ВОПРОСЫ СРАВНЕНИЯ РАЗНОТИПНЫХ ВЕКТОРОВ В ЗАДАЧАХ УПРАВЛЕНИЯ СОЦИАЛЬНЫМИ СИСТЕМАМИ

А.П. СЕРЫХ
Томский государственный университет
Sgd_46@mail.ru

Обсуждаются возможности и преимущества использования информационных мер различия и сходства для упорядочения объектов с разнотипным (разношкальным) описанием при непараметрическом уровне неопределенности.

1. ВВЕДЕНИЕ

«Всё познается в сравнении» – гласит народная мудрость. Другая, не менее мудрая максима, которую часто можно услышать, звучит так: «Наука начинается там, где начинаются измерения». Действительно, чтобы корректно, «научно» что-то сравнивать, необходимо иметь измерения, желательно полученные в «сильных» шкалах (что нечасто удается). Вопрос о том, кто быстрее пробежал стометровку, обычно не вызывает больших разногласий. А вот определить, кто лучший в художественной гимнастике или в фигурном катании, – предмет многочисленных дискуссий и закулисных игр. Причина сложностей в последней ситуации заключается, во-первых, в использовании более слабых (по сравнению со шкалой измерения времени для беговых видов спорта) порядковых шкал, и, во-вторых, процедура суммирования баллов (по двум качественно разным характеристикам) для получения итоговой оценки с точки зрения теории измерений далеко небезупречна. Надо ли говорить, насколько подобные проблемы усугубляются, когда объекты описываются многими разнотипными (разношкальными) переменными. Именно с такой ситуацией мы сталкиваемся при решении задач описания и управления в социальных системах, где процедура сравнения, упорядочения объектов играет клю-

чевую роль. Настоящая работа посвящена обсуждению упомянутых проблем на примере сравнения качества и уровня жизни. Решение этой задачи уже содержит многие особенности, с которыми могут столкнуться специалисты при управлении социальными системами.

Проблеме измерения, анализа качества жизни, её уровня, определению содержания этих понятий посвящены статьи в специализированном журнале «Уровень жизни населения регионов России» (см. [1, 2]). Разнообразие точек зрения по этим вопросам дает основание утверждать, что словосочетания «качество жизни» и «уровень жизни» относятся к понятиям с «мерцающими» смыслами. Можно назвать две основные причины, приводящие к нечёткости, расплывчатости, «мерцанию» смыслов этих понятий.

Первая из них – состав учитываемых показателей, свойств, характеристик, «причастных» к раскрытию смысла понятий качества или уровня жизни. Здесь желание перечислить как можно более полный перечень, который бы согласовался с интуитивными или научными представлениями об этих понятиях, сталкивается с невозможностью практического измерения части из этих показателей либо отсутствием статистических данных о показателях (в принципе измеряемых) для ряда стран или регионов. Компромиссный, минимальный набор характеристик («лучше уж что-то измерять и сравнивать, чем ничего») используется при расчёте ИРЧП (индекс развития человеческого потенциала, рассчитываемый на основе трех равнозначимых индексов: средней продолжительности жизни, уровня образования и ВВП на душу населения).

Вторая причина неопределённости связана с синтезом процедуры сравнения. Традиции многих наук стимулируют поиски единого интегрального показателя, удобного индекса, который бы позволил ранжировать, упорядочить изучаемые объекты. Однако ранжирование многомерных векторов представляет собой довольно сложную задачу из-за неоднозначности выбора упорядочивающей функции. В частности, при синтезе ИРЧП предпочтение отдано линейной форме (почему?), далее многие авторы отмечают произвол в выборе весовых коэффициентов. Кроме того, почему используются только средние значения показателей? Несмотря на важность этой числовой характеристики, средняя зарплата, например, неадекватно описывает положение дел о социальной справедливости в некотором регионе, а следовательно, и о качестве жизни в нём (вспомним шутку о средней температуре по больнице).

Судя по всему, дискуссии о составе показателей для сравнения качественно разнородных объектов ещё не скоро завершатся, поскольку, по большому счёту, они соприкасаются с «вечными вопросами», ответы на которые имеют не менее «мерцающие» смыслы. Заметим только, что, по

нашему мнению, набор показателей должен описывать и измерять уровень жизни, который является необходимым условием достижения высокого качества жизни. Что же касается второй части обсуждаемой проблемы – синтеза процедуры сравнения, её смысла, – то часть противоречий и трудностей можно избежать, используя арсенал методов многомерного статистического анализа, теории статистических решений и распознавания образов. Обсуждению этой части проблемы и посвящена данная работа.

2. МЕРЫ СХОДСТВА И РАЗЛИЧИЯ НА ОСНОВЕ СТОХАСТИЧЕСКОЙ МОДЕЛИ РЕАЛЬНОСТИ

Прежде всего, необходимо отметить, что рассматриваемая задача имеет сугубо статистический характер, поскольку речь идёт об уровне и качестве жизни большого количества людей, которое можно интерпретировать как некую генеральную совокупность. Самое полное статистическое описание таких совокупностей содержится в совместном распределении вероятностей выделенных показателей, признаков, характеристик. (Вопрос о том, насколько «полон», обоснован сам набор признаков, как уже отмечалось, мы сейчас не обсуждаем).

Пусть $x=(x_1, x_2, ..., x_n)$ – вектор показателей, $x \in X^n$, где X^n – пространство, которое в статистике называют пространством наблюдений. Пространство X^n в общем случае может быть разнотипным в том смысле, что компоненты вектора x измеряются в различных типах шкал (номинальных, ранговых, числовых и т.д.), т.е. иметь как качественный, так и количественный характер, а их декартово произведение и образует разнотипное пространство наблюдений.

Наряду с пространством наблюдений определим ещё вектор условий $y=(y_1, y_2, ..., y_m)$, в рамках которых наблюдается вектор x; $y \in Y^m$, где Y^m – пространство условий, которое также может быть разнотипным. В частности, при расчете ИРЧП: y – скаляр, $Y^1=\{1, 2,..., k\}$ – суть множество обозначений стран или регионов, подлежащих сравнению. В этом случае y фиксируется в номинальной шкале.

Таким образом, статистическим описанием (моделью) учитываемых показателей x, зафиксированных в условиях y, является распределение $p(x / y)$. Следует заметить, что оба вектора x и y и можно сокращать либо расширять, но если они фиксированы, то это означает, что всеми неучтёнными признаками или условиями мы пренебрегаем. Конечно, знание совместного распределения $p(x,y)$ было бы предпочтительней, чем условного, так как оно содержит в себе информацию обо всех связях между признаками и всеми условиями наблюдения. Принятие условных распределений в качестве основной модели – уступка реальному положению

вещей. Дело в том, что отдельные компоненты вектора могут не иметь статистической природы (как, например, наименование страны или региона), либо, в других случаях, распределение $p(y)$ нам неизвестно и, следовательно, совместное распределение $p(x, y) = p(x / y)\, p(y)$ не может быть вычислено.

Рассмотрим теперь процедуру сравнения объектов по показателям x при фиксированных условиях y. Для этого необходимо ввести некую меру сходства или различия. Меры сравнения статистических объектов задаются в виде функционалов от распределений. Они нашли широкое распространение в теории информации, статистических решений, распознавании образов, многомерном статистическом анализе и, на наш взгляд, должны найти применение в обсуждаемой проблеме – применительно к социальным системам.

Рассмотрим для простоты вариант сравнения двух объектов по характеристикам x. В наших обозначениях это означает: $Y^1=\{1, 2\}$, статистическое описание объектов имеет вид $p(x / y = 1) = p_1(x)$, $p(x / y = 2) = p_2(x)$. В качестве меры различия можно использовать, например, вариационное расстояние Колмогорова

$$K(1,2) = \frac{1}{2} \int_X |p_1(x) - p_2(x)|dx. \qquad (1)$$

Эта мера принимает значения из отрезка [0,1], причём равна нулю, когда распределения полностью совпадают, т.е. $p_1(x) = p_2(x)$, и равна единице, когда носитель первого распределения X_1 и носитель второго – X_2 удовлетворяют условиям:

$$X_1 \bigcup X_2 = X, \quad X_1 \bigcap X_2 = 0.$$

Мера (1) имеет однозначную связь с таким понятием, как средняя вероятность ошибки распознавания $p_{ош}$, и с её помощью удобно интерпретировать различие между статистическими объектами по набору признаков x. Связь между этими характеристиками имеет вид $K(1, 2) = 1–2\, p_{ош}$. В контексте обсуждаемой проблемы интерпретация различия между уровнем или качеством жизни двух регионов выглядит следующим образом: **если взять представительные (репрезентативные) выборки людей из первого и второго регионов, то как часто оптимальный алгоритм будет ошибаться в отнесении их к своему или чужому региону по набору признаков x?**

Если ошибок нет ($p_{ош} = 0$), то различие максимально и $K(1, 2) = 1$; если оптимальный алгоритм равновероятно относит представителей выборок к любому из регионов, т.е. ошибка максимальна ($p_{ош} = 1/2$), то раз-

личие регионов по признакам *x* отсутствует и *K(1, 2) = 0.* Оптимальность алгоритма здесь понимается в смысле минимума средней вероятности ошибок.

Для сравнения *m* регионов ($Y^1=\{1,..., m\}$) имеет смысл вычислить матрицу взаимных различий:

$$K(i,j)=\frac{1}{2}\int_X\left|p_i(x)-p_j(x)\right|dx,\quad i,j=1,2,...m. \tag{2}$$

Расстояние Колмогорова (1) учитывает различия между совместными распределениями признаков, однако для достижения тех же целей можно также использовать меры сходства между распределениями. Такой мерой является информационное «расстояние» Бхаттачарья

$$B(1,2)=\int_X p_1^{1/2}(x)p_2^{1/2}(x)dx\,, \tag{3}$$

связанное со средней вероятностью ошибок границами Чернова [3]

$$B^2(1,2)\le 2p_{\text{ош}}\le B(1,2)\,. \tag{4}$$

Мера (3) равна единице, когда распределения совпадают, и нулю, когда распределения имеют непересекающиеся носители X_1 и X_2. Для этих крайних случаев границы (4) являются точными. При использовании меры сходства для сравнения *m* регионов вычисляется матрица сходства:

$$B(i,j)=\int_X p_i^{1/2}(x)p_j^{1/2}(x)dx,\quad i,j=1,2,...,m. \tag{5}$$

Очевидно, что дополнение до единицы меры (3) даст меру различия, а дополнение до единицы меры (1) даст меру сходства. (Нами выбраны информационные расстояния, значения которых принадлежат отрезку [0,1], хотя в принципе в таких задачах можно использовать и другие меры различия, например дивергенцию Кульбака и др.).

Применение информационных мер различия или сходства позволяет обойти неоднозначную и спорную процедуру взвешивания признаков, которая неизбежно возникает при попытке синтезировать индекс сравнения традиционным способом. Далее, интерпретация мер (1) или (3) на языке вероятностей ошибок инвариантна по отношению к составу показателей *x*. Их можно вычислить и в том случае, когда показатели разно-

типны (об этом подробнее мы остановимся в разд. 4). Это устраняет ещё одну неестественную операцию – попытки привести все показатели к стоимостному эквиваленту или к единой измерительной шкале, искажая при этом природу и смысл исходных данных. Наконец, функционалы от совместных распределений содержат **связи между всеми показателями**, что является, возможно, самым существенным в предлагаемом способе сравнения. Если ставить задачу проведения реформ или преобразований в некотором регионе, то без учета этих связей нельзя говорить о комплексности или системности подобных мероприятий. Индексы, основанные на взвешенных средних значениях признаков, этих связей не учитывают.

3. ЗАДАЧИ СРАВНЕНИЯ МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

Рассмотрим теперь некоторые постановки задач, которые могут иметь смысл в рамках обсуждаемой проблемы.

1. **Ранжирование (упорядочивание) регионов**. В этой задаче существенным моментом является выбор эталона, образца, некоторого идеального распределения $p^*(x)$, относительно которого ранжируются все остальные распределения. Если эталон задать, синтезировать затруднительно, то в качестве $p^*(x)$ можно принять одно из имеющихся распределений, которые эксперты сочтут образцовым. Тогда, например, при использовании меры (3) вычисляются величины

$$B_j = \int_X p_j^{1/2}(x)\, p^{*1/2}(x)dx, \quad j = 1,2,...,m, \tag{6}$$

которые затем упорядочиваются. Для эталонного распределения величина $B^* = 1$; для распределений, носители которых не пересекаются с носителем эталонного распределения, эти величины равны нулю.

2. **Кластерный анализ**. Задача упорядочения может потерять смысл, когда сравниваемые объекты имеют серьёзные качественные различия. Тогда целесообразно предварительно разбить регионы на некоторое число качественно однородных классов (кластеров), внутри которых и решать задачу ранжирования. Исходными данными для алгоритмов кластерного анализа являются матрицы сходства или различия (2) или (5). Имеет смысл также сравнивать выделенные кластеры между собой. Для этого по формуле полной вероятности рассчитываются условные распределения признаков для каждого кластера и решается задача ранжирования кластеров аналогично п. 1.

3. **Определение информативных комплексов признаков**. Когда различия установлены, то при многомерном описании объектов представляет интерес выяснить, какие комплексы признаков вносят решающий вклад в эти различия. Для этого функционалы сходства или различия рассчитываются для различных сочетаний показателей с последующим сравнением полученных значений между собой. Анализ информативности позволяет выявить латентные связи между признаками. Кроме того, устанавливается подвектор признаков, который практически не влияет на меру различия и может быть опущен из дальнейшего анализа, тем самым уменьшая размерность описания объектов, что немаловажно для практической реализации алгоритмов обработки данных.

Как показывает опыт решения задач распознавания образов, анализ информативности создаёт семантическое поле, в рамках которого можно уточнить и смысловое ядро самого понятия «качество жизни». Ранее отмечалось, что разные исследователи используют различные наборы характеристик для описания уровня и качества жизни. Если провести анализ на информативность этих наборов, то теперь сравнение их между собой даст «информацию к размышлению» о создании некоторого компромиссного набора признаков, нивелирующего субъективизм отдельных авторов.

4. **Регрессионный анализ.** В предыдущей задаче связи между признаками изучались с точки зрения их влияния на сходство или различия между распределениями. Однако не меньший интерес представляет выяснение зависимости какого-либо признака от других признаков. (Например, как продолжительность жизни связана с доходами и образованностью человека?) Ответы на подобные вопросы можно получить средствами регрессионного анализа, которые хорошо развиты, особенно для случая обработки количественных признаков [4].

5. **Преобразование признаков. Задача реформирования.** Конечной целью исследования качества и уровня жизни является не столько упорядочение регионов в каком-либо отношении (чем, в основном, озабочены политики), а реальное их улучшение. В связи с этим актуальным становится вопрос: как преобразовать показатели, чтобы приблизить их распределение к эталонному распределению? Рассмотрим, как выглядят эти преобразования на простых примерах. Пусть x – скаляр, и его распределение для некоторого региона и образцового имеют, соответственно, вид $p(x)$ и $p^*(x)$. Преобразование признака $z = \varphi(x)$, которое приводит к равенству $p(x) = p^*(x)$, имеет следующий вид:

$$z = F^{*-1}\left[F(x)\right], \tag{7}$$

где $F^{*-1}(t)$ – квантильная функция распределения $p^*(x)$ и $F(x)$ – интегральная функция распределения для $p(x)$. Пусть теперь $x = (x_1, x_2)$. Для двух признаков соответствующие преобразования имеют следующий вид:

$$z_1 = F_1^{*-1}\left[F(x_1)\right], \quad z_2 = F_2^{*-1}\left[F(x_2 / x_1) / x_2\right]. \tag{8}$$

Следует отметить, что уже в случае двух признаков, преобразования не могут осуществляться независимым образом, если исходные признаки зависимы. Незнание связей между признаками не позволит грамотно и, следовательно, комплексно осуществить реформирование. Обобщение преобразований для большего числа переменных проводится аналогично и с учетом всех связей между признаками. Очевидно, что после преобразований типа (7) – (8) функционал Бхаттачарья (6) будет равен единице.

В обсуждаемой задаче реформирования, а также в задаче упорядочивания регионов относительно образцового региона более адекватной может являться так называемая «направленная мера» похожести. Дело в том, что информационные меры (2) – (3) фиксируют сходство или различие между распределениями, связанные со средней вероятностью ошибок, а это комбинация ошибок двух родов. Другой смысловой оттенок появляется, если за меру похожести принять вероятность события следующего рода – перепутать выборку сравниваемого региона с эталонной. Такая мера имеет вид

$$\alpha = \int\limits_{p^*(x) \geq p(x)} p(x)dx = \int\limits_X I\left[p^*(x) - p(x)\right]p(x)dx, \tag{9}$$

где $I[z]$ – индикаторная функция: { $I[z]=1$, если $z \geq 0$; $I(z)=0$, если $z<0$}.

Величина α даёт ответ на вопрос: **какая доля населения региона с распределением признаков $p(x)$ имеет уровень жизни образцового региона?**

4. УЧЕТ РАЗНОТИПНОСТИ ОПИСАНИЯ ОБЪЕКТОВ

Чтобы реализовать какую-либо из перечисленных выше задач, прежде всего необходимо знать многомерные распределения показателей, входящие в функционалы сходства или различия объектов. Надо заметить, что теоретические законы именно многомерных распределений с возможными связями между компонентами можно перечислить по пальцам одной руки. Вместе с тем основная масса теоретических исследований, посвященных изучению так называемого параметриче-

ского уровня неопределенности, связана с такими параметрическими семействами.

Аналоги теоретического описания распределений многомерных разнотипных случайных векторов, по-видимому, вовсе отсутствуют. На языке математической статистики мы находимся на так называемом непараметрическом уровне неопределенности, который предполагает восстановление (оценивание) неизвестных распределений. При оценивании распределений, кроме весьма общих теоретических ограничений (как правило, мало влияющих на конечный результат), основным источником для создания стохастической модели реальности являются репрезентативные, синхронные статистические наблюдения всех характеристик изучаемых объектов.

Технике построения непараметрических оценок плотностей распределений и изучению их свойств посвящена обширная библиография. Сошлемся здесь лишь на работу Е.А. Епанечникова [6], в которой влияние многомерности описания объектов отражено наиболее ярко. Однако в этих работах предполагается, что выборочные значения получены в «сильных» числовых шкалах. В задачах управления социальными системами, где наряду с числовыми данными почти всегда присутствуют качественные характеристики объектов, мы имеем дело с разнотипными векторами. Особенности построения непараметрических оценок распределений разнотипных случайных векторов рассмотрены в работе автора [5].

Основная идея распространения непараметрических методов на этот случай заключается в том, что оценку распределения можно построить, не вводя единой метрики в разнотипном пространстве. Попытки корректно определить расстояние в таком пространстве связаны с противоречиями, которые уже обсуждались во введении, когда речь шла о синтезе интегральных показателей. Более естественным представляется вычислять **объёмы** в разнотипных пространствах, что, собственно, и требуется при построении оценок распределений, основанных на непараметрических фактах. Рассмотрение вопросов непараметрического оценивания информационных мер различия или сходства для разнотипных измерений, а также синтеза алгоритмов обработки данных при решении задач, перечисленных в п. 3, требует отдельной статьи, что входит в намерения автора.

ЛИТЕРАТУРА

1. *Мстиславский П.С.* Вопросы теории и методологии анализа качества жизни // Уровень жизни населения регионов России.2002. № 2. С. 5–17.

2. *Маликов Н.С.* К вопросу о содержании понятия «качество жизни» и его измерению // Уровень жизни населении регионов России. 2002. № 2. С. 17–23.

3. *Фукунага К*. Введение в статистическую теорию распознавания образов. М.: Наука, 1979. 383 с.

4. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д*. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985. 487 с.

5. *Серых А.П*. Оценка распределения случайных векторов разнотипных данных // Научная сессия Том. ун-та (апрель 1992.). Ч. 2. Томск: Изд-во Том. ун-та, 1993.

6. *Епанечников В.А*. Многомерная непараметрическая оценка плотности распределения вероятности. Теория вероятностей и её применение. Т. 1. М., 1969.