$t^*=2$	

Рис. 4. Система уровневых кластеров в задаче (раздел 2.1)

На втором этапе исследования выборка респондентов была расширена и составляла 221 человек. В нее вошли представители разных профессиональных и возрастных групп.

Для группы в 221 человек с теми же 140 вопросами имеем:
Кластеры первого уровня содержат объемы:

1 – 4	16 – 4	31 – 2
2 – 6	17 – 2	32 – 3
3 – 3	18 – 2	33 – 4
4 – 2	19 – 2	34 – 4
5 – 2	20 – 5	35 – 2
6 – 3	21 – 3	36 – 5
7 – 3	22 – 3	37 – 4
8 – 2	23 – 2	38 – 5
9 – 2	24 – 2	39 – 3
10 – 4	25 – 2	40 – 3
11 – 2	26 – 3	41 – 2
12 – 4	27 – 4	42 – 2
13 – 4	28 – 5	43 – 2
14 – 4	29 – 2	44 – 3
15 – 2	30 – 4	45 – 2
		46 – 2

Кластеры второго уровня для h-меры:
1 = {1, 8, 23, 24};
2 = {4, 7, 10, 14};
3 = {12, 11, 29, 39};
4 = {17, 26};
5 = {21, 20, 25};
6 = {30, 31, 34};
7 = {2, 22};
8 = {46, 5, 16};
9 = {6, 15};
10 = {9, 41, 42};
11 = {18, 27};
12 = {38, 32, 35, 37};
13 = {13, 3, 36};
14 = {33, 28, 45};
15 = {19, 44};
16 = {40}.

Структурная матрица имеет вид:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	10																
2		13															
3			11														
4				5													
5					10							2					
6						10											
7							11										
8								8									$\kappa \approx 0,98$
9									5					1			
10										6							
11											6						
12												12					
13													12				
14														10			
15															5		
16																3	
Σ	10	13	11	5	10	10	11	8	5	6	6	14	12	11	5	3	

Кластеры третьего уровня для h-меры:
1 = {1, 6, 7};
2 = {8, 2, 10};
3 = {3, 4, 11, 14};
4 = {12, 5, 9, 13};
5 = {15, 16}.

Структурная матрица имеет вид:

	1	2	3	4	5	
1	31			3		
2		26	1			
3			32			$\kappa \approx 0,96$
4				38		
5		1			8	
Σ	31	27	33	41	8	

На четвертом уровне с h-мерой:

$$1=\{2, 3, 5\}$$
$$2=\{1,4\}$$

	1	2	
1	67	0	
2	1	72	$\kappa \approx 0,99$
Σ	68	72	

В целом уровневая схема имеет вид: 221×140 → 2×140 → → 140×140 → 46×46 → 16×16 → 5×5 → 2×2. Для нее минимумы функций $f_1(t)$ и $f_2(t)$ также достигаются в $t^* = 2$. Значения κ следующие:

$$\kappa_2 \approx 0,98;\ \kappa_3 \approx 0,96;\ \kappa_4 \approx 0,99.$$

Исследование явилось начальным вкладом в кандидатскую диссертацию Л.Я. Серебрийской по психологическим наукам. Результат исследования заключается в описании психосоциальных компонентов стигматизации психически больных обществом и их связи с возрастом, полом и профессией стигматизаторов [67].

2.2. Феномен самостигматизации психически больных

Самостигматизация психически больного человека является сложным феноменом, который представляет собой синтез как минимум трех составляющих. Первая – симптомы психического расстройства, вторая – приписывание человеком с психическим расстройством своих представлений о психически больных самому себе, и третья – реакция на все переживания, которые человек связывает с болезнью. Можно сказать, что если в случае стигматизации психически больных речь шла о феноменах общественного сознания, то в данном случае изучаются субъективные переживания больного.

Исследование самостигматизации проводилось с помощью опросника, включающего утверждения о роли психического расстройства в жизни больного и отражающие разнообразные представления о психически больных в целом.

Опросник состоял из 83 утверждений, на которые предполагалось 4 варианта ответа: «неверно», «скорее неверно», «скорее верно» и «верно».

На первом этапе работы с помощью описанного опросника из 83 утверждений (т.е. признаков) были обследованы 7 больных шизофренией, 20 – аффективными расстройствами и 31 – невротическими расстройствами, итого – 81 человек (объект) [59].

База данных выборки объема 81 описывается матрицей 81×83 $\|x_{ij}\|$, где x_{ij} – коды, принимающие значения: 3 – «верно», 2 – «скорее верно», 1 – «скорее неверно», 0 – «неверно».

Предлагается по каждому утверждению ограничиться набором абсолютных частот кодов 0, 1, 2, 3 по всем больным выборки: $n_0 + n_1 + n_2 + n_3 = 83$. Такое сжатие столбцов матрицы базы данных выявляет крупномасштабную картину распределения ответов на утверждения опросника по всем исследуемым больным. Вместо матрицы 81×83 при этом имеем матрицу 4×83, элементы которой x_{i0}, x_{i1}, x_{i2}, x_{i3}, в отличие от x_{ij}, принадлежавших шкале порядка, принадлежат теперь к шкале отношений. Применение к матрице 4×83 алгоритма Clust1 дает на первом уровне 19 кластеров со структурной матрицей:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	7																			
2		6																		
3			3																	
4				7																
5					5															
6						2														
7							5													
8								4												
9									3											
10										6										$\kappa \approx 0,92$
11											3									
12												3						2		
13													3							
14														3						
15												3			5					
16																3				
17																	3			
18																		4		
19																		1	2	
Σ	7	6	3	7	5	2	5	4	3	6	3	6	3	3	5	3	3	7	2	

На втором при s-мере близости – 6 кластеров, описываемых структурной матрицей:

	1	2	3	4	5	6
1	9	4				
2		33				
3			9			
4		1	1	6		
5					10	
6			1		2	7
Σ	9	38	11	6	12	7

$\kappa \approx 0,87$

Значение κ-критерия для нее составляет 0,87, что говорит о высоком соответствии оценок двух виртуальных экспертов.

Состав кластеров следующий:

1 = {11, 6, 8};
2 = {3, 2, 4, 5, 7, 9, 10, 16};
3 = {12, 15};
4 = {13, 14};
5 = {17, 1, 19};
6 = {18}.

Для третьего уровня при той же s-мере имеем 2 кластера:

1 = {3, 4, 5, 6};
2 = {1, 2}.

	1	2
1	35	10
2	1	37
Σ	36	47

$\kappa \approx 0,73$

Удобно назвать эту систему 6 кластеров второго уровня кластерами самостигматизации предъявленной выборки больных. Анализ содержания вошедших в кластеры утверждений показал, что каждый кластер отражает определенный аспект нозоцентрического (ориентированного на представление о болезни) изменения самосознания больного. В [59] эти кластеры получили следующие названия:

K_1 – «жертва стигмы»;
K_2 – деидентификация с психически здоровыми;
K_3 – готовность к идентификации с психически больными;
K_4 – дистанцирование от «психически больных»;
K_5 – оправдание болезнью;
K_6 – генерализация стигмы.

Система уровневых кластеров показана на рис. 5.

81×83 ↓ 4×83 ↓ 83×83 ↓ 19×19 ↓ 6×6 ↓ 2×2 ↓ 1×1	база данных уровень	$N_t/83$	ℓ_t	$N_t/83+\ell_t$	$\max(N_t/83, \ell_t)$	κ
	0					
	1	0,229	0,055	0,284	0,229	$\kappa_1 = 0,92$
	2	0,072	0,181	0,253	0,181	$\kappa_2 = 0,87$
	3	0,024	0,312	0,336	0,312	$\kappa_3 = 0,73$
	4			t*=2	t*=2	

Рис. 5. Система уровневых кластеров в задаче (раздел 2.2)

Такая система позволяет даже ввести некий балл самостигматизации, характеризующий состав кластера самостигматизации. В каждом кластере может быть выбрана «типичная» точка (его центр Чебышева). Баллом самостигматизации, принадлежащим шкале порядка, естественно назвать величину arg max (n^0_0, n^0_1, n^0_2, n^0_3), где абсолютные частоты, помеченные нулями, относятся к центру Чебышева рассматриваемого кластера.

В связи с недостаточностью информации для основных целей исследования на следующем его этапе была обследована более многочисленная группа больных из 129 человек [68]. Ее состав следующий: шизофрения – 69, аффективные расстройства – 29, невротические расстройства –31. Вопросник оставлен тот же. Применен тот же прием использования частотных описаний – вместо матрицы 129×83 исследуется редуцированная матрица 4×83.

На первом уровне получается 21 кластер, а на втором уровне при s-мере близости возникает 9 кластеров со структурной матрицей:

	1	2	3	4	5	6	7	8	9
1	9	1							
2		5			1				
3			7			1			
4		3		18					
5					8				
6	1					5			
7	1						11		
8								7	
9									5
Σ	11	9	7	18	9	6	11	7	5

Значение к-критерия для нее составляет 0,88, что говорит о высоком соответствии оценок двух виртуальных экспертов-диагностов. Эта система из 9 кластеров самостигматизации, которые, как и на первом этапе исследования, отражают различные аспекты нозоцентрического изменения самосознания. Но, в отличие от кластеров, полученных на первом этапе, данные кластеры оказались более целостными по смыслу. В [68] они получают следующие названия:

K_1 – переоценка самореализации;
K_2 – нарушение Я-идентичности;
K_3 – готовность категории «психически больных» в сфере трудовой активности;
K_4 – деидентификация от окружающих в социальной сфере
K_5 – дистанцирование от психически больных в сфере внутренней активности;
K_6 – готовность к дистанцированию от психически больных в социальной сфере;
K_7 – переоценка внутренней активности;
K_8 – принятие роли психически больного в сфере самореализации;
K_9 – «зеркальное Я психически больного» в сфере внутренней активности.

Наконец, на третьем уровне при той же s-мере близости возникает три кластера со структурной матрицей вида

	1	2	3
1	18	1	
2		36	
3	4	2	22
Σ	22	39	22

для которой к-критерий дает 0,87.

Анализ содержания утверждений, вошедших в кластеры третьего уровня, показал, что каждый из кластеров отражает определенную психологическую стратегию самостигматизации. В частности, первый кластер отражал преувеличение значимости собственного заболевания, второй, напротив, ее частичное игнорирование, а третий – приписывание связанных с заболеванием проблем влиянию окружающих. Указанные стратегии были обозначены как отдельные формы самостигматизации и получили следующие названия:

K_1 – аутопсихическая;

K_2 – компенсаторная;

K_3 – социореверсивная [68].

На четвертом уровне при s-мере получаем:

$1 = \{1, 3\}$;

$2 = \{2\}$;

	1	2	
1	42	5	
2	2	34	$\kappa = 0,88.$
	44	39	

Результат кластеризации четвертого уровня выявляет различную направленность стратегий самостигматизации. Первый кластер объединил два кластера третьего уровня, отражающих актуализацию болезни. Второй кластер содержит кластер третьего уровня, отражающий игнорирование болезни.

Таким образом, последовательная кластеризация ответов больных на утверждения опросника дала материал для создания психологической модели самостигматизации следующего вида:

Четвертый уровень кластеризации	**Третий уровень кластеризации**	**Второй уровень кластеризации**
Направленность стратегий самостигматизации	Стратегии самостигматизации	Реализация стратегии самостигматизации в конкретном аспекте самосознания
Игнорирование болезни	Компенсаторная	Дистанцирование от психически больных в сфере внутренней активности
		Готовность к дистанцированию от психически больных в социальной сфере
		Готовность категории «психически больных» в сфере трудовой активности
Актуализация болезни	Аутопсихическая	Переоценка самореализации
		Переоценка внутренней активности
	Социореверсивная	Нарушение Я-идентичности
		Деидентификация от окружающих в социальной сфере
		Принятие роли психически больного в сфере самореализации
		«Зеркальное Я психически больного» в сфере внутренней активности

Применение кластерного анализа на следующих этапах исследования позволило выбрать из большого массива данных наиболее значимые для формирования каждого кластера самостигматизации клинико-психологические и социальные факторы и определить характер их влияния. На основании полученных результатов была создана клинико-психологическая типология самостигматизации и определены основные мишени целевого воздействия при работе с больными.

Работа легла в основу диссертации И.И. Михайловой (канд. мед. наук). В целом проблема самостигматизации описана в работе [75].

	матрица данных частотное описание данных по объектам уровень	$N_t/83$	ℓ_t	$N_t/83+\ell_t$	$\max(N_t/83, \ell_t)$	κ
129×83						
$\downarrow$						
4×83						
$\downarrow$						
83×83	0					
$\downarrow$						
21×21	1	0,253	0,060	0,313	0,253	
$\downarrow$						
9×9	2	0,108	0,127	0,235	0,127	$\kappa_2 = 0,88$
$\downarrow$						
3×3	3	0,036	0,256	0,262	0,256	$\kappa_3 = 0,87$
$\downarrow$						
2×2	4	0,024	0,308	0,332	0,308	$\kappa_4 = 0,83$
$\downarrow$						
1×1				t*=2	t*=2	

Рис. 6. Система уровневых кластеров в задаче (раздел 2.2)

2.3. Феномен самостигматизации больных шизофренией

Шизофрения оказывает сильное влияние на личность и самосознание больного, что определяет необходимость отдельного исследования самостигматизации у лиц с данным заболеванием, ее связи с клиническими факторами. Исследование проводилось с использованием аналогичного пункту 2 методического аппарата и метода математической обработки информации.

Исследовалась выборка из 121 больного шизофренией. Опросник содержал те же 83 вопроса, что и для [59] (пункт 2) с теми же кодами ответов.

При кластеризации базы данных, редуцированной до 4×83 алгоритмом Clust1, на втором уровне, получается 6 кластеров при s-мере со структурной матрицей

	1	2	3	4	5	6
1	16					
2	1	15	1	1		
3	2		14			
4			3	5		
5		2			9	
6						4
Σ	19	17	18	15	10	4

$\kappa \approx 0,86$

Хотя к-критерий дает высокое качество, для разнесения системы центров Чебышева был применен алгоритм Clust2 с m = 6. Оказалось, что при этом оценка ошибки первого рода улучшается с $p \approx 0{,}0004$ до $p \approx 0{,}00007$ – почти на порядок.

В данном случае оценка величин p была получена с помощью таблиц сопряженности в каждой кластерной системе их центров Чебышева. Таким образом, получена система 6 кластеров, названных:

K_1 – деидентификация со здоровыми в социальной сфере;
K_2 – деидентификация с категорией здоровых в личностной сфере;
K_3 – готовность к самоидентификации в личностной сфере;
K_4 – готовность к принятию стигмы психически больного;
K_5 – самоидентификация с больными в профессиональной сфере;
K_6 – готовность категории «психически больных» как маргиналов.

Объемы этих 6 кластеров следующие: K_1 – 36; K_2 – 21; K_3 – 5; K_4 – 10; K_5 – 8; K_6 – 4.

Как и в случае [59], эти кластеры называются кластерами самостигматизации и отражают реализацию самостигматизации в определенном аспекте самосознания. Кластеры позволяют ввести в каждом из них балл самостигматизации, который определяется так, как было показано выше [70].

На следующем, не описанном здесь этапе исследования с помощью кластеризации больных (объектов) по клиническим признакам с учетом их самостигматизации были выявлены и описаны 5 клинико-психологических типов самостигматизации больных шизофренией, отличающиеся друг от друга как психологическими характеристиками самостигматизации, так и характером ее связи с заболеванием.

Результатом проведенного исследования стала разработка клинической типологии самостигматизации больных, страдающих шизофренией. Типология имеет прикладное значение, позволяя разработать дифференцированный подход к социализации различных категорий пациентов, а также повысить точность клинической диагностики и прогноза заболевания.

Результаты были использованы при защите кандидатской диссертации О.А. Гонжал по медицинским наукам.

Обобщение результатов трех описанных здесь исследований стигматизации и самостигматизации психически больных позволило создать биопсихосоциальную модель системы стигмы в психиатрии, которая может служить основанием для разработки дестигматизационного направления.

2.4. Кластерный анализ нейрофизиологических маркеров когнитивных функций у больных шизофренией и шизоаффективным психозом

В рамках научной работы лаборатории нейрофизиологии НЦПЗ РАМН была поставлена цель выделения нейрофизиологических процессов, наиболее тесно связанных с патогенезом шизофрении.

В качестве базового электрофизиологического метода была выбрана регистрация слуховых вызванных потенциалов (ВП) в парадигме избирательного внимания («oddball»). Причиной стала высокая информативность метода для объективной количественной оценки отклонений в процессах избирательного внимания, памяти, определения значимости поступающей информации [95, 96] – процессов, являющихся одними из основных психопатологических характеристик шизофрении [97].

Описанный выше метод был применен при обследовании больших групп больных, в том числе и во время первого приступа эндогенного психоза, и их родственников первой степени родства, также в ходе совместных исследований с подразделениями НЦПЗ и ряда других научных учреждений нейрофизиологические данные сопоставлялись с результатами иммунологического, молекулярно-генетического, структурно-морфологического, психологического анализов больных шизофренией. Полученные данные позволили заключить, что маркерами процессов, наиболее близких к патогенезу шизофрении и проявляющихся еще до манифестации заболевания, являются снижение амплитуды волн N100 (ВП на незначимый стимул) и P300 слуховых ВП, зарегистрированных в парадигме «oddball», хотя развитие и течение болезни сопровождается

и другими сложными, иногда разнонаправленными изменениями электрофизиологических характеристик [98, 99].

Очевидно, что подобные маркеры имеют высокий потенциал для практического применения в клинике шизофрении, а одной из задач, которые должны быть решены в этой связи, является определение структуры аномалий волн слуховых ВП и ее вариабельности в популяции психически больных людей. Именно эта проблема легла в основу исследования, описанного ниже.

Было высказано предположение, что индивидуально специфичные аномалии параметров волн слуховых ВП и их динамика в процессе лечения ассоциируются с особенностями патопсихологической симптоматики у данного больного и таким образом можно ожидать, что выделенные с помощью кластерного анализа подгруппы больных со сходными нейрофизиологическими аномалиями будут различаться и по ряду аспектов клинической картины болезни.

Следует отметить, что работа, проведенная в 2002–2003 гг. (и опубликованная в 2003 г. [57]), являлась по сути лишь первым шагом – анализ проводился в относительно небольших выборках испытуемых.

Для математического анализа были представлены нейрофизиологические показатели волн слуховых ВП, зарегистрированных в парадигме «oddball». Группа больных включала 16 больных шизофренией, 6 – шизоаффективным психозом [57].

В качестве основного параметра была выбрана амплитуда волны Р300 слуховых ВП как показатель, наиболее информативный для оценки когнитивных аномалий у больных эндогенными психозами [98, 100].

Объектом анализа были величины разности амплитуд Р300, зарегистрированных у больных в течение первой недели после поступления в клинику и повторно – перед выпиской на фоне значительной психопатологической симптоматики, а матрицы базы данных формировались в виде $X^1 = \|x1_{ij}\|$ $1 \le j \le 12$, поскольку использовались данные, зарегистрированные в 12 отведениях ЭЭГ.

Всего рассматривалось три варианта кластеризации:

1. На первом уровне образуется 6 кластеров объектов с диагональной структурной матрицей – (4, 4, 4, 5, 3, 1). На втором

уровне, как при h-, так и при s-мере близости, получается два кластера со структурной матрицей:

	1	2
1	18	
2	1	3
Σ	19	3

$\kappa \approx 0,95$

$1 = \{1, 2, 3, 4, 6\}$
$2 = \{5\}$

2. На первом уровне образуется 7 кластеров объектов с диагональной структурной матрицей – (4, 4, 3, 7, 1, 2, 1) и двумя кластерами на втором уровне.

Для h-меры:

	1	2
1	20	
2	1	1
Σ	21	1

$\kappa \approx 0,95$

$1 = \{1, 2, 3, 4, 5, 6\}$
$2 = \{7\}$

Для s-меры:

	1	2
1	19	–
2	–	3
Σ	19	3

$\kappa = 1$

$1 = \{1, 2, 3, 4, 5\}$
$2 = \{6, 7\}$

3. На первом уровне образуется 8 кластеров с диагональной структурной матрицей – (4, 2, 4, 2, 2, 2, 4, 2). На втором уровне образуется три кластера и для h-меры и для s-меры:

	1	2	3
1	8	1	
2		8	
3		3	2
Σ	8	12	2

$\kappa \approx 0,73$

$1 = \{2, 3, 5\}$
$2 = \{1, 4, 6, 7\}$
$3 = \{8\}$

На третьем уровне для h- и s-меры получается:

	1	2		
1	12	0	$\kappa = 0,82$	$1 = \{2,3\}$
2	2	8		$2 = \{1\}$
Σ	14	8		

Системы уровневых кластеров показаны на рис. 7, 8, 9.

22×12 ↓	база данных					
22×22 ↓	уровень – 0	$N_t / 22$	ℓ_t	$N_t / 22 + \ell_t$	$\max(N_t / 22, \ell_t)$	κ
6×6 ↓	1	0,273	0,151	0,424	0,273	$\kappa_1 = 1$
2×2 ↓	2	0,091	0,236	0,327	0,236	$\kappa_2 = 0,95$
1×1	3			t* = 2	t* = 2	

Рис. 7. Система уровневых кластеров в задаче (вариант 1)

22×12 ↓	база данных					
22×22 ↓	уровень – 0	$N_t / 22$	ℓ_t	$N_t / 22 + \ell_t$	$\max(N_t / 22, \ell_t)$	κ
7×7 ↓	1	0,318	0,095	0,414	0,318	$\kappa_1 = 1$
2×2 ↓	2	0,091	0,245	0,336	0,245	$\kappa_2 = 0,95$
1×1	3			t* = 2	t* = 2	

Рис. 8. Система уровневых кластеров в задаче (вариант 2)

22×12 ↓	база данных					
22×22 ↓	уровень – 0	$N_t/22$	ℓ_t	$N_t/22 + \ell_t$	$\max(N_t/22, \ell_t)$	κ
8×8 ↓	1	0,364	0,127	0,491	0,364	$\kappa_1 = 1$
3×3 ↓	2	0,136	0,236	0,372	0,236	$\kappa_2 = 0,73$
2×2 ↓	3	0,091	0,337	0,428	0,337	$\kappa_3 = 0,82$
1×1				t* = 2	t* = 2	

Рис. 9. Система уровневых кластеров в задаче (вариант 3)

Для дальнейшего анализа был выбран первый вариант кластеризации, включавший наименьшее число исходных кластеров.

Статистическое сопоставление (по t-критерию Стьюдента) средних величин нейрофизиологических показателей каждого кластера дало величины достоверности их различий в диапазоне от 10^{-7} до 0,01. Эти различия проиллюстрированы на рис. 10 (цв. вклейка).

Из рисунка видно, что в результате применения кластерного анализа группа больных разделилась на тех, у кого ко второму обследованию волна P300 нарастала (кластеры I–III), уменьшалась (кластеры V–VI) или имела сложный разнонаправленный характер изменений (кластер V). Предварительное сопоставление с клиническими оценками по шкале PANSS показало согласованность выявленных нейрофизиологических кластеров с динамикой суммарного балла по субшкале позитивных расстройств [57].

Следующий шаг исследования мог быть сделан только после набора значительно большей по объему выборки больных. Ожидалось, что в этой новой выборке будет, во-первых, подтверждена устойчивость выявленных кластеров, а во-вторых, выделенные по результатам математического анализа подгруппы больных будут сопоставлены друг с другом относительно широкого диапазона клинических характеристик. К сожалению, по объективным причинам, данная работа так и не была продолжена.

2.5. Кластеры в группе больных эндогенными манифестными психозами юношеского возраста

Одними из самых распространенных вариантов аффективной патологии являются юношеские эндогенные депрессии с преобладанием расстройств в познавательных процессах (память, внимание, мышление и др.). Они выявляются преимущественно у учащейся молодежи, могут возникать как при шизофрении, так и при циклотимии, длительно остаются нераспознанными и приводят к выраженным ограничениям в учебной деятельности и в целом к социальной дезадаптации.

В целях определения структуры когнитивных расстройств, ее зависимости от характера заболевания и его психопатологической картины было предпринято нейропсихологическое исследование высших психических функций у двух названных выше групп больных.

Одна из задач исследования состояла в разработке шкалы количественной оценки степени выраженности нарушений различных параметров психических функций. Создание такой шкалы было продиктовано не только необходимостью сравнительного анализа симптомов когнитивных расстройств и их конфигурации в синдром при шизофрении и циклотимии (в том числе и в сопоставлении с возрастной нормой), но и установления связи этих расстройств с дисфункцией определенных зон мозга.

В разработке шкалы, нормировании показателей нарушений психических функций в баллах, в компьютерной обработке данных применялись статистические методы анализа сопряженности признаков.

Матрица «больной×признак» имеет объем 75×68. Распределение признаков по шкалам измерения данных следующее: для шкалы наименований – 34, для шкалы порядка – 30, для шкалы отношений – 4.

Признаки Н-шкалы имеют вид (0, 1) – отсутствие или наличие некоторого состояния у больного.

Признаки П-шкалы имеют вид (0, 1, 2) – отсутствие, слабая и сильная выраженность некоторого состояния у больного.

Это типичный случай употребления Н- и П-шкал в психиатрических описаниях состояний больных.

На первом уровне алгоритм Clust1 дает 23 кластера и следующее распределение чисел больных по 23 кластерам – {4, 5, 2, 4, 8, 10, 4, 4, 3, 2 , 3, 4, 3, 2, 4, 4, 4, 2, 5, 1, 2, 1, 2}.

На втором уровне с h-мерой межкластерной близости получается 5 кластеров:

1 = {2, 1, 5, 7, 10, 13, 20, 22};
2 = {3, 4, 6, 8, 15};
3 = {17, 9, 14, 16, 19};
4 = {12, 11, 18, 23};
5 = {21}.

Ее структурная матрица есть

	1	2	3	4	5
1	27				
2		8	2		
3		8	16	2	
4				9	
5	1				2
	28	16	18	11	2

$\kappa \approx 0,78$

На третьем уровне с h-мерой структурная матрица имеет вид:

	1	2
1	48	0
2	25	2
Σ	73	2

$1 = \{3, 1, 2, 4\}$
$2 = \{5\}$
$\kappa \approx 0,33$

Система уровневых кластеров показана на рис. 11.

75×68 ↓ 75×75 ↓	База данных Уровень — 0	$N_t/75$	ℓ_t	$N_t/75 + \ell_t$	$\max(N_t/75, \ell_t)$	κ
23×23 ↓	1	0,307	0,075	0,382	0,307	
5×5 ↓	2	0,067	0,095	0,162	0,095	$\kappa_2 = 0,78$
3×3 ↓	3	0,040	0,115	0,155	0,115	$\kappa_3 = 0,89$
2×2 ↓	4	0,027	0,115	0,142	0,115	$\kappa_4 = 0,33$
1×1	5			t* = 4	t* = 2	

Рис. 11. Система уровневых кластеров в задаче (раздел 2.5)

На втором уровне с s-мерой межкластерной близости получаются 3 кластера:

1 = {1, 2, 5, 7, 10, 13, 20, 21};
2 = {3, 4, 6, 8, 9, 11, 12, 14, 15, 17, 18, 19, 23};
3 = {16, 22}.

Структурная матрица имеет вид:

	1	2	3
1	27	1	
2		37	
3	2	3	5
Σ	29	41	5

$\kappa \approx 0,89$

В обоих случаях достигается высокий уровень соответствия оценок двух виртуальных экспертов [61, 73].

На третьем уровне структурная матрица имеет вид:

	1	2
1	33	0
2	1	41
Σ	34	41

$\kappa \approx 0,97$

Таким образом, были получены данные о специфичности «профилей» когнитивных нарушений в зависимости от нозологической принадлежности заболевания. Эти результаты имеют принципиальное значение для понимания природы изучаемых расстройств и расширяют возможности в решении дифференциально-диагностических задач при обследовании эндогенными манифестными психозами юношеского возраста.

Автор благодарит А.А. Кузюкову за любезно предоставленный материал.

2.6. Кластеры в задаче исследования доманифестного периода приступообразной шизофрении

Любое психиатрическое исследование, посвященное изучению клиники и психопатологии тех или иных состояний, сталкивается с проблемой объективной оценки полученных результатов. В этом отношении не стало исключением и проведенное нами исследование доманифестных проявлений шизофрении.

Использованный в данном исследовании алгоритм Clust1 исключил поверхностность и субъективизм в анализе полученных данных и в то же время позволил избежать грубой формализации и схематизма, которые зачастую прослеживаются при некорректном использовании современных шкал и опросников.

В начале исследования обследовалось 50 больных по 59 признакам [61]. По принадлежности их к шкалам измерения данных имелась следующая картина: шкала наименований с (0, 1) – 42 признака, шкала порядка с (0, 1, 2) и (0, 1, 2, 3, 4) – 15 признаков и шкала отношений – 2 признака (из 59 признаков было оставлено 49). Алгоритм Clust1 при кластеризации больных дал на первом уровне 19 кластеров.

Структурная матрица первого уровня имеет вид:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	2																		
2		3																	
3			2																
4				2															
5					2		1												
6						2													
7							3												
8								2											
9									3										
10										2									
11											4								
12												4							
13													2						
14						1								4					
15															3				
16																3			
17																	2		
18																		2	
19																			1
Σ	2	3	2	2	2	3	4	2	3	2	4	4	2	4	3	3	2	2	1

На втором уровне при h-мере близости получилось 4 кластера с числами больных (10, 18, 20, 2).

Структурная матрица этой системы имеет вид:

	1	2	3	4
1	10	6	1	
2		10	2	
3		2	16	
4			1	2
Σ	10	18	20	2

$\kappa \approx 0,66$

$$K_1 = \{7, 3, 12\}$$
$$K_2 = \{11, 5, 6, 9, 10, 14\}$$
$$K_3 = \{15, 1, 2, 8, 13, 16, 17, 18, 19\}$$
$$K_4 = \{4\}$$

Клинически результат был оценен как вполне рациональная группировка.

Клинико-психопатологически два первых кластера объектов имели на доманифестном этапе непсихотическую структуру. Доманифестные проявления в первом кластере исчерпывались нарушениями исключительно аффективного регистра, во втором кластере отмечалась тенденция усложнения доманифестной аффективной симптоматики за счет присоединения неврозоподобных и деперсонализационных расстройств. Третий кластер содержал клинические наблюдения, при которых в структуре инициальных аффективных расстройств эндогенной природы, задолго до манифестного психоза, регистрировались единичные феномены галлюцинаторно-бредового регистра.

У пациентов, попавших в четвертый кластер, первые психотические сенсорные расстройства развивались на фоне диффузной симптоматики с отдельными чертами олигофреноподобных и подобных органическим изменений личности. Этот последний кластер, состоящий из двух больных, показывает своего рода «маргинальность» содержимого, так как эти больные принадлежат к некому особому типу по сравнению с остальными 48.

Расширенное исследование описывалось матрицей «больной×признак» объема 60×68 (34 мужчины и 26 женщин в возрасте от 16 до 55 лет). В ней были особо подчеркнуты доманифестный и манифестный составы признаков – 68 признаков,

которые были сгруппированы в следующие блоки: социально-демографический (6), личностный (8), симптомы доманифестного периода (33) и симптомы манифестного периода (21).

Возраст больного и возраст начала заболевания были исключены из базы данных.

Система уровневых кластеров показана на рис. 12.

Из 54 признаков-симптомов 52 относятся к шкале наименований, а 2 – к шкале порядка.

С помощью программы Clust1 была получена следующая структура второго уровня кластеризации больных для s-меры близости:

	1	2	3	4	5	
1	15		3			
2		18	1			
3	2		12			$\kappa \approx 0,9$
4	1			7		
5					1	
Σ	18	18	16	7	1	

60×68 $\downarrow$	база данных					
60×60 $\downarrow$	уровень – 0	$N_t/60$	ℓ_t	$N_t/60+\ell_t$	$\max(N_t/60, \ell_t)$	κ
18×18 $\downarrow$	1	0,300	0,115	0,415	0,300	$\kappa_1 = 1$
5×5 $\downarrow$	2	0,083	0,215	0,298	0,215	$\kappa_2 = 0,83$
2×2 $\downarrow$	3	0,033	0,320	0,353	0,320	$\kappa_3 = 1$
1×1	4			$t^* = 2$	$t^* = 2$	

Рис. 12. Система уровневых кластеров в задаче раздела 2.6, расширенное исследование

На первом уровне получено 18 кластеров с диагональной структурной матрицей ($\kappa = 1$).

На втором уровне при h-мере получено 5 кластеров:

1 = {2, 1, 4, 12, 13};

2 = {6, 5, 9, 16};

3 = {3, 8, 10, 15, 17};
4 = {7, 11, 4};
5 = {18}.

	1	2	3	4	5	
1	12		2			
2	2	14	1			
3	2		17			$\kappa = 0,83$
4	1			8		
5					1	
Σ	17	14	20	8	1	

На третьем уровне получается для h-меры 2 кластера:

	1	2		
1	59		$\kappa = 1$	$1 = \{1, 2, 3, 4\}$
2		1		$2 = \{5\}$
Σ	59	1		

На третьем уровне для s-меры получаются 2 кластера:

	1	2		
1	59		$\kappa = 1$	$1 = \{1, 2, 3, 4\}$
2		1		$2 = \{5\}$
Σ	59	1		

Пятый кластер второго уровня состоит из одного больного, отделившегося в клиническом смысле правильно, как некий переходный случай. На третьем уровне оказалось, что все больные, кроме указанного, собрались в один кластер, а он по-прежнему отделяется. Это можно считать глубокой рациональностью устройства алгоритма Clust1.

Кластеризация 54 признаков дала следующий результат. На первом уровне получено 12 кластеров с объемами (9, 15, 2, 3, 8, 5, 3, 2, 1, 2, 2), на втором – 2 кластера с объемами (52, 2). При этом в первый кластер попали все признаки шкалы наименований, а во второй – два признака из шкалы порядка. Каппа-критерий для обоих уровней дает наивысшую оценку – 1.

Пример использования алгоритма Clust3 для базы данных 50×59

Среди 59 признаков: 42 из Н-шкалы, 15 из П-шкалы и 2 из О-шкалы, $\lambda_н = \lambda_п = 0{,}3$, $\lambda_о = 0{,}4$.

Зависимость объемов кластеров от r следующая:

R	N(r)
0,011	47
0,051	30
0,08	12
0,085	8
0,09	5
0,095	5
0,1	4

Для случая r = 0,085 имеем структурную матрицу:

	1	2	3	4	5	6	7	8
1	12							
2		11						
3			2	1	1			
4				2				
5					3			
6						6		
7							2	
8								5
Σ	12	11	2	3	4	6	2	5

$\kappa \approx 0{,}84$

Для 57 признаков, из которых 46 принадлежат Н-шкале и 17 – П-шкале, получаем:

R	N(r)
0,1	17
0,15	6
0,2	3

Для $r \approx 0{,}15$ структурная матрица имеет диагональный вид (37, 3, 2, 4, 1, 3).

Автор благодарит Е.В. Андриенко за любезно предоставленный материал для совместного исследования.

2.7. Кластерный анализ в оценке когнитивного дизонтогенеза при шизофрении в детском возрасте

При проведении экспериментально-психологических диагностических и динамических исследований в детской психиатрической клинике материал традиционно характеризуют такие особенности выборки, как большое количество «учетных» признаков, относительная малочисленность изученных испытуемых, наличие случаев «ненаблюдения» признаков у части испытуемых [61, 62]. Причины такого непростого положения с полнотой базы данных в исследованиях по детской клинической психологии (патопсихологии) кроются в сложности и нестабильности состояния испытуемых, возрастном разнообразии. Свои трудности связаны и с проблемой диагностики, в частности, наличием случаев с дифференцированным диагнозом, реабилитационными моментами и т.п. Существенную помощь в научной обработке данных при таких обстоятельствах играет грамотное применение математических данных, позволяющее получить не только количественную оценку полученного массива данных, но и наметить пути дальнейшего теоретического анализа.

Прекрасный образец такого сотрудничества специалистов представляет собой работа по оценке – с применением кластерного анализа – проявлений когнитивного дизонтогенеза (диссоциации развития) на примере квалификации гармоничности развития мышления и восприятия у больных шизофренией детей.

Изучение особенностей психического дизонтогенеза при шизофрении представляет одну из традиционных и остающихся актуальными задач современной клинической психологии. Решение такого типа задач требует новых подходов не только с точки зрения теории нарушенного, аномального развития, технологии проведения экспериментально-психологического исследования, но и с точки зрения возможностей применения аппарата математического анализа. В частности, в разработке задач по психологической ква-

лификации когнитивного дизонтогенеза при шизофрении, начавшейся в детском возрасте, использован кластерный анализ.

На базе отдела по изучению проблем детской психиатрии НЦПЗ РАМН (руководитель – д-р мед. наук, проф. И.А. Козлова) обследовалось 85 детей (61 мальчик и 24 девочки с диагнозами шизофрения, детский тип, и шизотипическое расстройство, исследование проведено в 2001–2006 гг.).

Гармоничность когнитивного развития изучалась на примере соотнесения операционных компонентов мыслительной и перцептивной (зрительное восприятие) деятельности. Использовались методики из экспериментально-психологического комплекса, направленного на оценку своеобразия познавательного развития детей при шизофрении. Данный комплекс разработан в лаборатории патопсихологии НЦПЗ РАМН и включает набор методик для оценки операционного аспекта познавательной деятельности и характеристик ее избирательности (с позиций предметно-содержательных составляющих). Для анализа на данном этапе работы был выделен операционный аспект (в мышлении и восприятии). Качественно оценивалось выполнение 3–4 методик. Анализировались только случаи болезни, так как каждый результат испытуемого сравнивался с возрастной нормой выполнения. Операционный компонент оценивался в баллах для мышления и восприятия отдельно на основании качественного и количественного анализа выполнения соответствующих методик, направленных на изучение формирования особенностей познавательной деятельности больных шизофренией детей.

Итак, два компонента операционной деятельности – мыслительный и перцептивный – оценивали в баллах x_1 и x_2. Обе оценки относятся к шкале порядка и принимают целые значения от –2 до +2. Оценки соответствуют:

0 – норме возраста испытуемого;

–/+ 1 – умеренно выраженному отставанию/опережению;

–/+ 2 – сильно выраженному отставанию/опережению по сравнению с нормативной группой.

База данных описывается матрицей $\|x_{ij}\|$ объемом 85×5, соответствующей целым точкам квадрата $[-2,+2]^2$. Двумерная (x_1, x_2) выборка для 85 детей изображена на рис. 13.

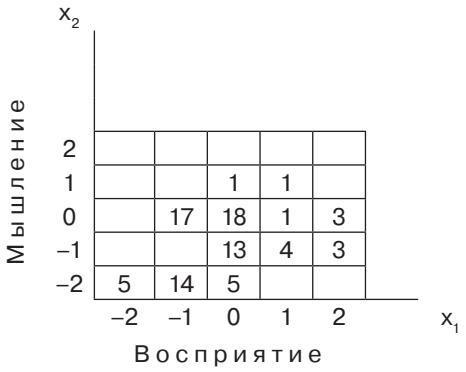

Рис. 13. Выборка (x_1, x_2) для 85 детей

Применение алгоритма Clust1 дало следующие результаты при кластеризации больных детей.

На первом уровне получается 9 кластеров (рис. 14) со структурной диагональной матрицей

[17, 19, 5, 13, 4, 3, 5, 14, 5].

На втором уровне при h- и s-мере возникают 5 кластеров, с диагональной структурной матрицей (к = 1) [5, 20, 19, 24, 17]. Их номера указаны на рис. 15 в левой части каждого кластера, причем левый нижний край малых квадратов соответствует точке (x_1, x_2).

Для третьего уровня при s-мере близости остается 2 кластера, также с диагональной структурной матрицей (к = 1): [44, 41], а при h-мере со структурной матрицей

	1	2
1	44	17
2	5	19
Σ	49	36

Система уровневых кластеров показана на рис. 16.

Рис. 14. Кластеры уровня 1 для группы из 85 детей

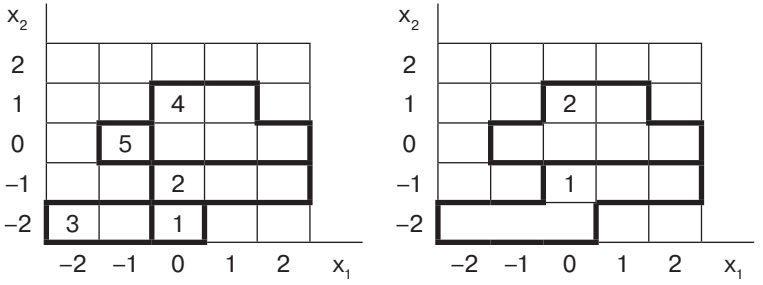

Рис. 15. Кластеры 2-го и 3-го уровней для группы из 85 детей

85×2 ↓	база данных					
85×85 ↓	уровень – 0	$N_t/85$	ℓ_t	$N_t/85+\ell_t$	$\max(N_t/85, \ell_t)$	κ
9×9 ↓	1	0,106	0,156	0,262	0,156	
5×5 ↓	2	0,059	0,282	0,341	0,282	$\kappa_2 = 1$
2×2 ↓	3	0,024	0,480	0,504	0,480	$\kappa_3 = 1$
1×1	4			$t^* = 1$	$t^* = 1$	

Рис. 16. Система уровневых кластеров в задаче (раздел 2.7)

Рассмотрим результат второго уровня иерархии, на котором получено 5 кластеров (рис. 15, слева), структурная матрица системы 5 кластеров имеет диагональный вид, а значение критерия κ равно 1. Содержательно это можно интерпретировать как разные варианты рассогласованного уровня развития мышления и восприятия у данных испытуемых – следует думать о вариативном проявлении диссоциации операционного компонента мышления и восприятия. Подобный вывод делался в предыдущих исследованиях только на основании сравнения средних показателей, без выделения различных видов рассогласования уровня развития операционного компонента мышления и восприятия [93].

Рассмотрим третий уровень иерархии, на котором получено всего 2 кластера, объемы которых составляют: 1 – 44 испытуемых, 2 – 41 испытуемый. Структура кластеров третьего уровня интерпретируется как два основных проявления когнитивного дизонтогенеза – преимущественно искаженного (кластер 2) и преимущественно задержанного типов (кластер 1). Эти термины на данном этапе – рабочие, их содержание, возможно, тесно связано с описанными клиницистами типами дизонтогенеза при шизофрении в детском возрасте – задержанный и искаженный типы (по [94]).

Дальнейшая работа над материалом дает основание полагать, что испытуемых, вошедших в каждый из кластеров третьего уровня иерархии, будут различать параметры, прямо не связанные с балльной оценкой состояния операционного компонента мышления и зрительного восприятия. К таким параметрам можно отнести длительность заболевания, время его начала, а также степень прогредиентности. В настоящее время разрабатывается схема качественной и количественной оценки факторов заболевания при оценке проявлений когнитивного дизонтогенеза при шизофрении у детей.

Таким образом, квалификация когнитивного дизонтогенеза при шизофрении в детском возрасте с применением аппарата кластерного анализа открывает новые возможности в количественной и качественной оценке разных параметров психического дизонтогенеза.

Результаты получены совместно с Н.В. Зверевой [92].

2.8. Кластеры в группе больных инфарктом мозжечка

В последние годы благодаря развитию новых технологий (КТ, МРТ, ПЭТ и др.) в поле зрения исследователей оказались структуры мозга, ранее трудно доступные для оценки их функционального состояния и роли в познавательной (когнитивной) деятельности человека. К таким «новым» для нейронауки структурам относится мозжечок, значение которого до настоящего времени рассматривалось большей частью в контексте координации движений.

В середине 90-х годов появились гипотезы о роли мозжечка в мышлении, речи и других психических процессах. В формировании и проверке этих гипотез особое место занимает нейропсихология, изучающая участие различных мозговых структур в обеспечении тех или иных составляющих психики и поведения. При этом нейропсихология обращается к заболеваниям мозга с соответствующими повреждениями его определенных зон.

Наиболее адекватной и корректной клинической моделью для изучения мозжечка являются его изолированные инфаркты. Пациенты с верифицированной локализацией инфаркта мозжечка встречаются крайне редко. В связи с этим количество обследованных пациентов, как правило, не достигает значений, необходимых для применения стандартных программных пакетов статистической обработки данных. Кроме того, не все больные могут быть обследованы по полной программе из-за их состояния на момент обследования. В этих случаях имеют место пропуски по целому ряду признаков (симптомов), которые выделены для нейропсихологической оценки состояния психологических функций. Это также затрудняет статистическую обработку данных.

Именно такая ситуация возникла при исследовании 25 больных с инфарктами мозжечка, находившихся на лечении в НИИ неврологии РАМН [62, 69].

Число признаков (нейропсихологических негативных симптомов) – 38. Степени их выраженности являются баллами в шкале порядка, кодируемыми как 0, 1, 2, 3. Общая база, опи-

сываемая матрицей $\|x_{ij}\|$ объема 25×38, расщепляется из-за особенностей неполноты информации о части больных на две, соответственно с объемами 18×38 и 22×33.

Примененный алгоритм Clust1 дал следующие результаты на втором уровне:

1) При частотном сжатии выборки больных до (n_0, n_1, n_2, n_3) для кластеров признаков получено 3 кластера объемом (22, 15, 1) для 18×38 и (15, 15, 3) для 22×33.

2) Аналогично для кластеров больных получено 3 кластера объемов (6, 8, 4) для 18×38 и 3 кластера (11, 9, 2) для 22×33.

3) При описании вида ($n_0 + n_1$, $n_2 + n_3$) для кластеров признаков получается при (22, 33) (11, 19, 3), а для кластеров больных — (15, 7).

4) При кластеризации 38 нейропсихологических признаков получено 7 кластеров первого уровня.

Для второго уровня получено при h-мере 2 кластера:

1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38};

2 = {37}.

Для третьего уровня:

	1	2	
1	36		$\kappa \approx 0,95$
2	1	1	
Σ	37	1	

Для s-меры получено на втором уровне 3 кластера:

1 = {1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35};

2 = {5, 6, 13, 21, 26, 36};

3 = {37}.

Для третьего уровня:

	1	2	3	
1	29			
2	2	6		$\kappa \approx 0,92$
3			1	
Σ	31	6	1	

Таким образом, благодаря применению процедуры кластеризации удалось решить следующую задачу – показать, как группируются симптомы нарушения психических функций при локализации инфаркта в правом или левом полушарии мозжечка, и на основании этих данных сделать заключение о собственном вкладе его в динамическую интегрированную организацию программирования и контроля деятельности человека с приоритетом правого полушария мозжечка в познавательной сфере.

Совместная работа положена в основу кандидатской диссертации по психологическим наукам Ю.В. Зуевой [69].

2.9. Кластеры объектов и признаков в группе больных рассеянным склерозом

Рассеянный склероз (РС) – распространенное инвалидизирующее неврологическое заболевание лиц преимущественно молодого возраста, при котором наблюдается многоочаговое поражение белого вещества ЦНС вследствие демиелинизации нервных волокон [79].

Нарушения исполнительных (или управляющих) функций (executive functions), возникающие у больных с когнитивным дефицитом (40–60% больных), могут иметь вторичный характер, как следствие снижения активационного обеспечения психической деятельности. Но в некоторых случаях нарушения произвольной регуляции являются первичными с преобладанием трудностей программирования, контроля, снижением регулирующей функции речи [80].

Актуальной представляется проблема влияния изменений психического функционирования, и в частности когнитивных и регуляторных функций, на реальную жизнь и поведение больного. Внимание исследователей направлено на изучение качества жизни, адаптационных возможностей, стратегий совладающего поведения в трудных жизненных ситуациях (копинг-стратегий) [81].

В данной работе была сделана попытка рассмотрения совладающего поведения как сложной функциональной системы, одной из составляющих которой является произвольная регуляция и контроль психической деятельности.

Проведенное нами исследование имело своей целью:

1) анализ регуляторных функций у больных РС с выделением больных со сходными симптомокомплексами нарушений произвольной регуляции;

2) выявление особенностей совладающего поведения в зависимости от специфики произвольной регуляции с учетом параметров когнитивной дисфункции, а также показателей агрессии.

Исследование проводилось в ГКБ № 11, являющейся базой Московского Центра рассеянного склероза.

Обследовано 26 больных (13 мужчин и 13 женщин), в возрасте от 17 до 47 лет, с достоверным рассеянным склерозом по критериям C. Poser [82]. Методическим инструментарием послужили методики общего нейропсихологического обследования А.Р. Лурия, Висконсинский тест сортировки карточек, словесно-цветовой тест Струпа, субтесты шкалы памяти Векслера, направленные на оценку состояния рабочей памяти, опросники совладающего поведения (тест Лазаруса) и агрессии (опросник Басса–Перри). По каждой из методик, исключая опросники, были выделены параметры, отражающие характер произвольной регуляции. Исходно предполагалось описывать больных 120 признаками, из которых было в конечном счете отобрано 19: 17 из П-шкалы и 2 из О-шкалы. П-признаки характеризуют группу инвалидности, результаты двигательных проб, счет, мышление, зрительный гнозис, память и агрессивность.

Был проведен поиск и выделение групп больных со сходными симптомокомплексами нарушений когнитивных и регуляторных функций.

При кластеризации объектов в базе данных 26×19 на первом уровне получено 10 кластеров. Для второго уровня особенно интересный результат получается для h-меры – 3 кластера со структурной матрицей:

	1	2	3
1	8		
2	4	11	
3	2		1
Σ	14	11	1

$\kappa \approx 0,65$

$1 = \{1, 8, 24, 9, 13, 15, 25, 7, 16, 17, 21, 23, 4, 5\};$

$2 = \{22,\ 3,\ 11,\ 18,\ 26,\ 6, 12, 19, 10, 14, 2\};$

$3 = \{20\}.$

На третьем уровне при h-мере получается структурная матрица для двух кластеров:

	1	2
1	18	
2	7	1
Σ	25	1

$\kappa \approx 0,46$

$1 = \{\text{все, кроме } 20\}$

$2 = \{20\}$

Хотя качество кластеризации очень невысоко, однако «маргинальное» поведение больного № 20 на обоих уровнях подчеркивается.

Для s-меры этот больной явно не выделяется.

Процесс кластеризации признаков быстро заканчивается уже на первом этапе, что описывается структурной матрицей

	1	2
1	12	
2		7
Σ	12	7

$\kappa = 1$

$1 = \{88, 32, 33, 35, 53, 55, 61, 72, 85, 89, 2, 3\};$

$2 = \{12, 1, 13, 14, 31, 68, 71\}.$

26×19 ↓ 26×26 ↓ 10×10 ↓ 3×3 ↓ 2×2 ↓ 1×1	база данных уровень – 0	$N_t / 26$	ℓ_t	$N_t / 26 + \ell_t$	$\max(N_t / 26, \ell_t)$	κ
	1	0,385	0,121	0,506	0,385	
	2	0,115	0,218	0,334	0,218	$\kappa_2 = 0,65$
	3	0,077	0,272	0,349	0,272	$\kappa_3 = 0,46$
	4			$t^* = 2$	$t^* = 2$	

Рис. 17. Система уровневых кластеров в задаче

Достаточно низкое качество кластеризации больных можно объяснить как малым числом выбранных признаков ($19 << 120$), скорее всего далеко не оптимальных, так и полиморфностью проявления рассеянного склероза. Задача сама по себе требует более глубокого изучения как объектов, так и признаков.

Таким образом, программа Clust показала свою эффективность в выявлении ярких единичных случаев. При кластеризации по параметрам когнитивных функций с использованием жесткой меры связей на 2-м уровне кластеризации было выделено 3 кластера. В первый вошли 14 больных, во второй — 11, а в третий кластер выделился один испытуемый. Это позволило уже с помощью идеографического подхода дать описание случая, выходящего за рамки выявленных закономерностей.

По всем параметрам больные второго кластера демонстрировали более низкие показатели регуляторных функций, чем больные первого. Далее проводился анализ различий групп испытуемых по параметрам совладающего поведения и агрессии.

Было показано предпочтение неконструктивных копинг-стратегий совладания с трудными жизненными ситуациями лицами с большей выраженностью регуляторных нарушений и в целом менее благоприятным течением заболевания. Трудности произвольной регуляции психической деятельности приводят к редкому использованию самоконтроля в стрессовой ситуации, а склонность к импульсивному реагированию – предпочтение конфронтационных способов совладания. По мере развития заболевания увеличивается частота использования стратегий избегающего типа. Кроме того, было показано, что скорость развития заболевания положительно связана с показателями враждебности – когнитивным компонентом агрессии в виде недоверия, подозрительности и обидчивости.

Выявленные особенности совладающего поведения у больных РС ставят проблему разработки методик коррекции произвольной регуляции психической деятельности, а также обучения адаптивным копинг-стратегиям (принятие ответственности, поиск социальной поддержки), что в свою очередь может привести к общей стабилизации состояния больного и повышению качества жизни.

Автор благодарит Е.В. Щербину за совместное исследование и любезно предоставленный материал.

2.10. Применение кластерного анализа для оценки роли МРТ-показателей при решении дифференциально-диагностических задач у больных с деменцией

Магнитно-резонансная томография (МРТ) все шире используется в диагностике заболеваний у пациентов с деменцией позднего возраста. Этот метод позволяет выявить потенциально излечимые случаи заболевания (например, объемные поражения), а также уточнить степень и локализацию атрофического процесса и выявить признаки цереброваскулярного заболевания.

Нами была поставлена задача проанализировать МРТ-картину пациентов, страдающих болезнью Альцгеймера, сосудистой деменцией, сочетанной сосудисто-атрофической деменцией, классифицировать изображения и провести сопоставление полученных данных с результатами клинических исследований.

Инструментом анализа был выбран метод кластерного анализа, позволяющий производить автоматизированную группировку наблюдений в однородные классы — кластеры. В данной работе задача кластерного анализа состояла в поиске сходных объектов.

Были изучены результаты МРТ головного мозга 166 пациентов [63, 72]. Изменения, выявленные у больных, включали расширение ликворных пространств, (желудочков, борозд) и очаговое поражение в вещества мозга. Для проведения анализа необходимо было формализовать данные МРТ. Для этого воспользовались реальными измерениями ширины желудочков, борозд больших полушарий и количеством выявленных очаговых изменений. Всего данные МРТ были описаны двадцатью параметрами, десять из которых описывали степень поражения вещества головного мозга в разных зонах и десять – степень расширения отделов мозга и борозд больших полушарий. Степень выраженности признака оценивалась в баллах от 0 до 3.

Каждый из 20 признаков относился к шкале порядка и характеризовался абсолютными частотами

$$(n_0,\ n_1,\ n_2,\ n_3), \quad \sum_{0}^{3} n = 166.$$

Так как степени интенсивности проявления признаков описываются кодами (0, 1, 2, 3), то исходная матрица «больной × признак» объема 166×20 может быть при кластеризации больных сжата до 166×4 при переходе к частотным описаниям признаков. Применение к ней алгоритма Clust1 дает 45 кластеров первого уровня, на втором – 9 (s-мера), на третьем – 3 (s-мера).

Кластеризация больных дает на первом уровне 45 кластеров. Их распределение по кластерам:

1 = {7}	16 = {4}	31 = {2}
2 = {3}	17 = {5}	32 = {2}
3 = {13}	18 = {4}	33 = {6}
4 = {4}	19 = {4}	34 = {5}
5 = {8}	20 = {3}	35 = {3}
6 = {4}	21 = {4}	36 = {2}
7 = {4}	22 = {3}	37 = {2}
8 = {4}	23 = {5}	38 = {2}
9 = {6}	24 = {4}	39 = {4}
10 = {2}	25 = {3}	40 = {3}
11 = {6}	26 = {3}	41 = {4}
12 = {5}	27 = {3}	42 = {1}
13 = {4}	28 = {4}	43 = {2}
14 = {2}	29 = {5}	44 = {1}
15 = {3}	30 = {7}	45 = {1}

На втором уровне при s-мере:

1 = {3, 1, 4, 7, 15, 17, 20, 23, 24, 30, 44};
2 = {10, 5, 8, 13, 14, 18, 19, 22, 32, 38};
3 = {6, 2, 12, 26, 28};
4 = {9, 11, 16, 21, 37};
5 = {25, 27, 29, 33, 34, 36, 40, 41};
6 = {31, 35};
7 = {42, 45};
8 = {39};
9 = {43}.

На третьем уровне при s-мере:
1 = {1, 2, 3, 4, 7};
2 = {5, 6, 8};
3 = {9}.
Соответствующие структурные матрицы имеют вид:

	1	2	3	4	5	6	7	8	9	
1	33	9	4	4	3					
2	1	21	1	2						
3	9	2	8	3	3					
4			6	12	3					
5				1	10			2		$\kappa \approx 0,51$
6	1	1			5	5				
7	1	2			3		2			
8	1							1		
9					4			1	2	
Σ	46	35	19	22	31	5	2	4	2	

Структурная матрица третьего уровня имеет вид:

	1	2	3	
1	119	4	0	
2	4	30	0	$\kappa \approx 0,86$
3	1	6	2	
Σ	124	40	2	

На четвертом уровне с s-мерой получается:

	1	2		
1	36	3	$\kappa \approx 0,9$	$1 = \{2, 3\}$;
2	6	121		$2 = \{1\}$.
Σ	42	124		

Система уровневых кластеров показана на рис. 18.

166×20 ↓ 166×4 ↓ 166×166 ↓ 45×45 ↓ 9×9 ↓ 3×3 ↓ 2×2 ↓ 1×1	база данных уровень	$N_t/166$	ℓ_t	$N_t/166+\ell_t$	$\max(N_t/166, \ell_t)$	κ
	0					
	1	0,271	0,155	0,426	0,271	
	2	0,054	0,260	0,314	0,260	$\kappa_2 = 0,51$
	3	0,018	0,300	0,318	0,300	$\kappa_3 = 0,86$
	4	0,012	0,430	0,442	0,430	$\kappa_4 = 0,90$
	5			t* = 2	t* = 2	

Рис. 18. Система уровневых кластеров в задаче

Описание признаков поражения мозга в баллах П-шкалы.

Редукция матрицы данных 166×20 до 4×20 путем перехода к частотному описанию признаков дает 20 частотных описаний:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	155	116	111	99	133	121	153	159	161	130	50	5	10	22	11	0	0	46	6	0
1	10	25	29	23	17	22	5	3	2	20	71	68	112	82	76	61	50	115	118	31
2	1	16	18	34	9	21	8	4	3	12	35	60	43	53	63	73	93	4	35	86
3	0	9	8	10	7	2	0	0	0	4	10	35	1	9	16	32	23	1	7	49

На первом уровне кластеризация признаков дает 6 кластеров со структурной матрицей:

	1	2	3	4	5	6
1	4					
2		3				
3		1	2			
4				2		
5					5	
6				1		2
Σ	4	4	2	3	5	2

$\kappa \approx 0,88$

На втором уровне при h-мере получаем 3 кластера:

	1	2	3
1	10		
2		5	
3		3	2
Σ	10	8	2

$\kappa \approx 0,76$

$1 = \{1, 2, 3\}$
$2 = \{4, 5\}$
$3 = \{6\}$

На втором уровне при s-мере получаются 3 кластера:

	1	2	3
1	10		
2		5	
3			5
Σ	10	5	5

$\kappa = 1$

$1 = \{1, 2, 3\}$;
$2 = \{4, 6\}$;
$3 = \{5\}$.

На третьем уровне получается:
1 = {2, 3};
2 = {1}.

С клинической точки зрения ценным получается результат кластеризации до третьего уровня иерархии. Из полученных кластеров первый кластер включал пациентов с различной степенью выраженности гидроцефалии, не имеющих очаговых изменений или с единичными очагами в веществе мозга больших полушариях. Преобладающее число пациентов не имело артериальной гипертензии, неврологических расстройств. Во втором кластере у всех пациентов было выявлено многоочаговое поражение, гидроцефалия различной степени выраженности. Бóльшая часть пациентов страдала гипертонической болезнью, сердечными заболеваниями. В третьем кластере оказалось два пациента с множественным поражением подкорковых структур.

Таким образом, применение кластерного анализа в настоящей работе позволило дифференцированно проанализировать значимость МРТ-характеристик для дифференциации пациентов пожилого возраста, страдающих деменцией.

В дальнейшем для уточнения нозологической значимости МРТ-показателей у больных с деменцией целесообразно ис-

пользовать данный вариант кластерного анализа с большим количеством параметров, описывающих МР-картину подкорковых изменений.

Автор благодарит О.В. Божко за совместную работу и любезно предоставленные данные.

2.11. Преимущества применения кластерного анализа при обработке данных нейрохимических исследований мозга

В биологической психиатрии нейрохимические исследования мозга, полученного после аутопсии, проводятся с целью обнаружения патологических изменений при психических заболеваниях. При этом сравниваются две (или несколько) группы – больные психическим(и) заболеванием(ями) и психически здоровые – и между этими группами ищутся достоверные различия по нейрохимическим признакам. Для такого рода задач характерно то, что объекты в группах немногочисленны, поэтому в качестве основного инструмента сравнения используется непараметрический анализ [83, 85]. Вторая особенность — большие индивидуальные различия нейрохимических признаков внутри групп. Они превышают различия между медианами по группам. Например, сильный разброс характерен для таких нейрохимических данных, как количество ферментов и других белков в экстрактах мозга [76, 84, 85].

Рассмотрим пример сравнения образцов мозжечка от двух групп — психически здоровых и больных шизофренией (по 22 объекта в каждой) — по нейрохимическим признакам (количеству ферментов глутаматдегидрогеназы (ГДГ), глутаминсинтетазы (ГС) и подобного ей белка (ГСПБ)). На диаграмме (см. рис. 19, цв. вклейка) различными символами (кружочками – ГДГ, ромбиками – ГС, треугольниками – ГСПБ) показано количество этих ферментов (в относительных единицах) для каждого объекта. Символы распределены в парные колонки: слева для каждой пары символов – больные шизофренией, справа – психически здоровые, а горизонтальные линии соответствуют медианам по группам. При сравнении «больных» со «здоровыми» с помощью непараметрического критерия U

Вилкоксона–Манна–Уитни в данном случае были получены достоверные различия групп [84]: по количеству ГДГ $p = 0{,}002$, по ГС $p = 0{,}0005$, по ГСПБ $p = 0{,}0001$. На диаграмме видно, как сильно «перекрываются» между собой две группы объектов в случае рассмотрения любого признака: в зону инверсий [86], или «перекрытия», попадает более половины объектов каждой группы. Такое же или еще большее «перекрытие» наблюдается, если сравнивать группы по другим нейрохимическим признакам — количеству ферментов и других белков [76, 84].

В результате дальнейшего применения непараметрических критериев (определение коэффициентов корреляции рангов Спирмена [83]) было обнаружено, что корреляции между уровнями отдельных ферментов, характерные для мозга психически здоровых, отличаются от связей в мозге больных шизофренией (причем подобный феномен наблюдался и при сравнении мозга пациентов с болезнью Альцгеймера с мозгом психически здоровых [85]). Эти наблюдения позволили выдвинуть гипотезу о системных нейрохимических изменениях в мозге при психических заболеваниях [84].

Исходя из этих предпосылок, кластерный анализ представлялся адекватным способом обработки информации, полученной в нейрохимических исследованиях. Этот подход (кластеризация) принципиально отличается от предыдущих способов, которыми группы **сравнивались** между собой. В случае кластер-анализа база данных (объекты×признаки) заранее не разделена на группы. Поскольку предполагалось, что кластерный анализ позволит выделить скопления «похожих» объектов, задача ставилась таким образом, чтобы выяснить, не разделятся ли объекты на «верхних» уровнях кластеризации на кластеры с преобладанием «больных» и – кластеры – с преобладанием «здоровых».

Так, кластерный анализ был применен к базе данных, куда были внесены 44 объекта (больные шизофренией и психически здоровые) с соответствующими нейрохимическими признаками (количеством ферментов – глутаминсинтетазы и подобного ей белка (ГС и ГСПБ), изоформ глутаматдегидрогеназы (ГДГ) и креатинфосфокиназы, а также количеством глиального фибриллярного кислого и основного миелиново-

го белков, измеренным в трех структурах мозга) и адресными признаками.

Параметрами описания 44 объектов являлись: 2 признака из шкалы наименований – пол, индекс здоровье/болезнь и 23 признака из шкалы отношений – состав белков в экстрактах мозга, полученного после аутопсии.

База данных задается матрицей порядка 44×25 (44 обследованных — 23 больных и 21 психически здоровый).

При кластеризации объектов Clust1 на первом уровне дает 14 кластеров со структурной матрицей

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	5													
2		2				1								
3			3											
4				6										
5					3									
6						5								
7							3							
8								2						
9									2					
10						1				4				
11											4			
12												1		
13													1	
14														1
Σ	5	2	3	6	3	7	3	2	2	4	4	1	1	1

$\kappa \approx 0,95$

На втором уровне с h-мерой получается 3 кластера:

1 = {1, 2, 3, 4, 5, 6, 7, 13};
2 = {11, 8, 9, 10, 12};
3 = {14}.

	1	2	3
1	23	1	0
2	7	12	0
3	0	0	1
Σ	30	13	1

$\kappa \approx 0,73$

С s-мерой получается:
1 = {1, 2, 3, 5, 7, 8, 13};
2 = {11, 4, 6, 9, 10, 12};
3 = {14}.

	1	2	3
1	17	5	0
2	2	19	0
3	0	0	1
Σ	19	24	1

$\kappa \approx 0,76$

На третьем уровне для обеих мер получается:
1 = {1, 2};
2 = {3}.

	1	2
1	43	0
2	0	1
Σ	43	1

$\kappa = 1$

Несколько иная картина возникает при кластеризации признаков.

На первом уровне получается 4 кластера со структурной матрицей диагонального вида:
1 = {4, 1, 5, 15, 18, 22, 25};
2 = {3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 19, 20, 21, 23, 24};
3 = {16};
4 = {2}.

	1	2	3	4	
1	7				
2		16			$\kappa = 1$
3			1		
4				1	
Σ	7	16	1	1	

Второй уровень для h-меры дает:
1 = {1, 2, 4};
2 = {3}.

	1	2	
1	23		$\kappa \approx 0,92$
2	1	1	
Σ	24	1	

Для s-меры получается:
1 = {2, 3, 4};
2 = {1}.

	1	2	
1	16		$\kappa \approx 0,84$
2	2	7	
Σ	18	7	

44×25	база данных					
↓						
44×44	уровень – 0	$N_t / 44$	ℓ_t	$N_t / 44 + \ell_t$	$\max(N_t / 44, \ell_t)$	κ
14×14	1	0,318	0,128	0,446	0,318	$\kappa_2 = 0,95$
3×3	2	0,068	0,207	0,275	0,207	$\kappa_3 = 0,73$
2×2	3	0,045	0,245	0,290	0,245	$\kappa_4 = 1$
1×1				$t^* = 2$	$t^* = 2$	

Рис. 20. Система уровневых кластеров в задаче

Таким образом, для признаков первый и второй уровни обеспечивают высокое соответствие оценок экспертов.

Интересная картина выявляется при анализе на втором уровне кластеризации объектов. При s-мере таблица сопряженности признаков «больной–норма» имеет вид

	K_1	K_2	
Б	3	19	22
Н	16	5	21
Σ	19	24	43

Для нее χ-квадрат с одной степенью свободы дает $\chi^2 \approx 17$; $\rho \approx 4 \cdot 10^{-5} << 0{,}05$.

Как видим, в результате анализа объекты разделились на два кластера с высокой степенью достоверности (средняя ошибка смешения < 20%, $p \approx 0{,}00004$), при этом один из этих кластеров составили в основном психически здоровые, а другой – больные шизофренией. Очень схематично получившееся разделение на кластеры можно представить так, как показано на рис. 19 (см. цв. вклейку) (под диаграммой): видно, что в «смешанный кластер» попали лишь трое «больных» и пять «здоровых» объектов, а один «маргинальный» объект выделился в отдельный кластер. Следует отметить интересную особенность Clust1 в данной задаче. Он подчеркивает «маргинальные» случаи – для обеих мер на втором уровне в третий кластер попадает больной с максимальным (20 ч) интервалом времени, прошедшим с момента смерти до замораживания образцов ткани для анализов.

Поскольку у представителей кластера «больных» в целом количество ферментов ГДГ, ГС и ГСПБ оказалось выше, чем у представителей кластера «здоровых», было сделано предположение о том, что рассмотренные группы (больные шизофренией и психически здоровые) объективно представляют различные метаболические типы: у больных тип метаболизма характеризуется повышением количества ключевых ферментов глутаматного обмена (ГДГ, ГС, ГСПБ) по сравнению с психически здоровыми.

Биохимическая часть исследования выполнена Г.Ш. Бурбаевой, И.С. Бокша и другими сотрудниками лаборатории ней-

рохимии, которым автор выражает благодарность за любезно предоставленный материал.

2.12. Кластеры объектов при исследовании смертности от алкогольных отравлений в Российской Федерации в 1991–1997 гг.

С помощью многолетних годичных данных по заболеваемости алкогольными психозами и смертности при отравлении алкоголем по всем областям России предполагалось «картировать» алкогольную ситуацию в стране.

Исходная посылка работы состояла в том, что больные алкогольными психозами (БАП) и умершие при отравлении алкоголем (УОА) рекрутируются из одной когорты людей: все БАП и почти все УОА исходно больны алкоголизмом. Из этого следует, что между двумя явлениями должна быть высокая корреляция при условии, что учет этих явлений поставлен удовлетворительно (собственно это и предстояло проверить на первом этапе работы). Гигантский рост смертности от отравления алкоголем (ОА) в РФ – с 1991 по 1994 г. на 238% — от 11,2 до 37,8 на 100 тыс. населения – настоятельно ставил задачу изучить распределение этого показателя по отдельным регионам. Для этого исследовался временной ряд показателя смертности в 1991–1997 гг., представленный как функция целочисленного аргумента:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7).$$

Полная матрица данных — $\|x_{ij}\|$, $1 \le i \le 79$, $1 \le j \le 7$, т.е. описание каждой области есть семимерный вектор с компонентами из шкалы отношений [65].

Сначала была использована некоторая аппроксимация, основанная на естественном соображении о поведении временного ряда смертности.

Простая форма динамики показателя смертности – все функции x одновершинные, с модой, приходящейся на 1994 г.; пример для Краснодарского края показан на рис. 21 а, б) – позволяет упростить описание этого показателя, введя меры скорости его возрастания и убывания как тангенсов углов наклона соответствующих прямых. Тем самым семимерный вектор

x превращается в пятимерный. Например, для Краснодарского края это вектор

$$y = (10, 3, 18, -2, 11).$$

Алгоритм Clust1 для матрицы 79×5 дал на первом уровне 16 кластеров, а на втором – 4 при кластеризации объектов. Состав этих кластеров с их выборочными средними (стандартная ошибка) приведен в табл. 1 [65].

Заметим, что если приближение y, использующее тангенсы углов подъема и спада в окрестности точки t = 1994, довольно точно передает поведение временного ряда смертности для «регулярных» случаев, то для «нерегулярных» оно грубо. Поэтому естественная картина системы кластеров всех областей РФ, полученная средствами Clust1, дается ниже.

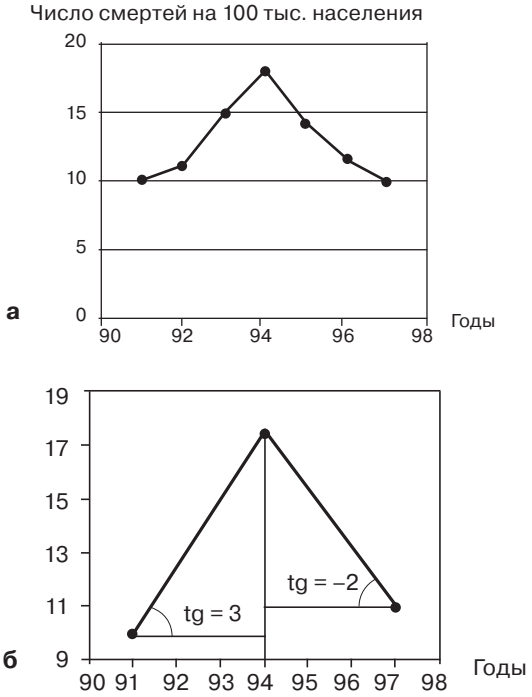

Рис. 21. Временной ряд за 1991–1997 гг. для Краснодарского края (а); Приближенный график временного ряда для Краснодарского края (б)

Таблица 1. Кластеры смертности от отравления алкоголем в регионах РФ за 1991–1997 гг.

	Кластер 1	**Кластер 2**	**Кластер 3**	**Кластер 4**
1	Москва	Вологодская обл.	Псковская обл.	Мурманская обл.
2	Респ. Мордовия	Смоленская обл.	Ивановская обл.	Респ. Карелия
3	Респ. Татарстан	Воронежская обл.	Владимирская обл.	Архангельская обл.
4	Курская обл.	Ульяновская обл.	Калужская обл.	Респ. Коми
5	Респ. Калмыкия	Волгоградская обл.	Нижегородская обл.	Санкт-Петербург
6	Ростовская обл.	Респ. Башкортостан	Респ. Чувашия	Калининградская обл.
7	Краснодарский край	Респ. Адыгея	Липецкая обл.	Ленинградская обл.
8	Ставропольский край	Респ. Саха	Белгородская обл.	Новгородская обл.
9	Респ. Ингушетия	Чукотский авт. округ	Тамбовская обл.	Костромская обл.
10	Респ. Кабардино-Балкария	Еврейская авт. обл.	Самарская обл.	Тверская обл.
11	Респ. Карачаево-Черкессия		Саратовская обл.	Ярославская обл.
12	Респ. Северная Осетия		Астраханская обл.	Московская обл.
13	Респ. Дагестан		Оренбургская обл.	Кировская обл.
14			Челябинская обл.	Пермская обл.
15			Курганская обл.	Респ. Удмуртская
16			Тюменская обл.	Свердловская обл.
17			Омская обл.	Брянская обл.
18			Томская обл.	Орловская обл.
19			Новосибирская обл.	Рязанская обл.
20			Красноярский край	Тульская обл.
21			Алтайский край	Пензенская обл.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4
22			Респ. Бурятия	Респ. Марий Эл
23			Магаданская обл.	Кемеровская обл.
24				Респ. Хакассия
25				Иркутская обл.
26				Респ. Алтай
27				Респ. Тыва
28				Читинская обл.
29				Амурская обл.
30				Хабаровский край
31				Камчатская обл.
32				Приморский край
33				Сахалинская обл.
1994	13	10	23	33
$X \pm s$	$8{,}4 \pm 1{,}4$	$24{,}8 \pm 1{,}1$	$34{,}7 \pm 1{,}3$	$65{,}0 \pm 2{,}0$

Таблица 2. Смертность от отравления алкоголем в субъектах РФ, 1991–1997 гг.

	n	Годы						
		1991	1992	1993	1994	1995	1996	1997
Респ. Карелия	1	11,6	45,0	65,0	80,8	79,4	52,9	44,7
Респ. Коми	2	11,5	37,8	76,2	86,5	73,9	55,2	36,3
Архангельская обл.	3	6,2	23,9	46,2	59,8	49,6	35,1	25,7
Вологодская обл.	4	10,1	16,3	24,6	11,3	18,1	20,8	10,2
Мурманская	5	10,6	23,1	45,3	48,2	35,2	21,1	11,1
Санкт-Петербург	6	11,6	28,0	49,1	46,3	28,3	21,5	15,9
Ленинградская обл.	7	16,1	35,8	75,7	87,2	55,4	34,1	26,0
Новгородская обл.	8	11,1	22,3	41,6	52,5	33,6	27,0	20,7

Продолжение табл. 2 на след. странице

Продолжение табл. 2

	n	Годы						
		1991	1992	1993	1994	1995	1996	1997
Псковская обл.	9	9,3	17,7	25,9	31,1	22,8	16,6	15,6
Брянская обл.	10	18,6	27,7	41,3	52,5	39,9	37,4	38,9
Владимирская обл.	11	14,0	15,5	29,3	37,7	29,2	21,8	18,8
Ивановская обл.	12	15,3	25,3	40,5	41,5	31,6	26,9	20,1
Калужская обл.	13	8,7	11,9	26,7	27,5	34,2	27,7	14,0
Костромская обл.	14	18,3	32,8	47,1	56,1	51,5	37,0	27,3
Москва	15	2,2	2,6	5,2	7,0	6,0	5,0	3,5
Московская обл.	16	14,8	22,9	45,1	63,2	53,7	39,6	27,6
Орловская обл.	17	31,6	35,7	50,6	44,3	38,1	36,8	37,5
Рязанская обл.	18	20,6	28,5	43,1	56,9	41,8	34,4	28,4
Смоленская обл.	19	8,4	11,7	16,5	20,6	17,1	10,5	9,3
Тверская обл.	20	16,3	30,4	60,5	75,4	49,8	38,9	31,8
Тульская обл.	21	27,0	34,7	56,1	68,3	52,3	43,5	37,4
Ярославская обл.	22	21,1	37,6	68,5	81,0	60,3	50,1	43,1
Респ. Марий Эл	23	31,8	39,2	51,4	72,2	48,5	31,2	29,2
Респ. Мордовия	24	2,7	3,9	8,2	12,8	11,8	9,7	10,2
Респ. Чувашия	25	21,6	23,2	40,2	46,7	36,2	23,4	16,3
Кировская обл.	26	18,2	33,2	55,5	75,9	63,8	54,3	40,1
Нижегородская обл.	27	17,2	24,1	41,6	47,3	40,1	34,2	2,8
Белгородская обл.	28	14,6	21,4	27,7	29,8	28,0	22,9	21,4
Воронежская обл.	29	9,3	9,4	15,6	25,2	18,6	16,5	15,4

	n	Годы						
		1991	1992	1993	1994	1995	1996	1997
Курская обл.	30	8,3	8,4	9,6	7,9	5,9	4,4	5,4
Липецкая обл.	31	13,5	20,8	25,1	33,3	32,5	28,6	18,9
Тамбовская обл.	32	11,1	19,2	23,4	33,3	31,0	20,7	18,2
Респ. Калмыкия	33	3,7	6,2	6,8	10,6	13,2	6,6	4,1
Респ. Татарстан	34	7,5	6,8	11,1	16,2	13,8	10,2	9,1
Астраханская обл.	35	9,6	9,7	25,8	31,6	24,5	18,6	14,8
Волгоградская обл.	36	12,6	10,9	18,4	24,0	15,1	15,7	14,7
Пензенская обл.	37	20,5	25,2	39,6	50,1	41,1	35,2	37,4
Самарская обл.	38	4,6	8,2	19,6	30,8	24,1	26,7	16,9
Саратовская обл.	39	10,5	14,0	25,8	34,3	29,6	25,9	26,8
Ульяновская обл.	40	8,6	12,6	19,6	24,9	15,2	12,1	8,5
Респ. Адыгея	41	0	16,0	19,6	23,8	19,1	16,9	10,9
Респ. Дагестан	42	0,8	0,5	0,9	1,2	0,6	1,3	0,7
Респ. Ингушетия	43	0	0,9	0	0	1,7	0,7	0
Респ. Кабардино-Балкария	44	1,9	1,7	4,7	6,5	3,7	4,6	4,0
Респ. Карачаево-Черкессия	45	0	4,4	4,1	6,4	3,4	5,3	3,4
Респ. Северная Осетия	46	2,0	1,8	4,5	2,9	2,7	3,0	2,9
Краснодарский край	47	10,1	10,3	15,3	18,2	15,0	11,8	10,7
Ставропольский край	48	3,7	4,5	7,8	9,2	9,4	5,2	5,2
Ростовская обл.	49	2,1	2,0	1,8	2,5	2,8	2,2	1,6

Продолжение табл. 2 на след. странице

Окончание табл. 2

	n	Годы						
		1991	1992	1993	1994	1995	1996	1997
Респ. Башкортостан	50	4,3	9,4	17,1	20,9	13,4	12,5	7,9
Респ. Удмуртия	51	12,6	28,1	49,6	52,2	35,3	29,7	19,1
Курганская обл.	52	14,7	19,6	34,8	30,7	26,9	21,9	18,0
Оренбургская обл.	53	6,0	11,1	20,5	26,1	18,6	26,8	23,2
Пермская обл.	54	18,6	23,5	49,3	87,3	62,3	50,6	39,3
Свердловская обл.	55	11,0	26,4	42,3	51,6	46,1	40,3	29,2
Челябинская обл.	56	6,8	11,2	21,2	27,6	23,4	20,2	15,0
Респ. Алтай	57	0	33,4	55,1	83,3	93,1	74,9	42,1
Алтайский край	58	18,11	25,1	39,0	46,5	43,6	30,0	22,6
Кемеровская обл.	59	17,7	37,2	61,5	67,9	56,9	43,1	31,3
Новосибирская обл.	60	5,0	9,2	21,0	28,4	15,6	19,0	16,5
Омская обл.	61	13,2	18,9	29,4	35,3	24,9	21,0	12,6
Томская обл.	62	8,4	16,9	40,8	37,0	34,7	25,3	20,2
Тюменская обл.	63	8,6	18,2	26,6	27,5	21,7	21,8	13,3
Респ. Бурятия	64	6,2	13,1	24,7	29,2	19,7	13,4	9,9
Респ. Тыва	65	8,4	29,4	39,2	56,3	50,5	46,8	58,3
Респ. Хакассия	66	0	14,9	29,6	52,0	26,5	22,4	25,4
Красноярский край	67	6,3	10,2	18,5	42,4	32,9	32,0	18,7
Иркутская обл.	68	18,8	29,1	47,4	59,3	40,8	20,6	15,1
Читинская обл.	69	0	15,7	31,3	57,8	38,6	34,2	29,5
Респ. Саха	70	7,3	12,3	14,1	19,7	16,6	14,0	10,2
Приморский край	71	5,9	14,5	58,7	23,9	20,2	16,4	12,9
Хабаровский край	72	5,8	9,9	65	15	8,9	9,6	7,7

	n	Годы						
		1991	1992	1993	1994	1995	1996	1997
Еврейская авт. обл.	73	0	11,3	31,6	6,1	0,9	3,4	4,8
Амурская обл.	74	12,6	38,0	57,5	62,4	44,5	43,9	37,2
Камчатская обл.	75	22,3	41,4	48,9	70,6	57,3	32,2	22,1
Магаданская обл.	76	6,0	13,6	11,0	26,6	29,0	19,6	13,9
Чукотский авт. округ	77	0	13,3	19,1	26,3	30,5	30,5	9,4
Сахалинская обл.	78	20,8	30,6	15,1	70,1	56,2	30,7	28,1
Калининградская обл.	79	24,2	33,9	64,8	71,8	49,7	52,3	49,0

Для точных данных x_i, $1 \le i \le 7$ (табл. 2), получается следующий результат.

На первом уровне получается 24 кластера: со структурной матрицей с высоким $\kappa \approx 0{,}97$. Структурные матрицы систем кластеров второго уровня также обладают высоким κ. Для третьего уровня качество заметно понижается лишь для h-меры ($\kappa \approx 0{,}52$), но остается высоким для s-меры.

Список 24 кластеров:

1 = {Москва, Кабардино-Балкария, Карачаево-Черкессия} – 3;
2 = {Северная Осетия, Ростовская} – 2;
3 = {Дагестан, Ингушетия} – 2;
4 = {Мордовия, Курская, Ставропольская, Калмыкия} – 4;
5 = {Вологодская, Смоленская, Татарстан, Ульяновская, Краснодарский, Башкортостан, Саха} – 7;
6 = {Псковская, Архангельская, Тюменская, Бурятия} – 4;
7 = {Рязанская, Брянская, Костромская, Орловская, Пензенская, Свердловская, Алтайский, Иркутская, Сахалинская} – 9;
8 = {Воронежская, Волгоградская, Адыгея} – 3;
9 = {Челябинская, Калужская, Самарская, Оренбургская, Новосибирская, Магаданская} – 6;
10 = {Белгородская, Курганская} – 2;
11 = {Липецкая, Тамбовская, Красноярский} – 3;
12 = {Владимирская, Саратовская, Омская} – 3;
13 = {Новгородская, Ивановская, Удмуртия, Томская} – 4;

14 = {Мурманская, Санкт-Петербург, Чувашия} – 3;
15 = {Архангельская, Московская, Тыва} – 3;
16 = {Амурская, Кемеровская, Ленинградская, Тверская, Тульская} – 5;
17 = {Марий Эл, Камчатская} – 2;
18 = {Нижегородская} – 1;
19 = {Калининградская, Кировская, Ярославская, Пермская, Алтай} – 5;
20 = {Хакасия, Читинская} – 2;
21 = {Чукотский авт. окр.} – 1;
22 = {Карелия, Коми} – 2;
23 = {Приморский край, Хабаровский край} – 2;
24 = {Еврейская авт. обл.} – 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	3																							
2		2																						
3			2																					
4				4																				
5					7																			
6						4																		
7							8																	
8								3																
9									6															
10										2														
11											3													
12												3												
13													4											
14														3										
15															3									
16																5								
17							1										2							
18																		1						
19																			4					
20																				2				
21																					1			
22																			1			2		
23																							2	
24																								1
Σ	3	2	2	4	7	4	9	3	6	2	3	3	4	3	3	5	2	1	5	2	1	2	2	1

$\kappa \approx 0{,}97$

Второй уровень, h-мера:
1 = {4, 1, 2, 3, 5, 24};
2 = {6, 8, 9, 21};
3 = {10, 11, 12};
4 = {7, 13, 14, 15, 18};
5 = {16, 17, 19, 22};
6 = {20};
7 = {23}.

	1	2	3	4	5	6	7	
1	14							
2	5	14						
3			8	5				
4				15	3			$\kappa \approx 0,86$
5					11			
6						2		
7							2	
Σ	19	14	8	20	14	2	2	

Второй уровень s-мера:
1 = {1, 2, 3, 4};
2 = {12, 10, 11, 20};
3 = {5, 8, 23};
4 = {6, 9, 21};
5 = {13, 7, 14, 18};
6 = {16, 15, 17};
7 = {19, 22};
8 = {24}.

	1	2	3	4	5	6	7	8	
1	11								
2		8							
3			7						
4		1	5	11					$\kappa \approx 0,90$
5		1			15	1			
6						9	1		
7							6		
8								1	
Σ	11	10	12	11	17	10	7	1	

Третий уровень, h-мера:
1 = {4, 2, 3, 5, 6};
2 = {7, 1}.

	1	2
1	39	
2	19	21
Σ	58	21

$\kappa \approx 0,52$

Третий уровень, s-мера:
1 = {1, 8};
2 = {2, 3, 4, 5};
3 = {7, 6}.

	1	2	3
1	12	6	
2		40	
3		4	17
Σ	12	50	17

$\kappa \approx 0,81$

Если приписать каждому кластеру уровня 1, (K_i, $1 \le i \le 24$), величину $R_i = \max\limits_{1 \le j \le 7} x_{ij}$ — риск среди образующих его кластеров уровня О (за время 1991–1997 гг.), то в терминах риска представляется удобным выразить смертность от ОА в РФ по группам кластеров любого уровня.

Так, для уровня 3 при h-мере суммарный риск для всех 79 регионов составляет $R = \sum_i R_i = 1143,5$.

Для s-меры 4-го уровня получается 17 наиболее «неблагополучных» регионов, упорядоченных по степени убывания риска:

1. Респ. Алтай – 93,1;
2. Пермская обл. – 87,3;
3. Ленинградская обл. – 87,2;
4. Респ. Коми – 86,5;
5. Ярославская обл. – 81,0;
6. Респ. Карелия – 80,0;
7. Кировская обл. – 75,9;
8. Тверская обл. – 75,4;

9. Респ. Марий Эл – 72,2;
10. Калининградская обл. – 71,8;
11. Камчатская обл. – 70,6;
12. Тульская обл. – 68,3;
13. Кемеровская обл. – 67,9;
14. Московская обл. – 63,2;
15. Амурская обл. – 62,4;
16. Респ. Тыва – 56,3;
17. Архангельская обл. – 51,8.

Для h-меры 3-го уровня получается 21 наиболее «благополучный» регион, упорядоченный по степени возрастания риска:

1. Респ. Дагестан – 1,3;
2. Респ. Ингушетия – 1,7;
3. Ростовская обл. – 2,8;
4. Респ. Сев. Осетия – 4,5;
5. Респ. Карачаево-Черкессия — 6,4;
6. Респ. Кабардино-Балкария – 6,5;
7. Москва – 7,0;
8. Ставропольский край – 9,4;
9. Курская обл. – 9,6;
10. Респ. Калмыкия – 12,2;
11. Респ. Мордовия – 12,8;
12. Респ. Татарстан – 16,2;
13. Краснодарский край – 18,2;
14. Респ. Саха – 19,7;
15. Смоленская обл. – 20,6;
16. Респ. Башкортостан – 20,9;
17. Вологодская обл. – 24,6;
18. Ульяновская обл. – 24,9;
19. Еврейская авт. обл. – 31,6;
20. Приморский край. – 58,7;
21. Хабаровский край – 65,0.

Обращают на себя внимание 2 кластера предыдущих уровней — № 19 и (20, 21), которые явно выделяются своими рисками из предшествующей группы рисков, не превосходящих $\approx$ 30. Это следствие перехода к третьему уровню со второго.

При этом соответствующие этому уровню кластеры 1 и 2 дают $R_1 = 1085{,}9 \approx 89\%$, $R_2 = 122{,}6 \approx 11\%$ оценки смертности от ОА по РФ.

Аналогичная картина уровня 4 при s-мере дает структурную матрицу для двух кластеров:

	1	2
1	55	0
2	7	17
Σ	62	17

$\kappa = 0{,}82$

$1 = \{1, 2\}$
$2 = \{3\}$

с той разницей, что во второй кластер собираются самые «тяжелые» регионы.

Результат впечатляющий – 17 субъектов РФ несут риск 481,9, что составляет 42% от всего риска, численно составляя лишь 22% от всех регионов.

Таким образом, Clust1 выделяет скопление данных, несущих некоторое экстремальное свойство, подобно тому, как в задачах психиатрии выделялись маргинальные элементы.

На первый взгляд (для случая h-меры) в кластере 1 собираются все «сильно пьющие» регионы (58 субъектов РФ), а в кластере 2 – «слабо пьющие» (21 субъект РФ). Однако картина потребления алкоголя в РФ не столь однозначна. Здесь могут сказываться как национальные особенности, так и степень доступности скорой медицинской помощи (примером может служить Москва). Более детальные характеристики потребления алкоголя в РФ можно найти в работе [51]. Если рассматривать проблему алкогольных психозов (АП) и ОА в целом, на первом этапе (здесь этот этап не рассмотрен) применение кластерного анализа обогатило ее решение за счет нетривиального подхода в виде кластеризации 693 значений АП и ОА и создания эталона для сравнения. Эталон составили 6 областей с высокой корреляцией АП и ОА и небольшим разбросом отношений АП и ОА. АП и ОА остальных областей были ранжированы по мере отличия от эталона с присвоением балла отличия, а мерой ошибок учета – разница баллов АП и баллов ОА. В результате мы пришли к выводу, что в случае ОА по сравнению с АП действует какой-то дополнительный фактор или факторы. Объяснение такого расхождения

уровня смертности и заболеваемости психозами мы связали с ошибками учета ОА. Эти ошибки тем больше, чем «дальше» отстоят показатели областей от эталонных областей, и таким образом были выявлены области, где учет ОА существенно занижен.

В развитие результатов этого исследования – на втором этапе, рассмотренном здесь, было показано, что в России показатели смертности при ОА занижены в 1,6 раза [51].

79×7 $\downarrow$	база данных					
79×79 $\downarrow$	уровень – 0	$N_t / 79$	ℓ_t	$N_t / 79 + \ell_t$	$\max(N_t / 79, \ell_t)$	κ
24×24 $\downarrow$	1	0,304	0,096	0,403	0,304	$\kappa_1 = 0,97$
7×7 $\downarrow$	2	0,089	0,146	0,235	0,146	$\kappa_2 = 0,86$
2×2 $\downarrow$	3	0,025	0,342	0,367	0,342	$\kappa_3 = 0,52$
1×1	4			$t^* = 2$	$t^* = 2$	

Рис. 22. Система уровневых кластеров в задаче

Глава 3

Действие алгоритма Clust в задаче оценки устойчивости развития стран

3.1. Системы кластеров в задаче оценки устойчивости развития стран Европы и СНГ

Анализ мирового опыта в области разработки индикаторов устойчивого развития экономических систем позволяет выделить два подхода к их построению. Первый подход заключается в построении системы индикаторов, каждый из которых отражает отдельные аспекты устойчивого развития. Другой подход – построение комплексного показателя, на основе которого можно судить о степени устойчивости социально-экономического развития. Наиболее распространенным показателем второго типа является «Human Development Index» — «индекс человеческого развития» (ИЧР) или чаще употребляемое в русскоязычной литературе название этого показателя — «индекс развития человеческого потенциала» (ИРЧП). Он был разработан в рамках Программы развития Организации Объединенных Наций (ПРООН) и успешно используется более чем в 170 странах мира с 1990 г. [87, 88, 89, 90].

ИРЧП является комплексным показателем, оценивающим уровень средних достижений страны по трем основным направлениям в области развития человека: *долголетие, знания, уровень жизни*.

Для выявления мировых тенденций в динамике показателей человеческого потенциала использовался анализ статистических данных социально-экономического развития России и более чем 170 стран мира (с населением более 1 млн человек), которые являются официальными результатами ежегодно публикуемых в рамках ПРООН «Докладов о человеческом развитии» («Human Development Report»).

Выявленные закономерности в изменении базовых и комплексного показателей человеческого потенциала свидетельствуют о сложных взаимосвязанных процессах, протекающих в социально-экономической сфере. Ряд исследователей полагают, что для анализа таких процессов можно использовать кластерный анализ [40, 91].

Итак, ИРЧП, по данным ООН, описывается как вектор (x_1, x_2, x_3), где x_1 – уровень образования, x_2 – уровень долголетия, x_3 — внутренний валовой продукт на душу населения.

Более точно это:

x_1 – долголетие на основе здорового образа жизни, определяемое уровнем ожидаемой продолжительности жизни при рождении;

x_2 – знания, измеряемые уровнем грамотности взрослого населения и совокупным валовым коэффициентом поступивших в начальные, средние и высшие учебные заведения;

x_3 – достойный уровень жизни, оцениваемый валовым внутренним продуктом на душу населения в соответствии с паритетом покупательной способности в долларах США.

Величины x_1, x_2, x_3, принимающие значения от 0 до 1, представляются ООН.

В [66] проведена кластеризация двух региональных блоков — Европа + СНГ (40 стран) и Африки (44 страны) в 1992 и 2002 гг. Соответственно базы данных описываются матрицами 40×3 и 44×3, а признаки принадлежат шкале отношений. Обе они были обработаны алгоритмом Clust1.

База данных по «Европа + СНГ» представлена в табл. 3.

Таблица 3. Индекс развития человеческого потенциала в 1992 и 2002 гг., Европа + СНГ

	Ио (x_1)92	**Ид (x_2)92**	**Иввп (x_3) 92**	**Ио(x_1) 02**	**Ид (x_2)02**	**Иввп (x_3)02**
Беларусь	0,90	0,75	0,95	0,95	0,75	0,67
Российская Федерация	0,89	0,71	0,95	0,95	0,69	0,74
Украина	0,87	0,74	0,92	0,94	0,74	0,65

Продолжение табл. 3 на след. странице

Продолжение табл. 3

	Ио (x_1)92	**Ид (x_2)92**	**Иввп (x_3) 92**	**Ио(x_1) 02**	**Ид (x_2)02**	**Иввп (x_3)02**
Казахстан	0,87	0,74	0,78	0,93	0,69	0,68
Туркменистан	0,91	0,67	0,62	0,93	0,70	0,63
Грузия	0,92	0,80	0,41	0,89	0,81	0,52
Азербайджан	0,87	0,76	0,46	0,88	0,78	0,58
Армения	0,92	0,79	0,43	0,9	0,79	0,57
Узбекистан	0,90	0,74	0,48	0,91	0,74	0,47
Кыргызстан	0,90	0,73	0,51	0,92	0,72	0,46
Респ. Молдова	0,89	0,71	0,67	0,87	0,73	0,45
Норвегия	0,95	0,87	0,98	0,99	0,90	0,93
Швеция	0,92	0,89	0,64	0,99	0,92	0,65
Нидерланды	0,95	0,87	0,98	0,99	0,89	0,94
Бельгия	0,94	0,86	0,78	0,99	0,90	0,71
Швейцария	0,91	0,88	0,98	0,95	0,90	0,93
Дания	0,94	0,84	0,95	0,98	0,86	0,82
Ирландия	0,94	0,84	0,98	0,96	0,86	0,94
Великобритания	0,92	0,85	0,96	0,99	0,88	0,87
Австрия	0,98	0,85	0,98	0,99	0,89	0,96
Франция	0,94	0,87	0,97	0,96	0,90	0,98
Германия	0,95	0,85	0,97	0,96	0,89	0,90
Испания	0,93	0,88	0,98	0,95	0,90	0,93
Италия	0,94	0,88	0,67	0,97	0,89	0,77
Португалия	0,88	0,83	0,95	0,93	0,85	0,75
Греция	0,83	0,88	0,64	0,97	0,89	0,70
Словения	0,88	0,80	0,98	0,95	0,85	0,95
Чешская респ.	0,89	0,77	0,98	0,96	0,84	0,99
Польша	0,89	0,77	0,88	0,92	0,81	0,78
Венгрия	0,91	0,73	0,96	0,96	0,78	0,87
Словакия	0,88	0,77	0,51	0,95	0,81	0,70
Эстония	0,90	0,74	0,95	0,91	0,78	0,81
Литва	0,89	0,76	0,97	0,98	0,79	0,87
Хорватия	0,88	0,78	0,98	0,96	0,82	0,93
Латвия	0,88	0,74	0,98	0,90	0,76	0,93
Болгария	0,89	0,77	0,62	0,95	0,77	0,77
БЮР + Македония	0,84	0,78	0,96	0,91	0,81	0,84
Румыния	0,83	0,75	0,99	0,87	0,76	0,95

	Ио (x_1)92	Ид (x_2)92	Иввп (x_3) 92	Ио(x_1) 02	Ид (x_2)02	Иввп (x_3)02
Албания	0,80	0,78	0,98	0,89	0,81	0,93
Финляндия	0,85	0,85	0,95	0,88	0,88	0,80

Примечания:
Ио – образование – x_1.
Ид – долголетие – x_2.
Иввп – доход – x_3.

На первом уровне блок «Европа + СНГ» дает в 1992 г. 13 кластеров со структурной матрицей диагонального вида (к = 1):

1 = {Норвегия, Нидерланды, Австрия, Франция, Германия}.
2 = {Беларусь, РФ, Венгрия, Эстония}.
3 = {Швейцария, Испания}.
4 = {Словения, Чехия, Литва, Хорватия, Латвия}.
5 = {Грузия, Армения}.
6 = {Дания, Ирландия, Соединенное Королевство}.
7 = {Азербайджан, Узбекистан, Кыргызстан, Словакия}.
8 = {Португалия, Финляндия}.
9 = {Швеция, Бельгия, Италия, Греция}.
10 = {Македония, Румыния, Албания}.
11 = {Украина, Казахстан, Польша}.
12 = {Туркменистан, Молдова}.
13 = {Болгария}.

Кластеры второго уровня с h-мерой близости:

	1	2	3	4	5
1	10	1			
2		16			
3			6		
4				3	
5					4
Σ	10	17	6	3	4

$\kappa \approx 0,97$

$1 = \{1, 3, 6\}$;
$2 = \{4, 2, 8, 10, 11\}$;
$3 = \{5, 7\}$;
$4 = \{12, 13\}$;
$5 = \{9\}$.

Кластеры третьего уровня с h-мерой:

	1	2	3
1	27		
2		9	
3			4
Σ	27	9	4

$\kappa = 1$

$1 = \{1, 2\};$
$2 = \{3, 4\};$
$3 = \{5\}.$

Кластеры четвертого уровня с h-мерой имеют структурную матрицу:

	1	2
1	25	0
2	11	4
Σ	36	4

$\kappa = 0,45$

$1 = \{1, 2\};$
$2 = \{3\}.$

Во второй кластер попадают 4 страны – Швеция, Бельгия, Италия, Греция, хотя и весьма «благополучные», но имеющие $x_3 \approx 0,7$ при высоких x_1, x_2.

Кластеры второго уровня с s-мерой близости:

	1	2	3	4
1	12	1		
2		14		
3			9	
4				4
Σ	12	15	9	4

$\kappa \approx 0,97$

$1 = \{6, 1, 3, 8\};$
$2 = \{4, 2, 10, 11\};$
$3 = \{7, 5, 12, 13\};$
$4 = \{9\}.$

Кластеры третьего уровня с s-мерой:

	1	2
1	27	
2		13
Σ	27	13

$\kappa = 1$

$1 = \{1, 2\};$
$2 = \{3, 4\}.$

Тот же блок «Европа + СНГ» дает в 2002 г. на первом уровне 12 кластеров со структурной матрицей:

	1	2	3	4	5	6	7	8	9	10	11	12	
1	4												
2		4											
3			4										
4			1	2									
5					3								
6						3							$\kappa \approx 0,97$
7							3						
8								4					
9									3				
10										3			
11											5		
12												1	
Σ	4	4	5	2	3	3	3	4	3	3	5	1	

Со следующими кластерами:

1 = {Швейцария, Франция, Германия, Испания};
2 = {Норвегия, Нидерланды, Соединенное Королевство, Австрия};
3 = {Ирландия, Словения, Чехия, Хорватия, Дания};
4 = {Венгрия, Литва};
5 = {Беларусь, Украина, Словения};
6 = {Грузия, Азербайджан, Армения};
7 = {Узбекистан, Кыргызстан, Молдова};
8 = {Швеция, Бельгия, Италия, Греция};
9 = {Латвия, Румыния, Албания};
10 = {РФ, Казахстан, Туркменистан};
11 = {Португалия, Польша, Эстония, Болгария, Македония};
12 = {Финляндия}.

40×3	база данных					
40×40	уровень – 0	$N_t / 40$	ℓ_t	$N_t / 40 + \ell_t$	$\max(N_t / 40, \ell_t)$	κ
13×13	1	0,325	0,085	0,410	0,325	$\kappa_1 = 1$
5×5	2	0,125	0,120	0,245	0,125	$\kappa_2 = 0,97$
3×3	3	0,075	0,175	0,250	0,175	$\kappa_3 = 1$
2×2	4	0,050	0,285	0,335	0,285	$\kappa_4 = 0,45$
1×1	5			$t^* = 2$	$t^* = 2$	

Рис. 23. Система уровневых кластеров в задаче (1992 г.)

40×3	база данных					
40×40	уровень – 0	$N_t / 40$	ℓ_t	$N_t / 40 + \ell_t$	$\max(N_t / 40, \ell_t)$	κ
12×12	1	0,300	0,080	0,380	0,300	$\kappa_1 = 0,97$
3×3	2	0,075	0,250	0,325	0,250	$\kappa_2 = 0,76$
2×2	3	0,050	0,250	0,300	0,250	$\kappa_3 = 0,95$
1×1				$t^* = 3$	$t^* = 2,3$	

Рис. 24. Система уровневых кластеров в задаче (2002 г.)

Кластеры второго уровня с h-мерой:

	1	2	3
1	16		
2	6	12	
3			6
Σ	22	12	6

$\kappa \approx 0,76$

$1 = \{3, 1, 2, 4, 8, 9\}$;
$2 = \{11, 5, 10, 12\}$;
$3 = \{6, 7\}$.

Кластеры третьего уровня с h-мерой:

	1	2
1	33	
2	1	6
Σ	34	6

$\kappa \approx 0,95$

$1 = \{1, 2\}$;
$2 = \{3\}$.

Важно отметить, как за 10 лет существенно изменилась картина кластеризации – в кластере 2 остались лишь аутсайдеры

СНГ: Грузия, Азербайджан, Армения, Узбекистан, Кыргызстан, Молдова.

Кластеры второго уровня с s-мерой:

	1	2	3	4
1	16			
2		10		
3			6	
4	2	1		5
Σ	18	11	6	5

$\kappa \approx 0,9$

$1 = \{3, 1, 2, 4, 9\};$
$2 = \{5, 10, 11\};$
$3 = \{6, 7\};$
$4 = \{8, 12\}.$

Кластеры третьего уровня с s-мерой:

	1	2
1	33	
2	1	6
Σ	34	6

$\kappa \approx 0,95$

$1 = \{4, 1, 2\};$
$2 = \{3\}.$

Результаты кластер-анализа для стран Европы и СНГ, включая Россию, приведены на рис. 25.

По составу стран поддаются интерпретации результаты третьего уровня кластеризации. Динамика кластеров стран Европы и СНГ с использованием жестких мер близости практически аналогична динамике с использованием мягких мер близости.

На 1992 г. первый кластер стран Европы и СНГ включает 27 стран как с высоким значением ИРЧП (0,8 и выше), так и со средним (0,7–0,79). Во второй кластер входят страны со средним значением ИРЧП (0,69–0,79), в 3-й кластер — страны с наиболее высоким значением ИРЧП (0,91–0,93).

В 2002 г. для стран Европы и СНГ в отдельный кластер выделяется часть второго кластера 1992 г., а все остальные страны объединяются в первый кластер. Состав первого кластера можно объяснить снижением уровня дохода на душу населения для стран 3-го кластера 1992 г. (Бельгия, Греция, Италия, Швеция) и ростом уровня образования в Болгарии, Словакии и Туркменистане. Во второй кластер входят страны СНГ, имеющие наименьшие значения ИРЧП.

Рис. 25. Состав кластеров для стран Европы, включая Россию и СНГ, на третьем уровне кластеризации по 1992 г. и 2002 г., в зависимости от номера кластера, для жестких мер близости (схема)

3.2. Системы кластеров в задаче оценки устойчивости развития африканских стран

База данных по блоку «Африка» приведена в табл. 4. На первом уровне блок «Африка» дает в 1992 г. 14 кластеров со структурной матрицей

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	4													
2		2												
3			3											
4				3										
5					2									
6						3								
7							5							
8								2	1					
9									4					
10										2				
11											3			
12												2		
13													5	
14														3
Σ	4	2	3	3	2	3	5	2	5	2	3	2	5	3

$\kappa \approx 0{,}98$

со следующими кластерами:

1 = {Руанда, Малави, Уганда, Гвинея-Бисау};
2 = {Чад, Ангола};
3 = {Судан, Кот д'Ивуар, Мавритания};
4 = {Конго, Камерун, Гана};
5 = {Буркина-Фасо, Нигер};
6 = {Сенегал, Бенин, Гамбия};
7 = {Ботсвана, Тунис, Алжир, Маврикий, Ливия};
8 = {Бурунди, Эритрея};
9 = {Эфиопия, Гвинея, Мали, Сьерра-Леоне, Мозамбик};
10 = {Заир, Танзания};

11 = {Того, Нигерия, ЦАР};
12 = {Намибия, Габон};
13 = {Зимбабве, Кения, Лесото, Мадагаскар, Замбия};
14 = {ЮА, Египет, Марокко}.

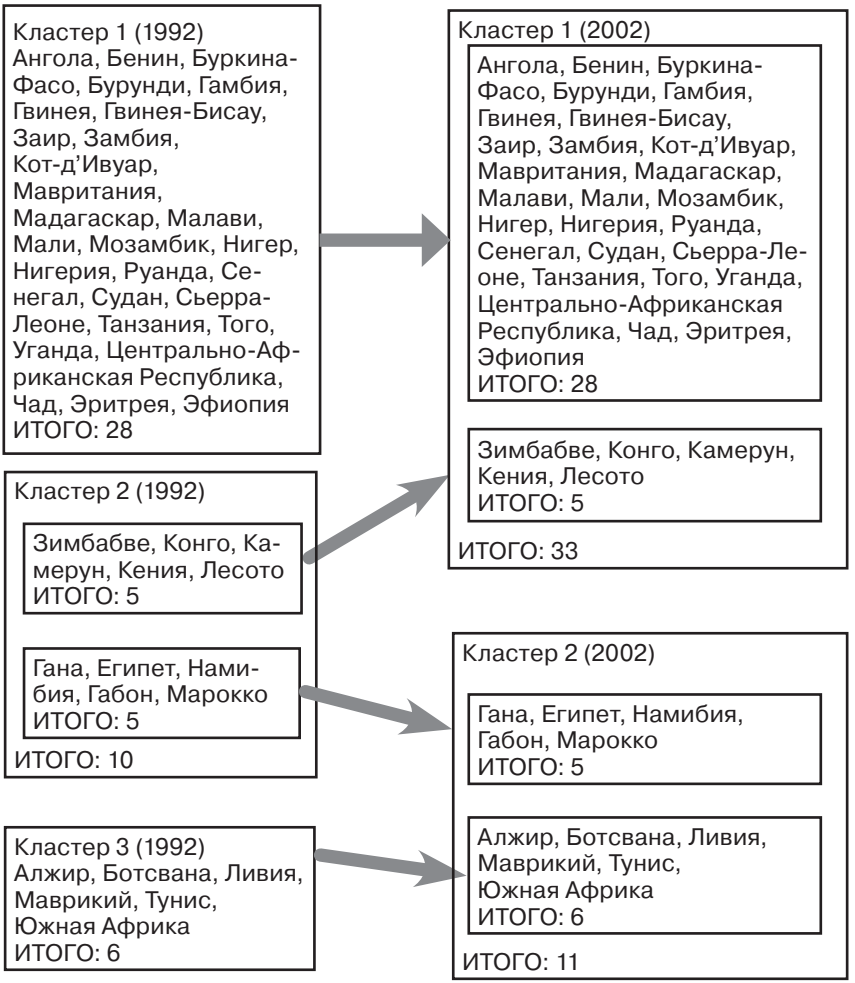

Рис. 26. Схема А. Состав кластеров для стран Африки на 3-м уровне кластеризации по 1992 г. и 2002 г. в зависимости от номера кластера для жестких мер близости

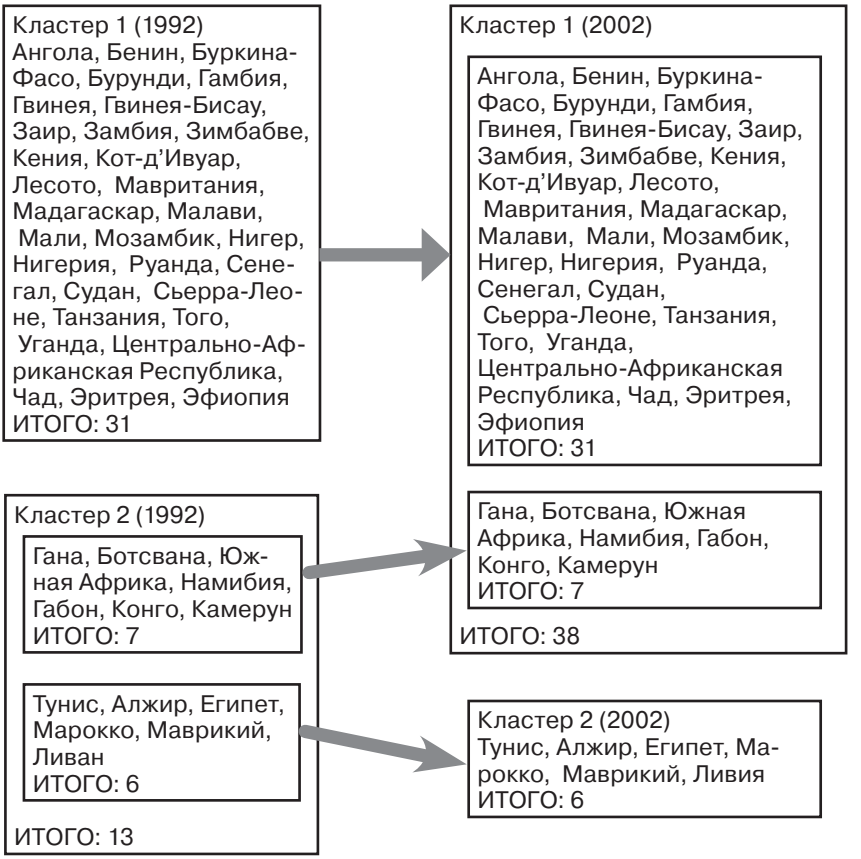

Рис. 26. Схема Б. Состав кластеров для стран Африки на 3-м уровне кластеризации по 1992 г. и 2002 г. в зависимости от номера кластера для мягких мер близости

Таблица 4. Индекс развития человеческого потенциала в 1992 и 2002 гг. (Африка)

	Ио (x_1) 92	Ид (x_2) 92	Иввп (x_3) 92	Ио (x_1) 02	Ид (x_2)02	Иввп (x_3) 02
Ботсвана	0,68	0,67	0,94	0,79	0,27	0,73
Тунис	0,63	0,71	0,94	0,73	0,79	0,70
Алжир	0,60	0,70	0,89	0,69	0,74	0,68

Продолжение табл. 4 на след. странице

Продолжение табл. 4

	Ио (x_1) 92	Ид (x_2) 92	Иввп (x_3) 92	Ио (x_1) 02	Ид (x_2)02	Иввп (x_3) 02
Южная Африка	0,79	0,63	0,69	0,83	0,40	0,77
Египет	0,55	0,64	0,64	0,63	0,73	0,61
Намибия	0,54	0,54	0,73	0,80	0,34	0,69
Габон	0,55	0,48	0,71	0,75	0,53	0,70
Марокко	0,41	0,64	0,61	0,50	0,72	0,61
Зимбабве	0,79	0,48	0,35	0,79	0,15	0,53
Конго	0,66	0,44	0,52	0,73	0,39	0,38
Камерун	0,56	0,52	0,43	0,64	0,36	0,50
Гана	0,55	0,52	0,38	0,64	0,55	0,51
Кения	0,69	0,51	0,24	0,73	0,34	0,39
Лесото	0,65	0,59	0,18	0,77	0,19	0,53
Мадагаскар	0,66	0,53	0,11	0,58	0,47	0,33
Замбия	0,67	0,4	0,21	0,68	0,13	0,36
Того	0,52	0,50	0,21	0,61	0,41	0,45
Нигерия	0,52	0,42	0,27	0,59	0,44	0,36
Заир	0,62	0,45	0,08	0,51	0,27	0,31
Судан	0,39	0,47	0,28	0,51	0,51	0,48
Кот-д'Ивуар	0,37	0,43	0,30	0,46	0,27	0,45
О.Р. Танзания	0,54	0,45	0,10	0,61	0,31	0,29
Центр.-Афр. Респ.	0,48	0,41	0,19	0,40	0,25	0,41
Мавритания	0,35	0,44	0,29	0,41	0,45	0,52
Сенегал	0,31	0,41	0,31	0,38	0,46	0,46
Бенин	0,33	0,38	0,29	0,42	0,43	0,40
Руанда	0,51	0,37	0,11	0,63	0,23	0,42
Малави	0,51	0,34	0,13	0,65	0,21	0,29
Уганда	0,51	0,33	0,14	0,69	0,34	0,44
Гамбия	0,35	0,33	0,22	0,41	0,48	0,47
Чад	0,39	0,38	0,12	0,41	0,33	0,39
Гвинея-Бисау	0,44	0,31	0,13	0,41	0,34	0,33

	Ио (x_1) 92	Ид (x_2) 92	Иввп (x_3) 92	Ио (x_1) 02	Ид (x_2)02	Иввп (x_3) 02
Ангола	0,39	0,36	0,12	0,38	0,25	0,51
Бурунди	0,32	0,42	0,12	0,43	0,26	0,31
Гвинея	0,29	0,33	0,09	0,39	0,40	0,51
Буркина-Фасо	0,18	0,37	0,13	0,24	0,35	0,40
Эфиопия	0,26	0,38	0,04	0,38	0,34	0,34
Мали	0,23	0,35	0,08	0,27	0,39	0,37
Сьерра-Леоне	0,28	0,23	0,15	0,41	0,16	0,28
Нигер	0,13	0,36	0,13	0,17	0,35	0,35
Мозамбик	0,33	0,36	0,05	0,43	0,22	0,39
Маврикий	0,74	0,75	0,97	0,80	0,78	0,78
Ливия	0,80	0,66	0,95	0,87	0,79	0,72
Эритрея	0,26	0,41	0,14	0,49	0,46	0,36

Примечания:
Ио – образование – x_1.
Ид – долголетие – x_2.
Иввп – доход – x_3.

Кластеры второго уровня с h-мерой:

	1	2	3	4
1	15	4		
2		13		
3		2	5	
4				5
Σ	15	19	5	5

$\kappa \approx 0,83$

$1 = \{3, 5, 6, 8, 9\}$;
$2 = \{11, 1, 2, 4, 10, 13\}$;
$3 = \{12, 14\}$;
$4 = \{7\}$.

Кластеры третьего уровня с h-мерой:

	1	2
1	33	
2	1	10
Σ	34	10

$\kappa \approx 0,95$

$1 = \{1, 2\}$;
$2 = \{3, 4\}$.

Кластеры второго уровня с s-мерой:

	1	2	3	4	5
1	10				
2	1	6			
3	1		10	1	
4				12	
5					2
Σ	13	6	10	13	2

$\kappa \approx 0,89$

$1 = \{2, 1, 8, 9\};$
$2 = \{3, 6\};$
$3 = \{10, 11, 13\};$
$4 = \{12, 4, 7, 14\}.$

Кластеры третьего уровня с s-мерой:

	1	2
1	31	1
2		12
Σ	31	13

$\kappa \approx 0,95$

$1 = \{1, 2, 3\};$
$2 = \{4\}.$

На первом уровне блок «Африка» дает в 2002 г. диагональную структурную матрицу 17×17:

(3, 2, 4, 3, 3, 2, 3, 3, 3, 2, 4, 2, 4, 2, 2, 1, 1) $\kappa = 1$

со следующими кластерами:

1 = {Чад, Гвинея-Бисау, Эфиопия};
2 = {Зимбабве, Лесото};
3 = {Судан, Сенегал, Бенин, Гамбия};
4 = {Кения, Конго, Уганда};
5 = {Мадагаскар, Нигерия, Эфиопия};
6 = {Мавритания, Гвинея};
7 = {Кот д'Ивуар, ЦАР, Мозамбик};
8 = {Заир, Бурунди, Сьерра-Леоне};
9 = {Буркина-Фасо, Мали, Нигер};
10 = {Тунис, Алжир};
11 = {Ботсвана, ЮА, Намибия, Габон};
12 = {Египет, Марокко};
13 = {Замбия, Танзания, Руанда, Малави};
14 = {Камерун, Того};
15 = {Маврикий, Ливия};
16 = {Гана};
17 = {Ангола}.

Кластеры второго уровня с h-мерой:

	1	2	3	4	5
1	11				
2	1	11			
3	1	3	5		
4				10	
5					2
Σ	13	14	5	10	2

$\kappa \approx 0,86$

$1 = \{3, 5, 6, 9, 16\};$
$2 = \{7, 1, 8, 13, 17\};$
$3 = \{14, 4\};$
$4 = \{10, 11, 12, 15\};$
$5 = \{2\}.$

Кластеры третьего уровня с h-мерой:

	1	2
1	31	–
2	1	10
Σ	34	10

$\kappa \approx 0,95$

$1 = \{2, 1, 3, 5\};$
$2 = \{4\}.$

44×3	база данных					
44×44	уровень – 0	$N_t / 44$	ℓ_t	$N_t / 44 + \ell_t$	$\max(N_t / 44, \ell_t)$	κ
12×12	1	0,318	0,119	0,437	0,318	$\kappa_1 = 0,98$
3×3	2	0,091	0,242	0,333	0,242	$\kappa_2 = 0,83$
2×2	3	0,045	0,412	0,466	0,421	$\kappa_3 = 0,95$
1×1	4			$t^* = 2$	$t^* = 2$	

Рис. 27. Система уровневых кластеров в задаче (2002 г.)

$44\underset{\downarrow}{\times}3$	база данных					
$44\underset{\downarrow}{\times}44$	уровень – 0	$N_t / 44$	ℓ_t	$N_t / 44 + \ell_t$	$\max(N_t / 44, \ell_t)$	κ
$14\underset{\downarrow}{\times}14$	1	0,386	0,101	0,487	0,386	$\kappa_1 = 1$
$4\underset{\downarrow}{\times}4$	2	0,114	0,238	0,352	0,238	$\kappa_2 = 0,86$
$2\underset{\downarrow}{\times}2$	3	0,045	0,481	0,526	0,481	$\kappa_3 = 0,95$
1×1	4			$t^* = 2$	$t^* = 2$	

Рис. 28. Система уровневых кластеров в задаче (1992 г.)

Кластеры второго уровня с s-мерой:

	1	2	3	4	5
1	9	2			
2		11			
3			10		
4				6	
5					6
Σ	9	13	10	6	6

$\kappa \approx 0,94$

$1 = \{3, 5, 6\}$;
$2 = \{1, 7, 8, 9, 17\}$;
$3 = \{14, 4, 13, 16\}$;
$4 = \{10, 12, 15\}$;
$5 = \{2, 11\}$.

Кластеры третьего уровня с s-мерой:

	1	2
1	36	
2	2	6
Σ	38	6

$\kappa \approx 0,91$

$1 = \{3, 1, 2, 5\}$;
$2 = \{4\}$.

Таким образом, на 1992 г. первый кластер стран Африки включает страны с самым низким значением ИРЧП (рис. 26, схемы А, Б). Во второй кластер для жестких (h) мер близости (рис. 26, схема А) входят страны со средним значением ИРЧП (0,49–0,65), для мягких (s) мер близости (рис. 26, схема Б) – со средним и высоким значением ИРЧП. В третий кластер для жестких мер близости (рис. 26,

схема А) входят страны со средним и высоким значением ИРЧП (0,59–0,79). В 2002 г. во второй кластер как для жестких, так и для мягких мер близости входят страны с наибольшими значениями ИРЧП.

В сравнении кластерной картины 1992 и 2002 гг. при h-мере развитые 10 стран (второй кластер уровня 3), а именно Ботсвана, Тунис, Алжир, ЮА, Намибия, Габон, Египет, Марокко, Маврикий, Ливия, сохраняют свое лидирующее место. Это объясняется низкими значениями x_3 у всех остальных 34 стран, поэтому некоторое его снижение у лидеров в 2002 г. не разрушает сложившийся в 1992 г. кластер.

3.3. Системы кластеров для региональных блоков с одномерным индексом человеческого потенциала

Кроме векторного описания $x = (x_1, x_2, x_3)$ рассмотрено и более простое – одномерный индекс человеческого потенциала, как их среднее

$$U = 1/3\ (x_1+x_2+x_3).$$

Для него в региональных блоках получаются (на втором уровне) следующие результаты:

1) Европа + СНГ, 1992 г.:

K_1 = {Венгрия, Словакия, Литва, Болгария, Беларусь, Греция, Чехия, Польша, Эстония, Украина, РФ} – 11;

K_2 = {Туркменистан, Армения, Кыргызстан, Албания, Казахстан, Финляндия, Грузия, Узбекистан, Азербайджан, Румыния, Молдова, Хорватия, Латвия, Македония} – 14;

K_3 = {Норвегия, Швеция, Нидерланды, Бельгия, Швейцария, Австрия, Франция, Германия, Италия, Дания, Ирландия, Соединенное Королевство, Испания} – 13;

K_4 = {Португалия, Словения}.

2) Европа + СНГ, 2002 г.:

K_1 = {РФ, Болгария, Финляндия, Беларусь, Македония, Украина, Казахстан, Румыния, Албания, Венгрия, Польша, Словакия, Литва, Эстония, Хорватия, Латвия} – 16;

K_2 = {Дания, Франция, Германия, Испания, Норвегия, Швеция, Нидерланды, Бельгия, Швейцария, Ирландия,

Соединенное Королевство, Австрия, Италия, Португалия, Греция, Словения, Чехия} – 17;

K_3 = {Узбекистан, Кыргызстан, Туркменистан, Грузия, Азербайджан, Армения, Молдова} – 7.

В обоих региональных блоках видно, даже вопреки переходу к одномерной картине, тенденцию к скоплению развитых стран в отдельные кластеры и «цепочечное» отставание остальных. Следует напомнить, что более ярко этот эффект виден при векторном представлении стран – (x_1, x_2, x_3).

3) Африка, 1992 г.:

K_1 = {Бенин, Сенегал, Руанда, Малави, Уганда} – 5;

K_2 = {Гвинея, Мозамбик, Буркина-Фасо, Эфиопия, Сьерра-Леоне, Мали, Нигер, Гвинея-Бисау, Гамбия, Чад, Ангола, Бурунди, Эритрея} – 13;

K_3 = {Танзания, Кот д'Ивуар, ЦАР, Мавритания, Заир, Судан} – 6;

K_4 = {Того, Нигерия, Мадагаскар, Замбия} – 4;

K_5 = {Ботсвана, Тунис, Маврикий, Ливия, Алжир, ЮА} – 6;

K_6 = {Марокко, Габон, Зимбабве, Конго, Египет, Намибия} – 6;

K_7 = {Кения, Камерун, Гана, Лесото} – 4.

4) Африка, 2002 г.:

K_1 = {Зимбабве, Конго, Кения, Лесото, Уганда, Камерун, Того, Судан, Мадагаскар, Нигерия, Мавритания} – 11;

K_2 = {Замбия, Кот д'Ивуар, Танзания, Малави, ЦАР, Заир, Эфиопия, Сенегал, Гамбия, Эритрея, Руанда, Бенин, Гвинея, Чад, Ангола} – 15;

K_3 = {Буркина-Фасо, Нигер, Гвинея-Бисау, Бурунди, Мали, Мозамбик, Сьерра-Леоне} – 7;

K_4 = {Египет, ЮА, Габон, Тунис, Алжир, Ливия, Мавритания} – 7;

K_5 = {Ботсвана, Намибия, Марокко, Гана} – 4.

Используя отношение квазипорядка, можно представить результаты таких одномерных кластеризаций в виде последовательностей кластеров:

1) для Европы + СНГ в 2002 г.:

$$K_3 \precsim K_1 \precsim K_2;$$

2) для Африки в 2002 г.:

$$K_3 \lesssim K_1 \lesssim K_2 \lesssim K_5 \lesssim K_4.$$

Уже на втором уровне видно, как страны начинают явно разделяться по степени развитости, понимаемой даже в общепринятом смысле. Тем более это заметно на третьем уровне.

Некоторые подробности влияния каждого компонента векторного представления индекса человеческого потенциала указаны в [66].

Глава 4

Закон Ципфа для укрупнения кластеров при росте уровня кластеризации

В публикации [28] достаточная структурированность теории, использующей кластер-анализ, иллюстрируется законом Ципфа – гиперболическим убыванием частот алгоритмов в зависимости от используемых параметров описания кластеризуемых объектов.

Сходная картина имеет место и для Clust1, с той разницей, что здесь речь идет о процессе укрупнения кластеров при росте уровня кластеризации. В табл. 5 сведено 15 задач, где с точностью до первого десятичного знака представлены величины

$$g(t) = \ln \frac{N_t}{N_{t+1}}, \quad N_t - \text{число кластеров уровня } t,\ t = 0, 1, 2, 3.$$

На рис. 29 показана линия линейной регрессии g(t) на t для 15 задач. Справа от точек вида (t, g(t)) представлены их абсолютные частоты. Уравнение линейной регрессии с точностью до второго знака есть

$$g(t) \cong 1{,}67 - 0{,}27 \cdot t,$$

уровень значимости $P \approx 0{,}00004$.

Хорошая логарифмически-линейная аппроксимация говорит о законе Ципфа.

Таким образом, поведение чисел кластеров в зависимости от уровня кластеризации соответствует закону Ципфа с высокой значимостью.

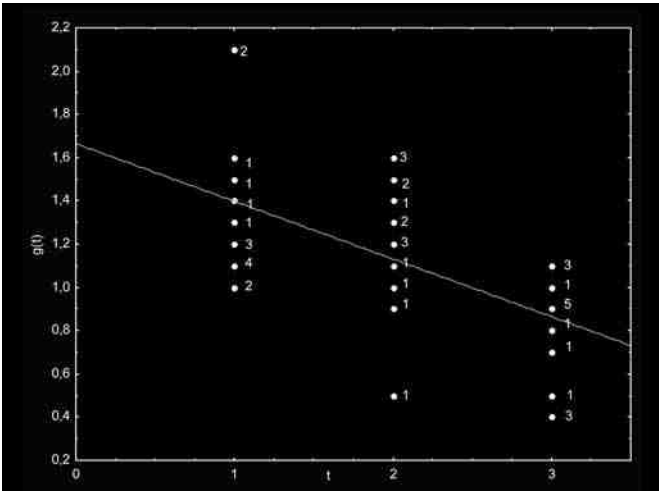

Рис. 29. Прямая линейной регрессии g(t) на t для 15 задач

Таблица 5. Число кластеров в зависимости от уровня t и функции g(t)

Гл., № задачи	**Число кластеров уровня t**				**Значения g(t)**		
	N_0	N_1	N_2	N_3	1	2	3
Гл. 2, 1	140	19	12	5	2,1	0,5	0,9
Гл. 2, 1	140	46	16	5	1,1	1,1	1,0
Гл. 2, 2	83	19	6	2	1,5	1,2	1,1
Гл. 2, 2	83	21	9	3	1,4	0,9	1,1
Гл. 2, 5	75	23	5	2	1,2	1,5	0,9
Гл. 2, 6	50	19	4	2	1,0	1,6	0,9
Гл. 2, 6	60	18	5	2	1,2	1,3	0,9
Гл. 2, 7	85	10	5	2	2,1	1,6	0,9
Гл. 2, 9	26	10	3	2	1,0	1,2	0,4
Гл. 2, 10	166	45	9	3	1,3	1,6	1,1
Гл. 2, 11	44	14	3	2	1,1	1,5	0,4
Гл. 2, 12	79	16	4	2	1,6	1,4	0,4
Гл. 3, 1	40	13	5	3	1,1	1,0	0,5
Гл. 3, 2	44	14	4	2	1,1	1,3	0,7

Глава 5

Описание программы Clust

С.Н. Мясоедов

В качестве бесплатного приложения к монографии предоставляется программа Clust. Программа написана С.Н. Мясоедовым под руководством С.А. Судакова в НЦПЗ РАМН. Авторские права на программу принадлежат НЦПЗ РАМН. Программа написана на языке фортран-95 и при помощи компилятора gfortran. Кодировка, в которой работает программа, – cp1251. В случае применения кодировки KOI-8r или cp866 результаты работы программы будут содержать символы в разных кодировках, что затруднит прочтение результата. При кодировке исходных данных в Unicode программа, скорее всего, не сможет прочитать исходный файл данных. Программа тестировалась в операционной системе Windows XP. В случае возникновения сбоев в работе программы автор просит присылать файл исходных данных, параметры запуска программы и сообщение об ошибке на адрес: smyasoe@mail.ru.

Программа Clust предназначена для задач кластер-анализа — автоматической группировки данных, представленных в виде прямоугольных таблиц – матриц вида «объект×признак», не содержащих пропусков элементов. В каждом столбце таблицы расположены данные по какому-либо одному признаку. В каждой строке таблицы расположены данные по какому-либо одному объекту. Разделителями данных (элементами, показывающими компьютеру, где кончается одно значение и начинается следующее), являются пробелы.

Максимальное число признаков – 9999. Максимальное число объектов – 32565. Память компьютера в 256 Мб позволяет проводить кластеризацию с ~ 1000 объектами или признаками.

Программа позволяет работать с данными сильной, порядковой и номинальной шкал измерения признаков одновременно. Номинальная шкала измерения содержит только качественную информацию о признаке, например кличка собаки или другого животного. Программа позволяет использовать любые символы в любом количестве в значениях номинальной шкалы. Порядковая шкала, наряду с качественной информацией, несет информацию о приоритете одних значений перед другими, например место, которое занимает спортсмен на соревновании. Данные порядковой шкалы, воспринимаемые программой, могут лежать в диапазоне от 0 до 999. Сильная шкала практически не несет никакой качественной информации, зато точно описывает количественную сторону информации, например, вес – можно представить в тоннах, а можно и в килограммах, и в граммах, и в миллиграммах. Данные сильной шкалы могут лежать в диапазоне $1 \cdot 10^{-34} - 1 \cdot 10^{34}$, максимум 4 значащие цифры для отрицательных чисел с отрицательным порядком, 6 значащих цифр для положительных чисел с положительным порядком. В таблице исходных данных порядок задается английской E.

Программа позволяет делать нормировку значений как по каждому столбцу в отдельности, так и по группам столбцов вместе. Нормировка является обязательной при работе программы и не позволяет отдельным признакам преобладать над другими при расчете. Нормировка по группам столбцов предпочтительна в случае сильной однородности признаков в этих столбцах, например один и тот же параметр, снятый в разное время.

При кластеризации объектов и признаков используются разные меры близости (расстояние между любыми двумя объектами или признаками), причем мера близости для кластеризации объектов более простая и дает более качественные результаты. Мера близости по объектам использует модуль разности для сильной шкалы (метод «City block» в программе Statistica). Мера близости по признакам использует парные отношения для шкал наименований и порядка и единица минус модуль коэффициент корреляции для сильной шкалы. Мера близости изменяется от 0 (расстояние равно 0) до 1 (расстоя-

ние равно ∞). Меры близости используются алгоритмом кластеризации.

В программе реализовано три алгоритма кластеризации.

Алгоритм-1 (ближайшая пара) находит группы кластеров вне зависимости от «рыхлости» кластера, т.е. размеры получаемых кластеров варьируют в очень больших пределах. На этапе «исходной кластеризации» первый кластер наименее «рыхлый», последний – наиболее. Этот алгоритм также позволяет группировать уже полученные кластеры (кластеризация n-го уровня). При этом используются «жесткая» мера близости («Single linkage» в программе Statistica) и «мягкая» мера близости («Unweighted pair-group average» в программе «Statistica»). Деление кластеров по уровням кластеризации производится исходя из структуры исходных данных.

Алгоритм-2 (максимальный каркас) находит заданное количество наиболее удаленных друг от друга групп.

Алгоритм-3 (оптимальная область) находит кластеры максимального объема с заданным радиусом мер близости (алгоритмы 1, 2, 3 основаны на аппарате, изложенном в работе [40]).

Во всех алгоритмах за этапом исходной кластеризации следует этап «центрирования», т.е. нахождения «центров» или, другими словами, «центроидов» кластеров и объединения точек вокруг этих элементов. В выходном файле центр кластера выводится первым по порядку в каждом кластере.

Программа также находит центры Чебышева на первом и последнем уровне кластеризации, а также меры разброса, если включен вывод значений. Меры разброса изменяются от 0 до 1.

При кластеризации объектов возможен вывод признаков, не участвующих в кластеризации. При кластеризации признаков вывод объектов, не участвующих в кластеризации, невозможен.

Для ввода набора параметров кластеризации необходимо заново запускать программу и вводить все параметры. Результаты (матрица расстояний или данные по составу кластеров) выдаются в текстовый файл с разделителями – пробелами.

Программа запускается из командной строки DOS путем перехода в рабочий каталог программы, набора последователь-

ности «Clust» и нажатия клавиши «Enter». Для упрощения работы с программой можно поместить файл исходных данных в рабочий каталог программы, и результирующие файлы также помещать в этот каталог.

Исходные данные вводятся в программу из файла (таблица исходных данных), а также в интерактивном режиме (параметры кластеризации).

Файл исходных данных в столбцах содержит признаки кластеризации, а в строках – объекты. Каждая строка, за исключением первой, содержит название объекта, состоящее из любых символов, кроме пробела и символа возврата каретки, и данные. Название объекта отделяется от данных, так же как и данные отделяются друг от друга, любым количеством пробелов. Названия объектов не должны повторяться. Тип данных определяется после запуска программы, единственным ограничением, в этом смысле накладываемым на файл, является однотипность данных в одном столбце. Данные шкалы наименований могут содержать любые символы, с теми же ограничениями, что и названия объектов, данные шкалы порядка могут быть в диапазоне от 0 до 999, данные сильной шкалы ~ от 10^{-34} до 10^{34} (п. 1.1), положительные и отрицательные, максимум 3 знака после запятой. Первая строка файла содержит названия признаков, причем на названия признаков накладываются такие же ограничения, как и на названия объектов. Первый столбец (названия объектов) не озаглавливается. Формат файла данных – любой текстовый. Не принимаются документы Word, таблицы Excel и т.п. На рис. 30 приведен правильный входной файл. Файл не должен также содержать пустых строк.

	1	2_3	в	г
а	бб	0	1.555E 30	0
бб	аа	11	–1.444E 30	1
ввввв	аа	12	1E –30	2
абв	г_г_г	0	–1E –30	2
г_д	аа	0	–100	2

Рис. 30. Образец входного файла

При кластеризации объектов таблица должна содержать минимум три объекта, при кластеризации признаков – минимум три признака и два объекта.

Программа выдает 4 типа выходных файлов:

— файл с поуровневым составом кластеров;

— файл с промежуточными результатами алгоритма-3;

— файл матрицы расстояний;

— файл с данными для последующего анализа.

Программа не позволяет загружать готовую исходную матрицу расстояний.

Пусть программа «Clust» помещена в рабочую папку «C:\Cl». Тогда последовательность команд для запуска программы приведена на рис. 31.

Пуск>Программы>Стандартные>Командная строка
В открывшемся окне командной строки набираем:
C:\Documents and Settings\имя пользователя>cd ..
C:\Documents and Settings>cd ..
C:\>cd cl
C:\Cl>clust

Рис. 31. Последовательность команд для запуска программы

Получаем приглашение программы:

Enter the name of input file:

1. Интерфейс интерактивного ввода параметров кластеризации.

В данном разделе вопросы, задаваемые программой, указываются в том же порядке, в каком задаются пользователю программой.

1.1. Ввод имени входного файла.

Первым вопросом, задаваемым программой пользователю, является

«Enter the name of input file:»

Допустим, файл из рис. 30 назван «test.txt» и помещен в ту же папку, что и программа Clust. Тогда в ответ на приглашение програмы ввести имя входного файла надо набрать имя этого файла и нажать **«Enter»**, например:

Enter the name of input file:test.txt+enter

Если файл находится не в папке запуска программы Clust, необходимо указать относительный или абсолютный путь к файлу, например файл находится в папке «C:\data».

Абсолютный путь:

Enter the name of input file:c:\data\test.txt+enter

Путь относительно папки «c:\cl»:

Enter the name of input file:.\..\data\test.txt+enter

В случае удачного ввода получаем примерно такие сообщения:

Scanning header of coordinates
4coordinates
Scanning 1 vector
4coordinates

Если количество признаков в первой строке равно количеству значений во второй (не считая названия объекта), то первый этап прошел нормально и можно переходить к п. 1.2.

В случае отсутствия файла данных по указанному пути выдается сообщение

No file found

1.2. Задание типа и количества областей нормировки данных.

Нормировка является обязательным действием при выполнении программы Clust, поэтому пользователь должен указывать программе, какие признаки и как нормировать. Наиболее часто употребляемой нормировкой является нормировка по столбцам, при этом все нормируемые признаки считаются неоднородными. Для улучшения работы алгоритма иногда применяется т.н. «общая» нормировка. Она пригодна в случае, когда одно и то же значение разных признаков имеет один и тот же смысл, например наименование продуктов, выращиваемых на поле, и наименование продуктов, перемещенных из поля в хранилище. Но если, например, один признак задает продукты растительного происхождения, используемые для приготовления пищи в ресторане, а другой задает продукты растительного происхождения, изображенные художником на натюрморте, то следует использовать нормировку по столб-

цам, так как вероятнее всего в первом случае нас интересует химический состав пищевых продуктов, а во втором – их внешний вид.

Программа Clust на этом этапе задает три вопроса о необходимости общей нормировки, ответить на которые можно положительно («**y**» и «**enter**») или отрицательно («**n**» и «**enter**»). В случае положительного ответа количество областей общей нормировки вводится как порядковое число + «**enter**»:

Common normalization for name scale? (y/n):y

Enter the number of scopes for common normalization of name scale:1

Common normalization for order scale? (y/n):y

Enter the number of scopes for common normalization of order scale:1

Common normalization for strong scale? (y/n):y

Enter the number of scopes for common normalization of strong scale:1

«**Name scale**» обозначает шкалу наименований, «**order scale**» – шкалу порядка, «**strong scale**» – сильную шкалу.

Напомним, что при общей нормировке одинаковое значение трактуется одинаково во всех признаках области общей нормировки, тогда как при нормировке по столбцам одинаковое значение трактуется по-разному в разных признаках. Из этого следует, что общую нормировку следует выбирать, когда точно известно, что признаки совершенно одинаковы, например цвет эмали ПФ-115 двух разных производителей. При различных признаках, например, цвет эмали и цветовое обозначение террористической опасности в США, следует выбирать нормировку по столбцам.

1.3. Задание используемых для расчета данных, их типа и нормировки.

В файле исходных данных не указано, к какой шкале принадлежат признаки. Для задания соответствия шкал и признаков исходного файла используются вопросы программы, описываемые в данном пункте. Также следует отметить, что программа производит расчет только по признакам, указанным пользователем в ответах на эти вопросы программы.

В исходной таблице (файле) признаки имеют свой порядковый номер, при счете слева направо. Условимся, что первый слева признак имеет номер один. Именно эти порядковые номера задают конкретные признаки из исходного файла. Можно указывать диапазоны признаков через тире без пробела, например **«2-10»**, **«2-»**, **«-10»**. В двух последних случаях задаются все признаки справа и слева, соответственно, от указанного.

Введение пустой строки осуществляет переход к вводу признака другой шкалы. Первыми вводятся признаки шкалы наименований (содержащие только качественные описания в виде текстовых строк), вторыми – порядковой (содержащие балльные описания в виде чисел в диапазоне от 0 до 999), последними – сильной (значения в диапазоне от -10^{34} до 10^{34}, не более 3 знаков после запятой, порядок от 10^{-34} до 10^{34}).

В случае если задается принадлежность признака одновременно двум шкалам, будет выдано сообщение об ошибке.

Приведем пример:

Enter the range of the calculated name scale (A-B, blank string-finish)1

Enter the range of the calculated name scale (A-B, blank string-finish)

Enter the range of the calculated order scale (A-B, blank string-finish)2

Enter the range of the calculated order scale (A-B, blank string-finish)

Enter the range of the calculated strong scale (A-B, blank string-finish)3-4

Enter the range of the calculated strong scale (A-B, blank string-finish)

В случае задания в п. 1.2 нескольких областей общей нормировки, например, для сильной шкалы, вначале будут вводиться признаки области общей нормировки 1 сильной шкалы, затем области общей нормировки 2 сильной шкалы и т.д. Это будет выглядеть примерно так:

Enter the range of the calculated strong scale scope 1 (A-B, blank string-finish)3

Enter the range of the calculated strong scale scope 1 (A-B, blank string-finish)

Enter the range of the calculated strong scale scope 2 (A-B, blank string-finish)4

Enter the range of the calculated strong scale scope 2 (A-B, blank string-finish)

Для перехода к введению признаков следующей области вводится пустая строка.

В случае ошибочного задания принадлежности признака двум шкалам выводится подобное сообщение:

Enter the range of the calculated name scale (A-B, blank string-finish)-

Enter the range of the calculated name scale (A-B, blank string-finish)

Enter the range of the calculated order scale (A-B, blank string-finish)-

Enter the range of the calculated order scale (A-B, blank string-finish)

Enter the range of the calculated strong scale (A-B, blank string-finish)-

Coordinate number 1 is already calculated order scale

Coordinate number 2 is already calculated order scale

Coordinate number 3 is already calculated order scale

Coordinate number 4 is already calculated order scale

Однако, вследствие ошибок в контроле правильности вводимых значений, контроль за двойной принадлежностью признака срабатывает не всегда. В этом случае выбирается последняя введенная шкальная принадлежность признака.

1.4. Задание типа кластеризации – объектов или признаков.

Под признаками подразумеваются столбцы исходного файла, под объектами – строки. Но различие кластеризации объектов и признаков состоит не только в том, как выбираются данные из таблицы. Для признаков и объектов используются совершенно разные способы нахождения расстояний. При кластеризации признаков частично учитывается взаимное влияние объектов друг на друга. При кластеризации объектов программа считает вклад каждого признака линейным. Взаимная зависимость признаков (нелинейный вклад групп признаков) при кластеризации объектов не учитывается.

На вопрос программы:

Clusterization of objects or attributes? (o/a):

Следует ввести «**o**»+**enter** (кластеризация объектов) или «**a**»+**enter** (кластеризация признаков). Буквы должны вводиться при включенной английской раскладке клавиатуры.

В случае кластеризации признаков — переход на п. 1.4.2.

1.4.1. Вопросы программы в случае кластеризации объектов.

1.4.1.1. Включение возможности задавать раздельный вклад в общую кластеризацию каждой из областей общей нормировки.

Программа задает следующий вопрос:

Use separate lambda for each scope of common normalization? (y/n):

В случае положительного ответа на него (английская **«y»**+ **«enter»**) впоследствии можно будет задать вклад в общую кластеризацию для каждой из областей общей нормировки отдельно (см п. 1.2, 1.3).

1.4.1.2. Задание выводимых признаков кластеров.

Данная группа вопросов используется, чтобы наглядно представить, почему точки группируются в кластеры, наблюдая за значениями признаков объектов, попавшими в один кластер. По выводимым значениям признаков также могут быть рассчитаны меры разброса – некоторая мера качества кластеризации. Возможна также такая комбинация ответов на вопросы, при которой значения признаков объектов не выводятся, но меры разброса рассчитываются. Выводимые на печать признаки не обязательно должны совпадать с признаками для расчета кластеризации (п. 1.3), но если совпадают, должны быть той же шкалы. Если не нужно выводить значения признаков, нужно просто три раза нажать клавишу «**enter**».

Вопросы, задаваемые программой, подобны вопросам в п. 1.3. Единственным исключением является то, что не различаются области нормировки, т.е. для вывода значений признака не важно, как нормируется он при расчете. Поэтому вводится только принадлежность признака к одной из трех шкал, например:

Enter the range of the output name scale (A-B, blank string-finish)1

Enter the range of the output name scale (A-B, blank string-finish)

Enter the range of the output order scale (A-B, blank string-finish)2

Enter the range of the output order scale (A-B, blank string-finish)

Enter the range of the output strong scale (A-B, blank string-finish)3-4

Enter the range of the output strong scale (A-B, blank string-finish)

В случае успешного окончания этого блока ввода выводится примерно следующее сообщение:

Scanning source file

5vectors

1.4.1.3. Возможность отключения вывода значений признаков, не отключая расчет мер разброса.

Вопросом, задаваемым программой пользователю на этом этапе, является:

Suppress output of object values? (y/n):

На вопрос можно ответить «Да/Нет» (**y**+**enter** или **n**+**enter**). В случае положительного ответа на вопрос вывод значений признаков отключается, но, тем не менее, по заданным в п. 1.4.1.2 признакам рассчитываются меры разброса.

При удачном задании параметров во всех предыдущих блоках ввода данных выводится такое сообщение:

Reading source file

100.%

При наличии пропусков в таблице выводится примерно такое сообщение:

Reading source file

60.% 3/ 4coordinates in vectorттттт

В этом случае в строке с названием «**ттттт**» не хватает данных (есть пропуск).

В случае ошибки в формате значений данных или не попадании данных в диапазон допустимых значений выводится примерно следующее сообщение:

Reading source file

50.%

Error in coordinate2_3 vectorттттт

В указанном случае ошибка произошла в строке с названием «ттттт», причем проблемы возникли при считывании признака с названием «2_3».

1.4.1.4. Использование специальной функции для расчета мер близости в случае наличия сильной шкалы с одной общей областью нормировки.

Эта функция использовалась для очень специфичной задачи и для пользователя не нужна. На вопрос:

Use function for scopes of strong scale during calculation of measures of

closeness? (y/n):

следует отвечать отрицательно (**n+enter**).

1.4.2. Вопросы, задаваемые в случае кластеризации признаков.

1.4.2.1. Предложение альтернативного расчета для шкалы наименований.

При кластеризации признаков особенно плохо получается кластеризация шкалы наименований при нормировке по столбцам. Для улучшения разделения признаков шкалы наименований по кластерам рекомендуется попробовать положительно ответить на следующий вопрос

Treat name scale as order scale? (y/n):

Вопрос задается только при указании кластеризации признаков шкалы наименований.

1.4.2.2. Задание выводимых значений объектов.

Кластеризация признаков производится по всем объектам, т.е. в отличие от кластеризации объектов, нельзя для расчета выбрать кусочек исходного файла. Для наглядного представления кластеризации признаков и расчета мер разброса при кластеризации признаков используется этот блок ввода данных. Программа задает пользователю следующий вопрос:

Enter the range of the output objects (A-B, blank string-finish)

Объекты нумеруются так же, как и признаки. Номер один соответствует самому верхнему объекту в файле. В ответ на

данный вопрос программы вводится диапазон номеров объектов или «-» для задания всех объектов. Пустая строка означает отсутствие вывода значений объектов. Ввод значений аналогичен п.п. 1.3, 1.4.1.2, например:

Enter the range of the output objects (A-B, blank string-finish)-

Для окончания ввода диапазона объектов используется введение пустой строки.

1.4.2.3. Задание возможности расчета мер разброса при отсутствии выведения значений объектов в выходной файл.

Реакция программы на этот ответ описана в п. 1.4.1.3. Вопросом программы является

Suppress output of object values? (y/n):

Ответ на вопрос аналогичен п. 1.4.1.3. На вопрос рекомендуется отвечать «нет», так как количество объектов обычно велико, что приведет к перегрузке выходного файла данными.

1.5. Введение вклада разных шкал в общую кластеризацию.

При наличии нескольких шкал одновременно иногда возникает ситуация, когда признаки разных шкал описывают исследуемую систему с совершенно разных сторон. В этом случае можно добиться улучшения качества кластеризации, наблюдая за изменением качества получаемых результатов при варьировании вклада разных шкал в результат.

Вопросом программы является

Enter lambda for name scale (3 digits max, others rejected):

Enter lambda for order scale (3 digits max, others rejected)::

В ответ надо вводить вклады в диапазоне от 0 до 1, не более 3 знаков после запятой, и нажимать клавишу «**enter**». Сумма всех вкладов не должна быть больше 1. По введенным вкладам для шкал наименований и порядка вычисляется вклад для сильной шкалы и выдается, например, следующее сообщение:

lambda for strong scale equal .300

В случае положительного ответа в п. 1.4.1.1 может быть выдан вопрос следующего вида:

Enter lambda for strong scale (3 digits max, others rejected): scope 1:

В данном вопросе подразумевается, что пользователь будет вводить вклад по первой области общей нормировки сильной шкалы.

1.6. Ввод поправочного коэффициента при кластеризации признаков шкалы наименований без общей нормировки.

При кластеризации признаков слабых шкал без общей нормировки все точки обычно объединяются в один-два кластера. Это объясняется тем, что при расчете мер близости используются не сами значения признаков, а их частотные описания. При этом маловероятно, что программа получит много равных частот для двух разных признаков. Поэтому, для шкалы наименований, парные отношения будут в основном «не равно» + «не равно». Это значит, что расстояние между двумя данными признаками будет практически 0. Это же будет выполняться для всех остальных пар признаков шкалы наименований.

Чтобы исключить эту ситуацию, мы используем нестрогое равенство. Возьмем некоторую долю от максимальной из двух сравниваемых частот, если эта частота больше единицы. Вычтем из максимальной частоты эту долю. Если меньшая частота будет больше полученной разности, то частоты будем считать равными. Если одна из частот равна 1, а другая 0, то также будем считать их равными. Например, имеем следующие частотные описания признаков 1 и 4:

	1	4
1	2	4
2	1	3
3	1	0
4	0	2

Тогда получим для них следующие вклады парных отношений в меру близости (табл. 6).

Таблица 6

Пара		Частоты				Парное отношение		Вклад в расстояние между признаками
		Объект 1		Объект 2				
Объект x	Объект y	Признак 1	Признак 2	Признак 1	Признак 2	Объект 1	Объект 2	
1	2	2	4	1	3	≠	≠	0
1	3	2	4	1	0	≠	≠	0
1	4	2	4	0	2	≠	≠	0
2	3	1	3	1	0	≠	≠	0
2	4	1	3	0	2	≠	≠	0
3	4	1	0	0	2	≠	≠	0

Возьмем долю, равную 0,5. Тогда парные отношения преобразуются в следующий вид (табл. 7).

Таблица 7

Пара		Частоты				Парное отношение		Вклад в расстояние между признаками
		Объект 1		Объект 2				
Объект x	Объект y	Признак 1	Признак 2	Признак 1	Признак 2	Объект 1	Объект 2	
1	2	2	4	1	3	=	≠	1
1	3	2	4	1	0	=	=	0
1	4	2	4	0	2	=	≠	1
2	3	1	3	1	0	≠	=	1
2	4	1	3	0	2	≠	≠	0
3	4	1	0	0	2	=	≠	1

Мера близости будет уже $3\times2/(4(4-1)) = 0,5$.

При кластеризации признаков слабых шкал максимально возможная мера близости может быть от 0,5 до 1 в зависимости от объем выборки. Для выборки из 4 объектов максимальная мера близости будет 2/3. Получив для частотного описания признаков 1 и 4 меру близости 0,5 при доле 0,5, мы показали, что эти признаки не сходны, тогда как

при отсутствии доли эти признаки получаются полностью сходными. Таким образом, при правильном подборе доли, в один кластер попадут два идентичных признака или два признака, один из которых не сильно меняется по объектам, или признаки, по объектам которых стоят близкие частоты. Если не существует системы соответствия частот двух признаков, они попадают в разные кластеры.

Доля в программе называется мерой мягкости. Вводится мера мягкости после следующего вопроса:

Enter degree of softness (0.200-0.950):

Обычно нормальная мера мягкости больше 0,5 и подбирается эмпирически. Можно ввести 0 или значение в диапазоне от 0,2 до 0,95. В случае введения 0 мера мягкости не используется.

1.7. Ввод точности представления меры близости.

Значения матрицы расстояний (промежуточный результат работы программы, который содержит расстояния между каждой парой объектов или признаков) изменяются от 0 до 1 и могут содержать от 0 до 4 знаков после запятой. Чем меньше исходная ошибка данных, чем меньше число используемых признаков при кластеризации объектов, чем больше признаков слабых шкал по сравнению с признаками сильных шкал, или чем меньше число объектов при кластеризации признаков, тем большее число знаков после запятой следует задавать. На вопрос:

Enter number of digits after comma for measures of closeness (0-4):

Следует ввести целочисленное значение от 0 до 4.

1.8. Группа вопросов по оформлению выходного файла.

Output closeness to centers of clusters? (y/n):

В случае положительного ответа на этот вопрос (**y**+**enter**) для каждого объекта выходного файла будет выводиться мера близости до центра кластера, локальная и глобальная

Output deviation from mean for strong scale? (y/n):

При положительном ответе на этот вопрос (**y**+**enter**) для каждого объекта сильной шкалы и для каждого его признака сильной шкалы будет выводится отклонение от средне-

го, рассчитанного для каждого признака сильной шкалы по кластеру.

Output correlation between vectors? (y/n):

Включение вывода таблиц коэффициентов корреляции между объектами внутри кластера, при условии, что кластеризация производится только по сильной шкале. Положительный ответ — **y+enter.**

Output correlation measure of closeness? (y/n):

Величина, равная 1 — коэффициент корреляции. Положительный ответ – **y+enter.**

Output cosine of angle between vectors? (y/n):

Коэффициент корреляции для центрированных векторов сильных значений внутри кластера. Положительный ответ – **y+enter.**

Output cosine measure of closeness? (y/n):

Величина, равная 1 – косинус. Положительный ответ — **y+enter.**

1.9. Выбор алгоритма кластеризации (п. 1.3).

Enter algorithm - closest pair/maximum frame/optimum domain (p/f/d):

p+enter – ближайшая пара (алгоритм-1), позволяет находить кластеры, самостоятельно учитывая структуру метрического пространства, а затем при необходимости объединять кластеры в более крупные.

f+enter – максимальный каркас (алгоритм-2), позволяет выделить задаваемое заранее число максимально удаленных друг от друга кластеров.

d+enter – максимальная область (алгоритм-3), находит кластеры с максимальной плотностью заселения, ограниченные задаваемым заранее радиусом.

Следует отметить, что алгоритм-1 пытается найти кластеры, одновременно максимально плотные и максимально удаленные друг от друга, но если требуются просто наиболее плотные кластеры или просто максимально удаленные, следует использовать алгоритм-2 или алгоритм-3 соответственно.

В случае алгоритма-2 или алгоритма-3 переход на п. 1.10.

1.9.1. Выбор наивысшего уровня кластеризации в случае выбора алгоритма-1.

Enter the level of clustering:

Ввести целочисленный уровень кластеризации + **enter**.

Уровень кластеризации может быть целым числом от 1 и выше. 1-й уровень кластеризации получается при работе алгоритма-1 без агрегирования кластеров. Более высокие уровни используются при бóльших выборках данных для упрощения анализа. Качество кластеризации на более высоких уровнях уступает качеству кластеризации на первом уровне.

1.9.2. Задание мер близости при кластеризации выше первого уровня.

Enter type of measure of closeness in clustering of clusters - hard/ soft/

both (h/s/b):m

h+enter – жесткие меры близости для всех уровней (грубая оценка мер близости между кластерами, полученными на младших уровнях. Применима при большой погрешности данных, большом количестве признаков при кластеризации объектов или при большом количестве объектов при кластеризации признаков, при малом размере кластера на последнем уровне кластеризации, а также при большой доле признаков сильной шкалы при кластеризации объектов).

s+enter – мягкие меры близости для всех уровней (применяются при малой погрешности данных, малом количестве признаков при кластеризации объектов, малом количестве объектов при кластеризации признаков, большом объеме кластера на последнем уровне, большой доле признаков слабых шкал).

b+enter – задание меры близости в зависимости от уровня

1.9.2.1. Поуровневое задание агрегированных мер близости для матриц расстояний.

В ответ на вопросы приведенного ниже вида:

Matrix of distances level 1: measures of closeness soft or minimum? (s/m):

Matrix of distances level 2: measures of closeness soft or minimum? (s/m):

Можно вводить **s+enter** (мягкая мера близости) или **m+enter** (жесткая мера близости).

1.10. Задание возможности вывода в файл матрицы расстояний.

Output matrix of lengths? (y/n):

В случае положительного ответа (**y+enter**) матрица расстояний (содержащая расстояния, полученные программой между любыми двумя объектами или признаками по используемой метрике) будет выведена в файл.

1.10.1. Задание количества выводимых матриц расстояний.

Output all levels of matrix of lengths? (y/n):

В случае задания вывода в файл матрицы расстояний положительный ответ на этот вопрос (**y+enter**) задает вывод в файлы матриц расстояний всех уровней.

1.10.1.1. Задание уровня выводимой в файл матрицы расстояний.

Enter the level of output matrix of length:

В случае отрицательного ответа на вопрос в п. 1.10.1 в ответ на данный вопрос можно ввести уровень выводимой в файл матрицы расстояний. Самая первая матрица расстояний имеет номер 0. Самая верхняя матрица расстояний (между полученными кластерами) имеет номер уровня кластеризации.

1.11. Задание вывода в файл мер разброса (нецентральных для сильной шкалы).

Calculate scattering characteristics? (y/n):

В случае положительного ответа на этот вопрос (**y+enter**) в файл будут выведены меры разброса по каждому признаку (для кластеризации объектов) или по каждому объекту (для кластеризации признаков), указанных в п.п. 1.4.1.2, 1.4.2.2.

1.12. Задание вывода в файл средних по всем объектам (или по всем признакам) мер разброса.

Calculate average scattering characteristics? (y/n):

y+enter – положительный ответ.

Для анализа качества кластеризации иногда проще посмотреть среднюю меру разброса по всем признакам данного кластера при кластеризации объектов, чем смотреть меру разброса по каждому признаку, особенно если признаков много. То же верно для кластеризации признаков.

1.12.1. Исключение признаков из средних мер разброса (только для кластеризации объектов).

Do you want to exclude some attributes from average scattering characteristics? (y/n):

Положительный ответ на вопрос в п. 1.12 вызывает данный вопрос. В случае положительного ответа (**y+enter**) впоследствии можно задать, какие признаки не нужно учитывать при расчете средних мер разброса. Например, в п. 1.4.1.2 задан признак, не участвующий в кластеризации и не несущий никакой информации, например порядковый номер. Включение меры разброса по порядковому номеру в среднюю меру разброса исказит ее, поэтому ее надо исключить из средней меры разброса.

1.13. Включение в выходной файл локально минимального и максимального расстояния между точками.

Calculate local minimum and maximum of closeness? (y/n):

В случае положительного ответа на вопрос (**y+enter**), экстремумы мер близости будут выводиться не только для всего кластера, но и для подкластеров.

1.14. Включение вывода выходного файла анализа по центрам Чебышева.

В случае положительного ответа на вопрос

Provide centers of Chebishev analysis? (y/n):

после вывода файла кластеризации будет выведен файл со стандартной шапкой, в котором будут перечислены по порядку объекты или признаки, кластеры, к которым они принадлежат, а также кластеры, к центрам Чебышева которых данные объекты/признаки ближе находятся.

При расчете по алгоритму-1,2 переход на п. 2.2.15.

1.15. В случае кластеризации по алгоритму-3 введение количества свободной ОЗУ компьютера для хранения временных результатов.

Enter amount of free operational memory, Mb:

Вводим целочисленное значение, примерно равное объему ОЗУ компьютера в мегабайтах, деленному пополам.

1.15.1. Возможность проведения черновых вычислений с целью подбора оптимального радиуса области по алгоритму-3.

Perform test calculations? (y/n):

Положительный ответ – **y+enter.**

При кластеризации по алгоритму-3 заранее неизвестно, какой радиус области позволит получить качественную кластеризацию. При неправильном задании радиуса получается либо много кластеров, содержащих мало точек, либо слишком мало кластеров, что неудобно при анализе результатов. Черновые вычисления позволяют увидеть, при каком радиусе кластеризация лучше.

1.15.1.1. Введение имени файла черновых расчетов для алгоритма-3.

Enter the name of file for intermediate results:

Введение имени файла описано в п.п. 1.1, 1.16.1.

1.15.1.2. Задание ориентировочного радиуса области.

Minimum radius of domain = .200 Maximum radius of domain = .600
Enter the radius of domain:

В данной строке показывается минимально возможный и максимально возможный радиус области. Пользователь должен ввести число в этом интервале с тремя знаками после запятой.

1.15.1.2.1. Точное задание чернового радиуса области.

Так как мы имеем дело с конечными значениями в матрице расстояний, введение радиуса области, не содержащегося в матрице расстояний, будет приводить к неопределенным результатам расчета. Более предсказуемым будет расчет с введением радиуса области, содержащегося

в матрице расстояний. В п. 1.15.1.2 введен ориентировочный радиус области. В текущем параграфе задается либо ближайший сверху, либо ближайший снизу радиус от ориентировочного:

Upper bound of radius: .400 Lower bound of radius: .200 Choose upper or lower bound? (l/u):

l+enter задает радиус, нижний ближайший к ориентировочному, **u+enter** – верхний.

1.15.1.3. Возможность повторить черновые вычисления для алгоритма-3.

Continue intermediate calculations? (y/n):

Положительный ответ – **y+enter**.

1.16. Введение имени выходного файла.

1.16.1. Введение имени файла получаемой матрицы расстояний.

Enter the name of matrix file (level 0):

В имени выходного файла запрещается использовать знаки *, ?,». Можно задавать путь к файлу, но если пути не существует, будет выдано сообщение:

Invalid path name

Если файл с таким же именем, как заданный, существует, будет выдана строка:

File already exist, overwrite? (y/n)

В случае положительного ответа предыдущий файл будет стерт. Вообще, по имени входного файла действуют те же правила, что и в п. 1.1. В случае необходимости выведения матриц расстояний для всех уровней такой вопрос будет задан для каждого уровня матриц расстояний.

1.16.2. Введение имени файла получаемой кластеризации.

Enter the name of clustering file:

Все аналогично п. 1.16.1, за исключением того, что кластеризация выводится только для последнего уровня. В случае кластеризации по алгоритму-1, переход на п. 1.16.2.3, для алгоритма-2 переход на п. 1.16.2.2.

1.16.2.1. В случае кластеризации по алгоритму-3 введение радиуса области.

Minimum radius of domain = .200 Maximum radius of domain = .600
Enter the radius of domain:

Аналогично п. 1.15.1.2.

1.16.2.2. В случае кластеризации по алгоритму-2 введения числа узлов максимальной решетки.

Enter the number of frame tangles:

Числом узлов максимальной решетки является число кластеров, которое пользователь хочет получить в результате.

1.16.2.3. Запрос на выведение в файл нецентрированной кластеризации.

Если имя выходного файла было задано правильно, программа спрашивает, выводить ли в файл нецентрированную кластеризацию последнего уровня.

Output initial clustering of clusters?(y/n):

Если вы отвечаете положительно (**y+enter**), в файл будет выведен состав кластеров последнего уровня без центрирования (центрирование обычно улучшает качество кластеризации, так как при этом форма кластеров более приближается к сфероиду).

1.16.2.4. Запрос на выведение в файл центрированной кластеризации).

Output centered clustering of clusters?(y/n):

Аналогично п. 1.16.2.3, только в файл выводится центрированная кластеризация.

1.17. Задание имени файла анализа близости к центрам Чебышева.

Enter the name of file for the centers of Chebishev analysis:

Все правила аналогичны п. 1.16.1. Файл выводится только для последнего уровня кластеризации.

1.18 Запрос о введении имени файла матрицы расстояний между полученными кластерами (самый верхний уровень).

Аналогично п. 1.16.1. Матрица выводится, если на последнем уровне больше одного кластера. В случае алгоритма-3 — конец работы, для алгоритма-1 переход на п. 1.20.

1.19. В случае кластеризации по алгоритму-2 возможность повторить расчет для другого числа узлов решетки.

Continue calculation? (y/n)

Положительный ответ (**y**+**enter**) позволяет сделать кластеризацию по алгоритму-2 еще раз, возможно с другим числом ожидаемых на выходе кластеров.

1.20. Возможность осуществить глобальное центрирование для многоуровневой кластеризации.

В случае положительного ответа на вопрос (**y**+**enter**)

Output clustering of clusters centered by the first level?(y/n):

для каждого кластера последнего уровня будет произведено центрирование и результаты будут выведены в выходной файл (п. 1.16.2). Центрирование обычно производится над кластерами предпоследнего уровня, с целью улучшения качества кластеризации. В данном случае положительный ответ включает центрирование не кластеров, а содержащихся в них точек, учитывая состав кластеров последнего уровня. Это позволяет улучшить качество кластеризации по сравнению с обычным центрированием, но при это теряется структурный состав кластеров.

1.20.1. Выведение информации о принадлежности центрам Чебышева.

Enter the name of file for the centers of Chebishev analysis:

1.20.2. Задание мер близости для матрицы расстояний между кластерами, случай п. 1.20, если задан положительный ответ в п. 1.10.1.

Enter type of measure of closeness in clustering of clusters centered

by the first level - hard/soft (h/s):

h+**enter** — жесткие меры близости, **s**+**enter** — мягкие меры близости.

1.20.3. Задание имени файла выходной матрицы расстояний последнего уровня по п. 1.18.

Аналогично п. 1.16.1.

1.21. Конец работы программы.

2. Описание выходных файлов программы.

2.1. Шапка выходных файлов программы.

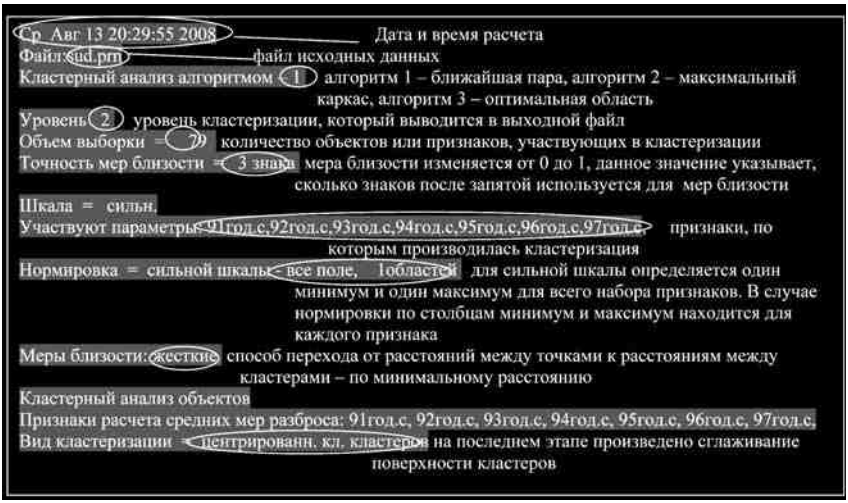

Рис. 32

Серым цветом указаны строки, присутствующие в выходном файле.

В зависимости от введенных параметров (п. 1), значение шапки может меняться, например, при кластеризации признаков появляется строка «число объектов», а объем выборки показывает число признаков. Также может быть выведен, в зависимости от заданных параметров, радиус области (алгоритм-3), степень мягкости (кластеризация признаков).

2.2. Файл матрицы расстояний.

В файле матрицы расстояний за шапкой следуют значения матрицы, разделенные пробелами (импортировать в Excel как

файл с разделителями данных). Названия точек присутствуют слева, справа, сверху и снизу.

2.3. Промежуточный файл алгоритма-3.

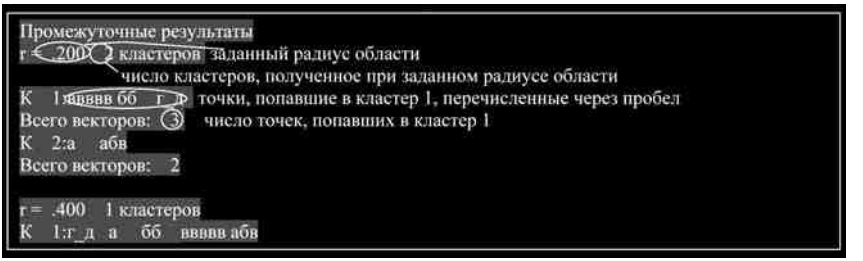

Рис. 33

Серым цветом указаны строки, присутствующие в выходном файле.

По такому файлу можно увидеть, при каком радиусе области точки перестают сливаться в один кластер и в то же время не распадаются на слишком маленькие кластеры.

2.4. Файл анализа центров Чебышева.

Рис. 34

После строки «Анализ центров Чебышева» следует шапка таблицы. Первый столбец (без заголовка) содержит кластеризованные точки. Второй столбец, озаглавленный «Прин.» — значения кластеров, в которые входят эти точки. Далее для каждого кластера имеется свой столбец. В нашем случае кластеров 2. Если в соответствующем столбце для данной точки стоит «*****» — данная точка ближе всего к центру Чебышева именно этого кластера, номер его можно смотреть в шапке этого столбца. Данные разделены пробелами и могут быть импортированы в Excel как данные с разделителями.

2.5. Файл кластеризации.

2.5.1. Отражение состава кластеров.

Рис. 35

2.5.1.1. Серым цветом обозначены строки в выходном файле.

Файл кластеризации отражает группировку подкластеров на всех уровнях кластеризации. Иерархия подкластеров в кластере отражается символом «.», например «подкластер 1.3.2» означает подкластер 2, входящий в подкластер 3, входящий в подкластер 1. Также отражается принадлежность точки к кластеру, например в строках по экстремуму расстояний в кластерах. При включении отклонений от среднего для сильной шкалы (п. 1.8) справа от значения признака для

данного объекта выводится отклонение, а последней строкой кластера – средние значения. Файл можно импортировать в Excel как файл со столбцами фиксированной ширины.

2.5.2. Итоги файла кластеризации.

```
ИТОГО ПО ВСЕМ КЛАСТЕРАМ:
Минимальная мера разброса по 1  .н = .000
Минимальная мера разброса по 2_3.п = .000
Минимальная мера разброса по в  .с = .227
Минимальная мера разброса по г  .с = .000
Минимум усредненной меры разброса:
По шкале наименований: .000
По шкале порядка: .000
По сильной шкале: .000
Максимальная мера разброса по 1  .н = .500
Максимальная мера разброса по 2_3.п = .667
Максимальная мера разброса по в  .с = .259
Максимальная мера разброса по г  .с = .000
Максимум усредненной меры разброса:
По шкале наименований: .500
По шкале порядка: .667
По сильной шкале: .000
Средняя мера разброса по 1  .н = .250
Средняя мера разброса по 2_3.п = .333
Средняя мера разброса по в  .с = .243
Средняя мера разброса по г  .с = .000
Среднее по кластерам усредненной меры разброса:
По шкале наименований: .250
По шкале порядка: .333
По сильной шкале: .000
```

Рис. 36

2.5.1.2. Пример итоговой таблицы в конце выходного файла кластеризации.

Как видно из рис. 36, по всем кластерам выбирается минимум, максимум и среднее по мерам разброса, если меры разброса включены. В зависимости от настроек программы также

могут выводиться и экстремумы по усредненным мерам разброса.

3. Основные особенности программной реализации комплекса Clust:

1) возможность нахождения ошибок при вводе исходных данных;
2) работа в режиме командной строки;
3) одновременная работа с тремя типами шкал измерения данных:
   Н-шкала с длинной строкой;
   П-шкала от 0 до 999;
   О-шкала данные от 10^{-34} до 10^{34}.

На материале рассмотренных выше задач Clust1 показал следующие важные свойства:

1) подчеркивание «маргинальных» объектов: Глава 2: 6, 9, 11;
2) выстраивание группы анализируемых объектов в виде пелотона, когда они скапливаются в головной и/или хвостовой части когорты: Глава 2: 12;
3) в большинстве кластерных систем оптимальность достигается для $t^* = 2$ при минимизации обеих функций

$$f_1(t) = N'_t + \ell_t; f_2(t) = \max(N'_t, \ell_t),$$

что означает уменьшение объемов кластеров объектов более чем в 10 раз.

Литература

1. *Судаков С.А.* О применении линейных и сферических разделителей к решению задачи распознавания изображений «Читающие устройства». — М.: ВИНИТИ АН СССР, 1965.

2. *Судаков С.А.* О некоторых процессах «обучения» в системах распознавания изображений «Читающие устройства». — М: ВИНИТИ АН СССР, 1965

3. *Судаков С.А., Катинский В.С., Романычева Т.К.* Анализ изображений путем представления их графами // Вопросы радиоэлектроники. — 1966. — Сер. 7, вып. 6.

4. *Судаков С.А.* О покрытиях точечных множеств при построении эталонов классов изображений // Вопросы радиоэлектроники. — 1967. — Сер. ЭВТ, вып. 8.

5. *Судаков С.А., Катинский В.С., Романычева Т.К.* Описание изображений с помощью их представления графами // Труды III Всесоюзной конференции по информационно-поисковым системам и автоматической обработке информации. — Т. 3. — М.: ВИНИТИ АН СССР, 1967.

6. *Судаков С.А.* О представлении многомерных распределений графами в задачах распознавании образов. «Структурные методы опознания и автоматического чтения». — М.: ВИНИТИ АН СССР, 1970. — С. 91–98.

7. *Судаков С.А., Зайцев-Зотов В.И.* Оптическое читающее устройство // Тезисы докладов IV Всесоюзной конференции «Автоматизация ввода письменных знаков в ЦВМ». — Т. 1. — Каунас: КПИ, 1977.

8. *Судаков С.А.* Об оптимизации параметров одной системы эталонов читающего устройства // Вопросы радиоэлектроники. — 1978. — Сер. ЭВТ, вып. 8. — С. 63–69.

9. *Судаков С.А.* Об одном классе моделей с квадратичными решающими правилами в задачах автоматической классификации // Вопросы радиоэлектроники. — 1982. — Сер. ЭВТ, вып. 9. — С. 46–52.

10. *Судаков С.А., Зайцев-Зотов В.И.* Свойства решающего правила в детерминистской модели машинописного знака с переменной толщиной гра-

фических элементов и контрастностью // Вопросы радиоэлектроники. — 1982. — Сер. ЭВТ, вып. 9. — С. 27–31.

11. *Судаков С.А., Зайцев-Зотов В.И.* Влияние формы границ на решающее правило в модели машинописного знака с переменной шириной графических элементов // Вопросы радиоэлектроники. — 1982. — Сер. ЭВТ, вып. 9. — С. 32–37.

12. *Судаков С.А.* Оптимальное решающее правило в статистической модели машинописного знака с переменной шириной графических элементов и контрастностью // Вопросы радиоэлектроники. — 1982. — Сер. ЭВТ, вып. 9. — С. 38–40.

13. *Судаков С.А.* Об аппроксимации многомерных распределений вероятностей на целочисленной решетке в задачах автоматической классификации // Вопросы радиоэлектроники. — 1982. — Сер. ЭВТ, вып. 9. — С. 111–117.

14. *Судаков С.А.* О сходимости к многомерному распределению на целочисленной решетке линейной комбинации его частных распределений // Вопросы радиоэлектроники. — 1984. — Сер. ЭВТ, вып. 8. — С. 33–39.

15. *Судаков С.А., Зайцев-Зотов В.И.* Математическая модель процесса контроля качества машинописных текстов с помощью оптического читающего устройства // Вопросы радиоэлектроники. — 1984. — Сер. ЭВТ, вып. 8. — С. 56–60.

16. *Судаков С.А.* Об одном классе приближений многомерных распределений для изображений машинописных знаков // Тезисы докладов V Всесоюзной конференции «Автоматизация ввода письменных знаков в ЦВМ». — Т. 2. — Каунас: КПИ, 1984.

17. *Судаков С.А., Каталов Ю.Т.* Анализ системы команд процессора с использованием структурных моделей // Вопросы радиоэлектроники. — 1986. — Сер. ЭВТ, вып. 12. — С. 24–28.

18. *Судаков С.А.* О верхней оценке погрешности аппроксимации дискретного многомерного распределения линейной комбинацией его частных распределений // Вопросы радиоэлектроники. — 1984. — Сер. ЭВТ, вып. 12.

19. *Судаков С.А., Калинин А.В.* Измерение параметров качества печати методом распознавания образов // Тезисы докладов III Всесоюзной конференции «Математические методы распознавания образов». — Т. 2. — Львов, 1987.

20. *Судаков С.А., Калинин А.В.* Оценка качества печати машинописных знаков на основе распределения яркости в оттиске // Вопросы радиоэлектроники. — 1988. — Сер. ЭВТ, вып. 8.

21. *Судаков С.А.* О приближении дискретного распределения вероятностей дискретным распределением с заданным числом одинаковых значений // Вопросы радиоэлектроники. — 1988. — Сер. ЭВТ, вып. 8. — С. 142–147.

22. *Судаков С.А., Томченко С.Н.* Качество изображений как оценка параметров распределений методом максимального правдоподобия // Вопросы радиоэлектроники. — 1989. — Сер. ЭВТ, вып. 8.

23. *Судаков С.А., Калинин А.В., Хатунцев А.Г.* Оценка качества печати по оцифрованным оттискам литер заданного шрифта // Вопросы радиоэлектроники. — 1988. — Сер. ЭВТ, вып. 8.

24. *Судаков С.А., Катинский В.С.* Метод классификации текстов, введенных в ЭВМ с помощью системы оптического распознавания // Вопросы радиоэлектроники. — 1991. — Сер. ЭВТ, вып. 8.

25. *Судаков С.А.* Автоматический поиск строчной структуры печатного текста как задача обнаружения сигнала в смеси с шумом // Вопросы радиоэлектроники. — 1992. — Сер. ЭВТ, вып. 8.

26. *Судаков С.А., Дыдычкин В.Е.* Об оптимальном представлении многоуровневого изображения знака двухуровневым в алгоритмах автоматического чтения // Вопросы радиоэлектроники. — 1992. — Сер. ЭВТ, вып. 8.

27. *Плюта В.* Сравнительный многомерный анализ в экономических исследованиях. — М.: Статистика, 1980.

28. *Мандель И.Д.* Кластерный анализ. — М.: Финансы и статистика, 1988.

29. *Каменский В.С.* Методы и модели неметрического шкалирования: Обзор // Автоматика и телемеханика. — 1977. — № 8.

30. *Уемов А.И.* Логические основы метода моделирования. — М.: Мысль, 1973.

31. *Миркин Б.Г.* Анализ качественных признаков и структур. — М.: Статистика, 1980.

32. *Орлов А.И.* Некоторые вероятностные вопросы теории классификации. Прикладная статистика. — М.: Наука, 1983.

33. *Апресян Ю.Д.* Алгоритм построения кластеров по матрице расстояний // Машинный перевод и прикладная лингвистика. — 1966. — Вып. 9.

34. *Миркин Б.Г.* Группировки в социально-экономических исследованиях. — М.: Финансы и статистика, 1985.

35. *Шлезингер М.И.* О самопроизвольном различении образов. Читающие автоматы. — Киев: Наукова думка, 1965.

36. *Диде Э.* Методы анализа данных / Пер. с франц. — М.: Финансы и статистика, 1985.

37. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. — М.: Статистика, 1974.

38. *Бауман Е.М.* Методы классификационной обработки в задачах экспертизы. I Всесоюзное совещание по статистическому и дискретному анализу нечисловой информации, экспертным оценкам и дискретной оптимизации. — М.: Алма-Ата, 1981.

39. *Розова С.С.* Классификационная проблема в современной науке. — Новосибирск: Наука, 1986.

40. *Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С.* Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.

41. *Кендал М.* Ранговые корреляции. — М.: Статистика, 1975.

42. *Кемени Дж., Снелл Дж.* Кибернетическое моделирование. — М.: Сов. радио, 1972.

43. *Закс Л.* Статистическое оценивание. — М.: Статистика, 1975.

44. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. — М.: Наука, 1974.

45. *Дюран Д., Оделл П.* Кластерный анализ. — М.: Статистика, 1977.

46. *Налимов В.В.* Теория эксперимента. — М.: Наука, 1971.

47. *Бесчастный А.А., Немцов А.В.* Состояние математизации в психиатрии // Журн. неврол. и психиатрии им. С.С. Корсакова. — 1990. — № 2. — С. 144–146.

48. *Зорин Н.А., Немцов А.В.* Тенденция использования математических методов в психиатрических статьях // Социальная и клиническая психиатрия. — 1998. — № 1. — С. 121–125.

49. *Зорин Н.А., Немцов А.В.* Формализованная экспертная оценка качества исследовательских публикаций в психиатрии // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2001. — № 3. — С. 64–68.

50. *Зорин Н.А., Калинин В.В., Немцов А.В.* Методы оценки качества исследовательских публикаций в психиатрии // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2001. — № 2. — С. 62–67.

51. *Немцов А.В.* Алкогольный урон регионов России. — М.: NALEX, 2003.

52. *Судаков С.А., Амосова А.М.* Нужна ли психиатрам математика? // Журн. неврол. и психиатрии им. С.С. Корсакова. — 1999. — Т. 99. — № 6. — С. 63–64.

53. *Судаков С.А., Трифонов Е.Г.* Применение аппарата проверки статистических гипотез к анализу особенностей контингентов психиатрических

стационаров // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2000. — Т. 100. — № 12. — С. 76–80.

54. *Зверева Н.В., Судаков С.А.* Об универсальной схеме статистической оценки латеральной организации различных функций в нормальной детской популяции // Вестник МГУ. — 2001. — Сер. 14. Психология, 4. — С. 32–38.

55. *Зверева Н.В., Каримулина Е.Г., Судаков С.А.* Функциональная асимметрия в разных модальностях у здоровых и «проблемных» детей // Вестник МГУ. — 2003. — Сер. 14. Психология, 1. — С. 29–34.

56. *Судаков С.А.* Особенности применения числовых статистических методов для нечисловых данных в задачах психиатрии // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2002. — № 2. — С. 51–53.

57. *Судаков С.А., Лебедева И.С., Каледа В.Г.* Применение кластерного анализа при исследовании клинико-нейрофизиологических корреляций // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2003. — № 7. — С. 40–43.

58. *Серебрийская Л.Я., Судаков С.А., Ениколопов С.Н., Мясоедов С.Н.* Кластерный анализ как метод определения феномена стигматизации психически больных // Психиатрия. — 2004. — № 3. — С. 50–54.

59. *Михайлова И.И., Судаков С.А., Ениколопов С.Н., Мясоедов С.Н.* Применение кластерного анализа для описания феномена самостигматизации // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2004. — № 7. — С. 61–65.

60. *Блохина О.А., Ениколопов С.Н., Судаков С.А., Оруджев Я.С.* Психологические аспекты самостигматизации больных шизофренией // Психиатрия. — 2005. — № 1. — С. 26–30.

61. *Судаков С.А., Мясоедов С.Н.* Неклассическая статистика в задачах психиатрии // Психиатрия. — 2005. — № 2. — С. 33–38.

62. *Корсакова Н.Н., Судаков С.А., Мясоедов С.Н.* Кластерный анализ в клинической психологии // Вестник МГУ. — 2006. — Сер. 14. Психология, 3. — С. 85–92.

63. *Божко О.В., Гаврилова С.И., Судаков С.А., Мясоедов С.Н.* Применение кластерного анализа для оценки значимости МРТ-показателей при решении дифференциально-диагностических задач у больных деменцией // Психиатрия. — 2006. — № 3. — С. 55–60.

64. *Немцов А.В., Судаков С.А., Мясоедов С.Н.* Областные показатели алкогольных отравлений и алкогольных психозов // Судебная медицинская экспертиза. — 2003. — № 4. — С. 37–41.

65. *Немцов А.В., Судаков С.А.* Смерть при отравлении алкоголем в регионах Российской Федерации в 1991–1997 годах // Вопросы наркологии. — 2002. — № 5. — С. 65–70.

66. *Тарасова Н.П., Кручина Е.Б., Судаков С.А., Мясоедов С.Н., Беляева М.П.* Оценка вклада базовых показателей в динамику комплексного показателя человеческого потенциала при формировании механизмов устойчивого развития экономических систем // Менеджмент в России и за рубежом. — 2006. — № 2. — С. 17–27.

67. *Серебрийская Л.Я.* Психологические факторы стигматизации психически больных: Автореф. дис. ... канд. психол. наук. — М., 2005.

68. *Михайлова И.И.* Самостигматизация психически больных: Автореф. дис. ... канд. мед. наук. — М., 2005.

69. *Зуева Ю.В.* Нарушение когнитивных процессов при изолированных инфарктах мозжечка: Автореф. дис. ... канд. психол. наук. — М., 2003.

70. *Гонжал О.А.* Клиническая типология самостигматизации при шизофрении: Автореф. дис. ... канд. мед. наук. — Волгоград, 2006.

71. *Кручина Е.Б.* Разработка и применение инструментов мониторинга развития экономических систем народного хозяйства России с использованием показателей качества человеческого потенциала: Автореф. дис. ... канд. эконом. наук.. — М., 2005.

72. *Божко О.В.* Магнитно-резонансная томография подкоркового поражения головного мозга при болезни Альцгеймера: Автореф. дис. ... канд. мед. наук. — М., 2007.

73. *Кузюкова А.А.* Клиника и психопатология манифестных эндогенных психозов юношеского возраста: Автореф. дис. ... канд. мед. наук. — М., 2007.

74. *De Groot, Vizzich J.E.* A celebration of statistics. — N.-Y., 1985. — P. 145–165.

75. *Ястребов В.С., Михайлова И.И., Судаков С.А.* Стигма в психиатрии: скрытая угроза. — М.: РБОО Центр социально-психологической и информационной поддержки «Семья и психическое здоровье», 2007.

76. *Бурбаева Г.Ш., Бокша И.С., Судаков С.А., Мясоедов С.Н. и соавт.* Комплексная нейрохимическая оценка мозговых белков в норме и при шизофрении // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2008. — Т. 102. — № 2. — С. 44–50.

77. *Финн В.К.* Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. — Сер. Информатика, Т. 15. — М.: ВИНИТИ, 1991. — С. 54–101.

78. *Финн В.К., Блинова В.Г., Панкратова Е.С., Фабрикантова Е.Ф.* Интеллектуальная система для анализа медицинских данных: Части 1, 2, 3 // Врач и информационные технологии. — 2006. — № 5. — С. 62–70; 2006. — № 6. — С. 50–60; 2007. — № 1. — С. 51–57.

79. *Гусев Е.И., Завалишин И.А., Бойко А.Н.* Рассеянный склероз и другие демиелинизирующие заболевания. — М.: Миклош, 2004.

80. *Алексеева Т.Г., Ениколопова Е.В., Садальская Е.В. и соавт.* Комплексный подход к оценке когнитивной и эмоционально-личностной сфер у больных рассеянным склерозом // Рассеянный склероз: Прил. к журн. — 2002. — Спец. выпуск. — С. 20–25.

81. *Lazarus R., Folkman S.* Stress, appraisal, and coping. N.-Y: Springer Publishing Co, 1984.

82. *Poser C.N., Paty D.W., Scheinberg L. et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols // Ann. Neurol. — 1983. — 13(3). — P. 227–231.

83. *Реброва О.Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. — М.: МедиаСфера, 2002. — 312 с.

84. *Burbaeva G.Sh., Boksha I.S., Tereshkina E.B. et al.* Systemic neurochemical alterations in schizophrenic brain: glutamate metabolism in focus // Neurochem. Res. — 2007. — V. 32. — № 9. — P. 1434–1444.

85. *Burbaeva G.Sh., Boksha I.S., Tereshkina E.B. et al.* Glutamate metabolizing enzymes in prefrontal cortex of Alzheimer's disease patients // Neurochem. Res. — 2005. — V. 30. — № 11. — P. 1443–1451.

86. *Ермолаев О.Ю.* Математическая статистика для психологов. — М.: Московский психолого-социальный институт, Флинта, 2004. — 102 с.

87. Индикаторы устойчивого развития России (эколого-экономические аспекты) / Под ред. С.Н. Бобылева, П.А. Макеенко — М.: ЦПРП, 2001. — 220 с.

88. Индекс человеческого развития: Проблемы и перспективы: Сб. статей / Под ред. А.А. Саградова. — М.: Макс-Пресс, 2002. — 96 с.

89. Устойчивое развитие: ресурсы России / Под ред. Н.П. Лаверова. — М.: Издательский центр РХТУ им. Д.И. Менделеева, 2004. — 212 с.

90. Человеческое развитие: новое измерение социально-экономического прогресса: Учебное пособие / Под ред. В.П. Колесова (экономич. ф-т МГУ) и Т. Маккинли (ПРООН, Нью-Йорк). — М.: Права человека. — 2000. — 464 с.

91. *Закс Л.* Статистическое оценивание. — М.: Статистика, 1976.

92. *Зверева Н.В.* Опыт применения кластерного анализа в детской патопсихологии // Сборник тезисов конференции, посвященной 40-летию ф-та психологии МГУ им. М.В. Ломоносова. — М., 2006.

93. *Мелешко Т.К., Алейникова С.М., Захарова Н.В.* Особенности формирования познавательной деятельности у детей, больных шизофренией. Проблемы шизофрении детского и подросткового возраста / Под ред. М.Ш. Вроно. — М., 1986.

94. *Юрьева О. П.* О типах дизонтогенеза у детей, больных шизофренией // Журн. неврол. и психиатр. им. С.С. Корсакова. — 1970. — № 8. — С. 1229–1235.

95. *Иваницкий А.М., Стрелец В.Б., Корсаков И.А.* Информационные процессы мозга и психическая деятельность. — М.: Наука, 1984. — 135 с.

96. *Boutros N., Nasrallah H., Leighty R. et al.* Auditory evoked potentials, clinical vs research applications // Psychiatry Res. — 1997. — V. 69. – P. 183–195.

97. *Gold J., Weinberger D.* Cognitive deficits and the neurobioilogy of schizophrenia // Current opinion in Neurobiology. – 1995. – V. 5. – P. 225–230.

98. *Лебедева И.С.* Нейрофизиологические механизмы обработки слуховой информации в условиях избирательного внимания в норме и их аномалии при шизофрении: Автореф. дис. ... д-ра биол. наук. — 2007. — 42 с.

99. *Лебедева И.С., Каледа В.Г., Абрамова Л.И., Бархатова А.Н., Омельченко М.А.* Нейрофизиологические аномалии в парадигме Р300 как возможные эндофенотипы шизофрении // Журн. неврол. и психиатрии им. С.С. Корсакова. — 2008. — Т. 108. — № 1. — С. 61–70.

100. *Mathalon D.H., Ford J.M., Pfefferbaum A.* Trait and state aspects of P300 amplitude reduction in schizophrenia: a retrospective longitudinal study // Biological Psychiatry. — 2000. — № 47. — P. 434–449.