первого класса $c(1)$ в объединенной контрольной выборке такие, что при $c(1) < d(1)$ рассматриваемый алгоритм хуже второго тривиального алгоритма, а при $c(1) > d(2)$ он хуже первого тривиального алгоритма.

Поэтому мы полагаем, что использовать в качестве показателя качества алгоритма диагностики долю правильной диагностики нецелесообразно.

Предлагаем применять *метод пересчета на модель линейного дискриминантного анализа*[1], согласно которому показателем качества алгоритма диагностики является т.н. «прогностическая сила», а статистической оценкой «прогностической силы» h является «эмпирическая прогностическая сила» $h^* = \Phi(d^*/2)$, $d^* = G(a) + G(b)$, где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1, а $G(y)$ — обратная ей функция.

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется классический линейный дискриминантный анализ Р. Фишера, то величина d^* представляет собой состоятельную статистическую оценку расстояния Махаланобиса между двумя рассматриваемыми совокупностями, причем независимо от порогового значения, определяющего конкретное решающее правило. В общем случае показатель h^* вводится как эвристический. Распределение статистики h^* является асимптотически нормальным, что позволяет строить доверительные интервалы для прогностической силы h[2].

Как проверить обоснованность пересчета на модель линейного дискриминантного анализа? Допустим, что классификация состоит в вычислении некоторого прогностического индекса y и сравнении его с заданным порогом c. Объект относят к первому классу, если $y \le c$, ко второму, если $y > c$. Прогностический индекс — это обычно линейная функция от характеристик рассматриваемых объектов. Возьмем два значения порога c_1 и c_2. Если пересчет на модель линейного дискриминантного анализа обоснован, то, как можно показать, «прогностические силы» для обоих правил совпадают: $h(c_1) = h(c_2)$. Выполнение этого равенства можно проверить как статистическую гипотезу. Расчетные алгоритмы предложены нами в цитированной работе 1987 г. и включены в наши учебники, цитированные на протяжении настоящей статьи.

Организационные меры, которые целесообразно использовать для исправления явно ненормальной ситуации в той научной области, которой посвящена настоящая конференция, достаточно очевидны. Надо проанализировать сделанное в самой этой области и в прилегающей к ней, в которых развиваются и применяются статистические методы. Не только необходима научная и учебная специальность «Математические методы в социологии», но и обеспечивающая ее инфраструктура — сеть научных учреждений и подразделений, журналов, конференций, диссертационных советов и т.д. Иначе можно умереть от жажды, сидя на каменном островке посередине реки и громко крича: «Ау, мы отстали! Где вода?».

Выбор типичных объектов в классификационных задачах

Буховец Алексей Георгиевич,
Бирючинская Татьяна Яковлевна,
Воронежский государственный аграрный университет им. К.Д. Глинки
Лаврова Екатерина Олеговна, *НИУ ВШЭ*

1. Разделение понятий типологии и классификации позволило акцентировать внимание исследователей на связи содержательной постановки задачи с общеметодологическими понятиями и методами получения типологии как таковой.

[1] Орлов А.И. О сравнении алгоритмов классификации по результатам обработки реальных данных // Доклады Московского Общества испытателей природы 1985 г. Общая биология: Новые данные исследований структуры и функций биологических систем. М.: Наука, 1987. С. 79–82.

[2] См.: Орлов А.И. Организационно-экономическое моделирование: учебник в 3 ч. Часть 1: Нечисловая статистика. М.: Изд-во МГТУ им. Н.Э. Баумана. 2009 и др. учебники.

Классификацию мы будем рассматривать как знание об объективно существующей внешней системе, функционирующей в рамках целостности и состоящей из подобных друг другу представителей целостностной общности. Проблема системности в данном случае — это проблема соотнесения наблюдаемого подобия с общей сущностью. В рамках такого подхода выделяемые классы соответствуют некоторым типам. Известны два принципиально различных подхода к пониманию и описанию типа: (1) тип как *среднее* (предельно обобщенное) и (2) тип как *крайнее* (предельно своеобразное)[1]. В первом случае типичным является объект со свойствами, близкими по своей выраженности к среднему значению класса выборки, во втором — с максимально выраженными свойствами. Эти различия, на наш взгляд, в большей степени связаны с представлением исследователя о рассматриваемой совокупности. Если класс рассматривается отдельно и изолировано от других выделенных классов, то в качестве типичного объекта обычно берется объект, значения показателей которого наиболее близки к среднему значению, т.е. тип интерпретируется как *среднее* (предельно обобщенное). А если классы и анализируются в рамках системного подхода, т.е. как некоторые паттерны целостной совокупности, то в качестве типичного объекта предпочтительнее взять объект, значения признаков которого *максимально* выражают свойства выделенного класса. Таким образом, выбор типичного объекта в значительной степени определяется точкой зрения наблюдателя.

Рассматривая классификационную задачу как задачу исследования структуры многомерных данных, направленную на выделение пространственно неоднородных стационарных распределений объектов[2], отметим, что сам результат, без глубокого понимания причин, его породивших, отражает не более чем образ мышления наблюдателя, его стереотипы или начальные идеологические установки, а не объективное положение дел. Таким образом, выявление и понимание отраженных в понятии *тип* противоречий позволяет развивать это понятие.

Мы разделяем полностью представленные выше взгляды на понятие *типа*. Однако использование таких математических процедур анализа многомерных данных, как кластерный анализ (или его различные модификации), позволяет получать типичные объекты только в виде средних значений, т.е. использовать понятие типа как некой усреднённой величины. В арсенале математических средств, используемых в анализе социологических данных, практически не были представлены методы, позволяющие производить конструирование (выделение) типа, соответствующего второму из отмеченных выше понятий. Такие типы выделялись на этапе теоретического анализа (т.е. при использовании абстрактно-теоретических методов). Однако этот подход всегда оставляет открытым вопрос о существовании таких (даже логически допустимых типов) в конкретном эмпирическом материале. Для определения типичных объектов нами предлагается подход, в основе которого лежит представление о фрактальной структуре многомерных данных и метод (алгоритм), позволяющий проводить выделение типичных объектов, отвечающих второму из приведенных выше определений.

2. Понятие фрактальных множеств (фрактальных структур) было введено сравнительно недавно, в 1975 г., американским математиком польского происхождения Б. Мандельбротом. Не вдаваясь в утомительные подробности строгого математического определения, и не пытаясь охватить все возможные случаи, представляющие фрактальные образования, мы лишь отметим, что для нашей работы вполне достаточно рассмотрения фрактальных объектов, представленных совокупностью точек некоторого признакового пространства. Такие пространства обычно встречаются в работах, где используются методы многомерного статистического анализа: факторного, кластерного или регрессионного. В этом случае фрактальное множество обычно представимо в виде совокупности точек, обладающей некоторыми свойствами, основными из которых являются дробная (фрактальная) размерность множества, меньшая, чем размерность признакового пространства, и наличие свойства самоподобия.

Большая часть работ Б. Мандельброта так или иначе связана с изучением фрактальных структур, их свойств, областей их распространения. В частности, приведем одну цитату из его

[1] Ганзен В.А., Фомин А.А. О понятии типа в психологии // Вестник СПбГУ. Сер. 6. 1993. Вып. 1 (№6).

[2] Типология и классификация в социологических исследованиях / Отв. ред. В.Г. Андреенков, Ю.Н. Толстова. М.: Наука, 1982. С. 156.

широко известной книги: «Я считаю совершенно необходимым — параллельно с продолжением попыток *объяснить* кластеризацию — найти способ *описать* ее и смоделировать реальность чисто геометрическими средствами (*курсив***. — Авторы**)[1]». Фактически здесь идет обсуждение одной важной проблемы современной науки — необходимости не только описывать результаты классификации (кластеризацию), но и моделировать эти структуры. Современный подход к исследованию каких-либо структур — физических, биологических или социальных — нацелен на описание процессов, порождающих эти структуры. Об этом направлении в современном системном мышлении хорошо сказано в известной книге Ф. Капра «Дао физики».

Для моделирования структур многомерных данных нами была предложена процедура, основанная на рандомизированной системе итерирующих функций. Оставляя в стороне подробное описание этой процедуры, мы отсылаем за подробностями читателя к работам, в которых приводится описание самой процедуры и исследование свойств получаемых в результате ее выполнения множеств[2]. Эту процедуру относительно легко можно реализовать в системе Mathcad, что и было сделано авторами.

Для выполнения этой процедуры необходимо указать количество генерируемых точек, точки, которые участвуют в порождении фрактального множества, *точки протофрактала*, вместе с заданным на них распределением вероятностей, а также значение входного параметра, характеризующего степень близости точек одного класса к соответствующей точке протофрактала. Было показано, что построение фрактального множества (предфрактал — это еще то предельное множество, которое называется фракталом, и поэтому предпочтительнее говорить о предфрактале, а не о фрактале) можно выполнить и другим способом, при котором учитывается указанное выше значение входного параметра, число классов в классификации и численности классов[3]. Возможность выполнения процедуры предфрактального множества построения двумя различными способами — с и без использования множества протофрактала — позволяет изменить (обратить) саму схему выполнения алгоритма. Для этого достаточно предположить, что имеющиеся в распоряжении исследователя данные представляют собой некоторое фрактальное образование — предфрактал. Проведенное исследование структуры многомерных данных с помощью алгоритмов кластерного анализа позволяет получить классификационное разбиение, т.е. выделить классы, указать их состав, численность и оценить степень их рассеяния в признаковом пространстве. Полученные результаты выполнения классификационных алгоритмов в дальнейшем будут использоваться для оценки входных параметров указанной выше процедуры построения протофрактала. Результат выполнения процедуры — в данном случае в обратном порядке, т.е. не для построения фрактального множества, а для определения протофрактала — позволяет получить значения признаков объектов, которые в процедуре при выполнении прямого хода играют роль точек протофрактала. Другими словами, если бы выделенные объекты рассматривались как объекты, порождающие все множество посредством процедуры построения фрактала, то эти объекты следует признать как типичные во втором указанном выше определении типичного объекта. В ходе исследования свойств предложенной процедуры было выявлено, что эти объекты могут быть рассмотрены как типичные объекты выделенных классов. При этом понятие типичного объекта, выделенного этой процедурой, будет соответствовать именно второму из рассмотренных нами типов[4].

3. В качестве примера, иллюстрирующего изложенные выше теоретические положения, рассмотрим задачу о выделении типичных объектов при классификации регионов России. Мы будем использовать именно этот термин, т.к. содержательный анализ, выполненный нами в ходе построения типологии, в этой работе не приводится ввиду ограниченности объема.

---

[1] Мандельброт Б. Фрактальная геометрия природы. М: Институт компьютерных исследований, 2002. С. 127.

[2] Буховец А.Г. Системная интерпретация результатов классификации // Социология: методология, методы, математические модели. 2006. № 22. С. 114–144; Буховец А.Г. Математические модели и методы типологического анализа // Социологические методы в современной исследовательской практике: Материалы Всероссийской научной конференции. М., 2007. С. 16–31.

[3] Буховец А.Г. Математические модели и методы…

[4] Буховец А.Г. Системная интерпретация…; Буховец А.Г. Математические модели….

Как известно, задача выделения типичных объектов играет большую роль в построении классификационной (районированной) выборки. Кроме этого, эта задача может иметь самостоятельное значение, поскольку в классификационных задачах типичный объект не всегда совпадает с некоторым усреднением по классу, а скорее определяет главную тенденцию эволюционного развития выделенного класса.

Исходя из анализа основных теоретических работ, посвященных типологии российских регионов были выбраны следующие блоки показателей:

— уровень жизни в регионе;

— инвестиционная активность;

— экономический потенциал.

Были получены статистические данные по 83 регионам, каждый регион характеризовался следующими 8 показателями:

- доля населения с доходами ниже прожиточного минимума (индикатор бедности);
- отношение среднедушевых доходов к прожиточному минимуму;
- отношение среднедушевых расходов к прожиточному минимуму;
- доля инвестиций в валовом региональном продукте;
- темпы роста инвестиций в 2009 г. по отношению к среднероссийскому уровню на соответствующий период 2008 г.;
- отношение иностранных инвестиций к валовому региональному продукту;
- уровень безработицы в регионе (доля безработных от экономически активного населения региона);
- темпы роста валового регионального продукта по отношению к ВВП.

Все показатели, кроме прожиточного минимума (2010 г.) и доли инвестиций в основной капитал (2007 г.), были рассчитаны за 2008 г. (ввиду того, что большинство из них имеют привязку к ВРП, а последнее значение ВРП есть только за 2008 г.).

Все статистические данные были предварительно стандартизированы.

При построения классификации регионов использовались методы кластерного анализа: алгоритмы иерархической классификации и метод k-средних. Как показали исследования, структура данных была такова, что ее хорошо можно было представить по результатам метода k-средних.

Исследование структуры данных дополнялось изучением проекций данных в пространстве первых двух главных компонент, которые представляли 67,4 % обобщенной дисперсии.

Итоговое классификационное разбиение на 8 кластеров приведено в Приложении 1. Средние значения выделенных кластеров приведены в таблице 1.

Приведем очень краткие характеристики выделенных кластеров.

1-й кластер условно можно назвать **«самым благополучным»**, если исходить из высокого уровня жизни и экономического потенциала Ближе всего к кластерному центру находится **Свердловская область**.

2-й кластер составили регионы, которые можно определить как «**потенциально неблагополучные**». Эти регионы быстрыми темпами приближаются к «неблагополучным». Самый яркий представитель — **Амурская** область.

3-й кластер образуют два региона, которые условно можно назвать «самыми **неблагополучными»**. Здесь самая высокая доля населения с доходами ниже прожиточного минимума, отношение среднедушевых доходов и расходов также самое низкое, причем отношение расходов к прожиточному минимуму меньше 1.

4-й кластер состоит из одного региона, который выделен в отдельный кластер из-за низких темпов роста инвестиций и высокой доли иностранных инвестиций.

5-й кластер — «**неблагополучные**» — один из самых многочисленных. Регионы этого кластера имеют довольно высокую долю населения с доходами ниже прожиточного минимума, низкие среднедушевые доходы и расходы. Темпы роста инвестиций в основной бюджет по сравнению со среднероссийским уровнем невысоки. Регионы данного кластера похожи на регионы второго кластера, однако во втором кластере наблюдается другая структура распределения данных, и эта структура, видимо, обусловлена главным различием этих кластеров — высо-

кими темпами роста во втором кластере. Самые яркие представители «**неблагополучных**» регионов — **Забайкальский** край, **Чувашская** республика и **Ульяновская** область.

6-й кластер составили регионы **«потенциально благополучные»**. Самый типичный пример — **Калужская** область.

7-й кластер — самый многочисленный — составляют регионы, занимающие положение между «неблагополучными» и «потенциально благополучными». Можно сказать, что это **«средние»** по уровню развития регионы без ярко выраженных тенденций перехода к неблагополучным или благополучным регионам. Среднедушевые доходы и расходы выше среднего уровня по России, но ниже, же у «благополучных» и «потенциально благополучных» регионов. Типичными представителями кластера служат **Смоленская о**бласть и **Пермский** край.

В отдельный 8-й кластер выделен г**. Москва**, где наблюдается самый высокий уровень жизни и экономического потенциала.

Для определения типичных объектов в рамках фрактального подхода была построена матрица А размером 83х8 при значении параметра процедуры, равного 17,8. Значение параметра выбиралось экспертным путем.

С помощью матрицы А были получены точки признакового пространства представляющие протофрактал. Результаты построения приводятся в Приложении 2. Каждую строку этой матрицы Z можно соотносить с ближайшим объектом классификации.

Точки выделенного протофрактала отражают структуру распределения объектов — регионов России — в пространстве выбранных для классификации признаков. Результаты вычислений для сравнения их со средними значениями приводится в таблице 2.

Можно показать, что в рамках фрактального подхода складывается ситуация, когда значения точек протофрактала, соответствующие кластерам, где средние значения высокие, будет еще выше, а в тех кластерах, где они низкие — еще ниже. Так, к примеру, при разбиении населения по уровню дохода типичным представителем состоятельных людей будет самый богатый, а представителем бедных — самый бедный регион, и лишь среди людей со средним доходом самым ярким представителем будет человек со средним доходом[1]. Однако, такая упрощенная ситуация при анализе классификационного разбиения складывается не часто. Обычно кластеров бывает больше трех, и главной задачей аналитика при содержательной интерпретации является сравнение кластеров между собой, т.е. определение отношения кластеров друг к другу, их взаимосвязи. В нашем случае использование фрактального подхода позволяет оценить различия в структуре кластеров, обнаружить скрытые закономерности в общей структуре и оценить, насколько «потенциальные» регионы отличаются от более однородных.

Ближайшие к точкам протофрактала объекты кластеров в некоторых случаях отличаются от объектов, ближайших к кластерным центрам (см. Таблица 2). Так, самым ярким представителем «самых благополучных» регионов с точки зрения фрактального подхода является г. Санкт-Петербург (была Свердловская область), «потенциально благополучных» — Приморский край (была Амурская область), «неблагополучных» — республика Хакасия и Чувашская республика (был Забайкальский край и Чувашская республика). В кластерах «потенциально благополучных» и «средних» типичные представители не изменились (Калужская и Смоленская области соответственно).

Как видно, различия в положении средних выборочных значений кластеров и вычисленных значений признаков точек протофракалов связано с тем положением, которое занимает кластер во всей исследуемой системе объектов.

В заключении еще раз отметим, что рассмотрение (исследование) эмпирических данных с точки зрения фрактального подхода позволяет практически реализовать метод выделения типичных объектов, наиболее ярко и полно выражающих свойства классов.

Результаты классификационных построений

Приведены: (численность / процентный состав кластеров), список входящих в кластер регионов.

[1] Буховец А.Г. Математические модели….

Жирным шрифтом выделены регионы, типичные в смысле близости к средним значениям показателей кластера; подчеркиванием выделены регионы, типичность которых была установлена в рамках фрактального подхода.

Кластер 1 (9 / 11) Московская область, г. Санкт-Петербург, Татарстан, Самарская, **Свердловская**, Тюменская области, Ханты-Мансийский и Ямало-Ненецкий автономные округа, Челябинская область.

Кластер 2 (3 / 4) **Амурская область**, Приморский край, Якутия.

Кластер 3 (2 / 2) **Ингушетия** и **Чеченская Республика**.

Кластер 4 (1 / 1) ***Ненецкий автономный округ***.

Кластер 5 (27 / 33) Владимирская, Воронежская, Ивановская, Костромская, Рязанская области, Карелия, Калмыкия, Кабардино-Балкария, Карачаево-Черкессия, Марий Эл, Мордовия, **Чувашия**, Кировская, Саратовская, **Ульяновская** области, республика Алтай, Бурятия, Тыва, Хакасия, Алтайский край, **Забайкальский край**, Иркутская и Томская области, Камчатский край, Хабаровский край, Магаданская область, Еврейская автономная область.

Кластер 6 (7 / 8) Чукотский автономный округ, Сахалинская, Вологодская, Архангельская, Липецкая, **Калужская**, Белгородская области.

Кластер 7 (33 / 40) Брянская, Курская, Орловская, **Смоленская**, Тамбовская, Тверская, Тульская и Ярославская области, Республика Коми, Калининградская, Ленинградская, Мурманская, Новгородская и Псковская области, Адыгея, Краснодарский край, Астраханская, Волгоградская и Ростовская области, Дагестан, Северная Осетия-Алания, Ставропольский край, Башкортостан, Удмуртия, **Пермский край**, Нижегородская, Оренбургская, Пензенская и Курганская области, Красноярский край, Кемеровская, Новосибирская и Омская области.

Кластер 8 (1 / 1) **г. Москва**.

Таблица 1

Средние значения показателей кластеров

№ кластера	Численность кластеров	Доля населения с доходами ниже прожиточного минимума	Отношение среднедушевого дохода к прожиточному минимуму	Отношение среднедушевых расходов к прожиточному минимуму.	Темпы роста инвестиций в % к 2008 г. по сравнению со среднерос. уровнем 2009 г.
1	9	10,59	3,25	2,55	96,27
2	3	20,67	1,71	1,31	196,86
3	2	36,10	1,10	0,52	169,99
4	1	10,20	1,74	1,23	43,44
5	27	21,19	1,69	1,36	100,56
6	7	12,49	2,29	1,83	96,80
7	33	15,08	2,23	1,88	109,75
8	1	10,00	4,17	3,84	92,84

Продолжение таблицы 1

№ кластера	Численность кластеров	Доля инвестиций в основной капитал в ВРП_2007	Отношение иностранных инвестиций к ВРП	Отношение темпов роста ВРП и ВВП	Уровень безработицы
1	9	27,80	0,07	0,03	6,46
2	3	35,47	0,05	0,01	9,07
3	2	70,05	0,00	0,00	43,95
4	1	92,70	0,37	0,00	9,70
5	27	27,07	0,02	0,00	10,81
6	7	34,69	0,27	0,01	6,59
7	33	26,88	0,03	0,01	9,17
8	1	11,50	0,12	0,20	2,70

$$Z = \begin{pmatrix} 10.276 & 3.313 & 2.59 & 94.907 & 27.899 & 0.065 & 0.035 & 6.253 \\ 20.638 & 1.688 & 1.289 & 200.858 & 34.995 & 0.046 & 0.006 & 8.455 \\ 37.262 & 0.442 & 0.164 & 173.095 & 71.949 & -0.003 & -0.001 & 45.915 \\ 9.955 & 1.709 & 1.191 & 39.852 & 96.355 & 0.389 & -0 & 9.738 \\ 21.437 & 1.667 & 1.335 & 100.208 & 26.923 & 0.023 & 0.002 & 10.876 \\ 12.33 & 2.296 & 1.827 & 96.794 & 35.277 & 0.279 & 0.007 & 6.367 \\ 15.02 & 2.23 & 1.883 & 109.989 & 26.779 & 0.029 & 0.007 & 9.177 \\ 9.695 & 4.28 & 3.953 & 91.895 & 10.636 & 0.125 & 0.211 & 2.321 \end{pmatrix}$$

Рис. Матрица протофрактала, полученная после выполнения процедуры построения фрактального множества.

Таблица 2

Сравнение значений признаков средних значений (левая колонка) и типичных объектов, полученных в рамках фрактальной теории (правая колонка)

№ кластера	Доля населения с доходами ниже прожиточного минимума		Отношение среднедушевого дохода к прожиточному минимуму		Отношение среднедушевых расходов к прожиточному минимуму		Темпы роста инвестиций в % к 2008 г. по сравнению со среднероссийск. уровнем 2009 г.		Доля инвестиций в основной капитал в ВРП_2007	
1	10,59	10,29	3,25	3,30	2,55	2,59	96,27	95,03	27,80	27,81
2	20,67	20,87	1,71	1,69	1,31	1,29	196,86	201,2	35,47	35,98
3	36,10	36,64	1,10	0,55	0,52	0,21	169,99	174,1	70,05	70,29
4	10,20	9,62	1,74	1,75	1,23	1,23	43,44	35,51	92,70	95,5
5	21,19	21,35	1,69	1,67	1,36	1,34	100,56	100,3	27,07	27,04
6	12,49	12,09	2,29	2,31	1,83	1,84	96,80	96,16	34,69	34,78
7	15,08	15,02	2,23	2,23	1,88	1,88	109,75	109,6	26,88	26,8
8	10,00	9,75	4,17	4,27	3,84	3,94	92,84	91,99	11,50	10,71

Продолжение таблицы 2

№ кластера	Отношение иностранных инвестиций к ВРП		Отношение темпов роста ВРП и ВВП		Уровень безработицы	
1	0,065	0,067	0,033	0,034	6,46	6,17
2	0,047	0,049	0,006	0,006	9,07	8,97
3	0,00	0,001	0,001	0,001	43,95	44,91
4	0,371	0,386	0,002	0,003	9,70	9,65
5	0,025	0,023	0,003	0,002	10,81	10,87
6	0,266	0,279	0,006	0,007	6,59	6,19
7	0,029	0,029	0,007	0,007	9,17	9,17
8	0,121	0,125	0,204	0,21	2,70	2,37

Получение представительных данных по группам на основе представительного обследования входящих в них индивидов

Вейхер Андрей Алексеевич,
СПб. филиал НИУ ВШЭ

Проведение представительного обследования объектов, в которые входит несколько людей (домохозяйств-семей, малых предприятий и т.п.) в большинстве случаев оказывается значительно более трудным, чем представительное обследование общей совокупности людей, входящих во все эти группы.

Как известно, создание выборочных совокупностей адресов мест проживания, по которым проводятся опросы населения, считающиеся наиболее представительными, не являются опросами домохозяйств. При таком опросе населения многоступенчатая выборка с различными спо-