Выбор опорного множества при построении устойчивых интегральных индикаторов*

Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.
strijov@ccas.ru
Москва, Вычислительный центр РАН
Берлин, Технический университет

Исследуется задача построения интегрального индикатора множества объектов, устойчивого к выбросам в описаниях объектов. Объекты описаны в линейных шкалах. Для построения интегрального индикатора из множества всех описаний с помощью критерия принадлежности выбирается множество опорных описаний. Интегральный индикатор строится методом «без учителя». Предложенный алгоритм использован для получения интегрального индикатора уровня загрязнений основных продуктов питания в регионах России.

Введение

Построение интегрального индикатора — введение отношения порядка на множестве сравнимых объектов. Выбор алгоритма построения индикатора зависит от тех свойств, которыми обладают объекты. Предполагается, что каждый объект описан вектором, компоненты которого являются результатами измерений соответствующих показателей. Все измерения выполнены в линейных шкалах. Интегральный индикатор — скаляр, поставленный в соответствие объекту. Говоря о наборе объектов, будем называть интегральным индикатором вектор, компоненты которого поставлены в соответствие сравниваемым объектам.

Распространенным алгоритмом построения интегральных индикаторов для объектов, описанных в линейных шкалах, является линейная комбинация значений показателей [1]. Веса при этом вычисляются исходя из некоторого заданного критерия информативности описаний. Принятый в данной работе критерий наибольшей информативности, введенный С. Р. Рао, рассмотрен в первом разделе в связи с методом главных компонент. Однако этот метод вызывает, при наличии выбросов в описаниях объектов, проблему адекватной сравнимости объектов. Эксперты, определяющие множество объектов, предполагают все объекты сравнимыми и ожидают от алгоритма адекватные значения интегральных индикаторов. Однако если некоторые отдельные объекты имеют значения показателей, существенно отличающиеся от значений показателей основного числа объектов, то, в рамках линейной модели, объекты-выбросы имеют большее влияние на веса показателей, чем прочие объекты. При исключении таких объектов можно наблюдать изменение значений индикаторов, существенное не только для линейных, но даже и для ранговых шкал.

Ранее были предложены алгоритмы получения устойчивых интегральных индикаторов с использованием как линейных [2], так и нелинейных моделей [3, 4].

В данной работе исследуется задача построения устойчивых интегральных индикаторов. Решением этой задачи является алгоритм построения индикатора для всего множества объектов, построенный на основе его подмножества, называемого *опорным множеством*. Алгоритм разделяет исходное множество описаний объектов на два подмножества — опорное и множество выбросов. При этом используется критерий вероятности принадлежности описаний объекта одному из двух подмножеств. По опорному множеству, с помощью метода главных компонент, вычисляются веса. Эти веса используются для получения интегральных индикаторов всей выборки.

Алгоритм построения интегральных индикаторов

Задано множество, состоящее из m объектов, которые описаны набором из n показателей. Задана матрица описаний $A \in \mathbb{R}^{m\times n}$. Элемент матрицы a_{ij} — значение j-го показателя i-го объекта. Вектор $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{in})$ — описание i-го объекта.

Интегральный индикатор объекта — это свертка вида

$$q_i = \sum_{j=1}^{n} w_j g_j(a_{ij}), \tag{1}$$

где g_j — функция приведения показателей в единую шкалу:

$$g_j : a_{ij} \mapsto (a_{ij} - \min_i a_{ij})(\max_i a_{ij} - \min_i a_{ij})^{-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{2}$$

Если в формуле (2) знаменатель равен нулю, то это означает, что значения j-го показателя для всех объектов равны. При этом показатель не может быть использован для построения интегрального индикатора и должен быть исключен из дальнейшего рассмотрения.

Без ограничения общности будем считать, что выполнено условие монотонности такое, что

*Работа выполнена при финансовой поддержке РФФИ, проекты № 07-07-00181, 08-01-12022.

из $a_{ij} \geqslant a_{\xi j}$ следует $q_i \geqslant q_\xi$ для $j = 1, \ldots, n$. Выполнение этого условия вместе с выполнением (2) влечет неотрицательность значений $w_1, \ldots, w_n$. Так как на практике выставляется требование инвариантности интегрального индикатора к линейным преобразованиям, введем еще одно условие, накладываемое на веса: $\sum\limits_{j=1}^{n} w_j^2 = 1$.

Выполнение вышеперечисленных условий включено в предварительную обработку данных с целью их приведения в соответствие с принципом «чем больше, тем лучше». Исходя из этого принципа, эксперт ожидает, что увеличение значения некоторого показателя объекта приведет к увеличению его интегрального индикатора. Объект, имеющий максимальный по значению интегральный индикатор, называется наилучшим, а показатель, имеющий максимальный по значению вес, называется важнейшим в произвольных подмножествах соответственно объектов и показателей.

Результатом работы алгоритма построения интегрального индикатора методом «без учителя» является отыскание оптимального, по отношению к критерию информативности, вектора весов $\boldsymbol{w} = (w_1, \ldots, w_n)^\intercal$ свертки (1). Рассмотрим алгоритм получения интегрального индикатора «без учителя». Метод главных компонент, используемый для вычисления интегральных индикаторов [5], заключается в том, что к множеству описаний объектов применяется преобразование вращения, которое соответствует критерию *наибольшей информативности* С. Р. Рао [6]. Согласно этому критерию, наибольшая информативность есть минимальное значение суммы квадратов расстояния от описаний объектов до их проекций на первую главную компоненту.

Наилучшим выбором линейных функций, для которых остаточная дисперсия, предсказания с помощью линейного предиктора, минимальна, является выбор первых k главных компонент случайной величины A.

Для нахождения первой главной компоненты требуется найти такие линейные комбинации $Z^\intercal = WA^\intercal$ векторов-столбцов матрицы A, что векторы-столбцы $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ матрицы Z обладали бы наибольшей дисперсией: $\max \sum\limits_{j=1}^{n} \mathsf{D}\boldsymbol{z}_j$ при ограничениях нормировки $WW^\intercal = I$ — единичная матрица. Рао было показано, что строки матрицы W есть собственные векторы ковариационной матрицы $\Sigma = A^\intercal A$. Значение интегрального индикатора $\boldsymbol{q}$ вычисляется как проекция векторов-строк матрицы A на первую главную компоненту, $\boldsymbol{q} = A\boldsymbol{w}$, где $\boldsymbol{w}$ — вектор-столбец матрицы $W^\intercal$, соответствующий наибольшему собственному значению матрицы Σ.

Поиск устойчивых интегральных индикаторов

Для получения интегральных индикаторов, устойчивых к выбросам, в рамках линейной модели ранее было предложено использовать регуляризацию. А. М. Шурыгин в работе [2] рассмотрел два способа регуляризации ковариационной матрицы Σ. Первый способ — регуляризация посредством ридж-регрессии, $\Sigma_{r\beta} = \Sigma + \beta I$, где β — регуляризующий множитель. Второй способ — диагональная регуляризация $\Sigma_{d\nu} = (1 - \nu)\Sigma + \nu \mathrm{diag}(\Sigma)$, где $\nu \in [0, 1]$ — регуляризующий множитель. Было показано, что второй способ дает лучшую устойчивость к выбросам.

Использование регуляризации приводит к потере информативности. Поставим задачу так, чтобы сохранить значение критерия наибольшей информативности на опорном множестве описаний.

Задано множество описаний объектов, $S_0 = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$. Обозначим $\mathcal{S} = \{S_1, \ldots, S_l\}$ — множество всех подмножеств S_0, в котором число элементов $l = 2^m$. Алгоритм, вычисляющий наиболее информативный линейный предиктор, использует множество S_ξ, отыскивает веса $\boldsymbol{w}_\xi = \boldsymbol{w}(S_\xi) \in \mathbb{R}^n$ и возвращает интегральный индикатор $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi \in \mathbb{R}^m$. Обозначим $\bar{S}_\xi$ дополнение S_ξ до S_0. Исключим из рассмотрения тривиальные пары $(S_\xi, \bar{S}_\xi)$, в которых $\#S_\xi = 1$ и $\bar{S}_\xi = \varnothing$. Будем считать, что значения показателей объектов являются независимыми случайными величинами и принята гипотеза Гауссовского распределения этих величин.

Пусть $p_\xi = P(\boldsymbol{a}_i \in S_\xi)$ обозначает вероятность принадлежности некоторого объекта из S_0 множеству S_ξ, и $\bar{p}_\xi$ — вероятность того, что этот объект принадлежит дополнению до S_0. Найдем в $\mathcal{S}$ такое опорное множество S_ξ, для которого отношение $f_\xi = p_\xi / \bar{p}_\xi$ максимально.

Рассмотрим суммарные дисперсии σ_ξ и $\bar{\sigma}_\xi$ проекций $\boldsymbol{p}_i$ элементов $\boldsymbol{a}_i$ множеств S_ξ и $\bar{S}_\xi$ на первые главные компоненты, определяемые матрицей S_ξ. Обозначим $n_\xi, \bar{n}_\xi, n_0$ — число элементов во множествах $S_\xi, \bar{S}_\xi, S_0$ соответственно. Суммарная дисперсия проекций $\boldsymbol{p}_i$ элементов множеств S_ξ и $\bar{S}_\xi$ всей выборки $\sigma^2(S_0)$ равна сумме дисперсий каждой выборки, взвешенных вероятностями принадлежности вектора $\boldsymbol{a}_i$ с проекцией $\boldsymbol{p}_i$ множествам $S_\xi, \bar{S}_\xi$,

$$\sigma^2(S_0) = p_\xi^2 \sigma^2(S_\xi) + p_\xi^2 \sigma^2(\bar{S}_\xi) = \frac{p_\xi^2 \sigma_\xi^2}{n_\xi} + \frac{\bar{p}_\xi^2 \bar{\sigma}_\xi^2}{\bar{n}_\xi}. \quad (3)$$

Для получения выражения отношения вероятностей f_ξ минимизируем дисперсию $\sigma^2(S_0)$. Так как выражение (3) должно удовлетворять равенству $n_\xi + \bar{n}_\xi = n_0$, при дифференцировании используем метод множителей Лагранжа, обозначив мно-

житель λ. Тогда

$$L = \sigma^2(S_0) + \lambda(n_\xi + \bar{n}_\xi - n_0) = \\ = \frac{p_\xi^2\sigma_\xi^2}{n_\xi} + \frac{\bar{p}_\xi^2\bar{\sigma}_\xi^2}{\bar{n}_\xi} + \lambda(n_\xi + \bar{n}_\xi - n_0).$$

Приравняв частные производные по λ и по n_ξ к нулю, получаем

$$\frac{\partial L}{\partial n_\xi} = -\frac{p_\xi^2\sigma_\xi^2}{n_\xi^2} + \lambda = 0, \quad \frac{\partial L}{\partial \lambda} = n_\xi + \bar{n}_\xi - n_0 = 0,$$

откуда получаем $p_\xi\sigma_\xi = n_\xi\sqrt{\lambda}$. Из двух последних выражений $n_0\sqrt{\lambda} = (p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)$ и $p_\xi = n_\xi(p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)(n_0\sigma_\xi)^{-1}$. Продифференцировав лагранжиан L по $\bar{n}_\xi$, получим аналогичное отношение для вероятности $\bar{p}_\xi$. Искомое отношение вероятностей равно

$$\frac{p_\xi}{\bar{p}_\xi} = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}. \tag{4}$$

Таким образом, вероятность принадлежности описания объекта одному из множеств прямо пропорциональна мощности этого множества и обратно пропорциональна среднеквадратичному отклонению. Искомый интегральный индикатор $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi$ доставляется таким множеством S_ξ, для которого отношение $f_\xi = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}$ максимально.

Результаты

Был выполнен сравнительный анализ регионов России по уровню загрязнения ртутью основных продуктов питания. Каждому региону был поставлен в соответствие интегральный индикатор, указывающий на загрязненность продуктов. Были рассмотрены три показателя загрязненности: мясные продукты, молочные продукты и хлебобулочные изделия. Использовались данные 29 регионов. Данные нормированы следующим образом. В каждом регионе для каждого из трех показателей был проведен ряд стандартизованных измерений. Элемент a_{ij} матрицы описаний — величина загрязнения j-го продукта в i-м регионе. Его значение есть отношение квантиля уровня 0,9 распределения содержания ртути в серии измерений к величине предельно допустимой концентрации ртути в данном продукте.

Предложенный алгоритм отыскивает опорное множество S_ξ с целью вычисления весов показателей $\boldsymbol{w}_\xi$ для получения интегральных индикаторов, устойчивых к выбросам. Алгоритм состоит из трех шагов: назначения ядра опорного множества, отыскания опорного множества и вычисления интегрального индикатора.

1. Отыскивается центр исходного множества. Для этого находится вектор-среднее по всем компонентам векторов $\boldsymbol{a}_i$, вошедших в выборку S_0, и изымается вектор, наиболее удаленный в евклидовой метрике. Это действие производится итеративно, до получения последнего вектора, который и является центром. Для сокращения времени работы алгоритма, две трети описаний объектов, наименее удаленных от центра, были занесены в ядро опорного множества.

2. Исходное множества S_0 разбивается на множества S_ξ и $\bar{S}_\xi$ таких, что S_ξ включает ядро опорного множества в качестве собственного подмножества, а $\bar{S}_\xi$ являются объектами-выбросами. Для каждого разбиения вычисляется целевая функция $f_\xi = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}$, где $n_\xi, \bar{n}_\xi$ — мощности множеств $S_\xi, \bar{S}_\xi$; и $\sigma_\xi, \bar{\sigma}_\xi$ — суммарная дисперсия проекций объектов множеств $S_\xi, \bar{S}_\xi$ на собственные векторы ковариационной матрицы, определяемой множествами $S_\xi, \bar{S}_\xi$. Из множества полученных функций f_ξ выбираем функцию, на которой достигается максимум.

3. Объекты выбранного опорного множества S_ξ задают матрицу «объект–показатель» A_ξ. Для нее вычисляется ковариационная матрица $\Sigma = A_\xi^\intercal A_\xi$. Первый собственный вектор матрицы Σ определяет веса $\boldsymbol{w}_\xi$ показателей исходного множества [7]. Интегральный индикатор объектов, вычисленный с помощью предложенного алгоритма, есть $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi$.

Множество исходных данных — описаний регионов — содержит три выброса по второму показателю (молочные продукты) в трех регионах: республика Карелия, г. Санкт-Петербург, Московская область. Данные Карелии, кроме того, содержат выброс по всем трем показателям. Эти три региона не вошли в опорное множество объектов.

Таблица 1. Веса показателей до и после применения алгоритма.

$\boldsymbol{w}$	Без регуляризации	С регуляризацией	С опорным множеством
w_1	0,0204	0,2264	0,4693
w_2	0,9983	0,7687	0,7706
w_3	0,0548	0,5982	0,4312

В таблице 1 показано распределение весов показателей, полученных для трех алгоритмов построения интегральных индикаторов. Первый алгоритм — применение метода главных компонент к исходным данным без использования регуляризации. Второй алгоритм — метод главных компонент с регуляризацией. Был выбран метод диагональной регуляризации, так как полученные с помощью его результаты доставили большее значение критерию устойчивости, чем результаты, полученные с помощью регуляризации посредством ридж-регрессии. Третий алгоритм — метод главных компонент для опорного множества описаний объектов. При использовании первого алгоритма выбро-

сы по второму показателю приводили к неадекватному увеличению вклада этого показателя в интегральный индикатор. Предложенный метод доставляет более адекватные значения весов показателей, как показано в последнем столбце таблицы.

Для иллюстрации результатов работы алгоритмов был введен критерий устойчивости $\varphi = \arg\min_{\Phi} \|\boldsymbol{w}_A - \boldsymbol{w}_{A^*}\|_2$, где множество Φ определено как

$$\Phi = \left\{\boldsymbol{a}^* \colon \|\boldsymbol{a}^*\|_2 = \max \|\boldsymbol{a}_i\|_2,\ i = 1, \ldots, m\right\}.$$

Вектор $\boldsymbol{w}_A$ был получен с помощью метода главных компонент для исходной матрицы A. Вектор $\boldsymbol{w}_{A^*}$ получен был получен с помощью метода главных компонент для матрицы A с присоединенным вектором-столбцом $\boldsymbol{a}^*$, который рассматривался как выброс. Значение критерия устойчивости было вычислено для трех алгоритмов: без использования регуляризации, с диагональной регуляризацией и с предложенным алгоритмом выбора опорного множества. В первом случае значение критерия устойчивости составило $\varphi = 0{,}4727$, во втором $\varphi = 0{,}0962$ и в третьем $\varphi = 0{,}0$.

Следует отметить, что алгоритм, использующий диагональную регуляризацию, позволяет получить адекватный индикатор, но тем не менее влияние объектов-выбросов на индикатор полностью не исключено. Вектор $\boldsymbol{q}_2$ — индикатор, полученный с помощью диагональной регуляризации, вектор $\boldsymbol{q}_3$ — индикатор, полученный с помощью алгоритма выбора опорного множества описаний объектов. Коэффициент ранговой корреляции был использован для сравнения в связи с тем, что он инвариантен относительно монотонных преобразований интегральных индикаторов и учитывает только порядок их значений, игнорируя при этом величину выбросов.

Алгоритм, не использующий регуляризацию, вычисляет интегральный индикатор, который существенно зависит от наличия в выборке объектов-выбросов. Коэффициент ранговой корреляции между интегральным индикатором, полученным посредством такого алгоритма, и между интегральным индикатором, полученным с помощью опорного множества, равен 0,82. Это означает, что у 37 пар, из всех возможных пар элементов двух индикаторов, порядок следования объектов отличается. В таблице 2 приведены примеры таких пар. В столбцах $\boldsymbol{q}_1$ и $\boldsymbol{q}_3$ приведены значения интегральных индикаторов указанных регионов. В столбцах $r(\boldsymbol{q}_1)$ и $r(\boldsymbol{q}_3)$ приведены ранговые номера регионов.

Таблица 2. Значения интегрального индикатора без регуляризации и интегрального индикатора, построенного на основе опорного множества.

Регион РФ	$\boldsymbol{q}_1$	$r(\boldsymbol{q}_1)$	$\boldsymbol{q}_3$	$r(\boldsymbol{q}_3)$
Архангельская обл.	0,5367	19	0,8356	23
Хабаровский край	0,7986	21	0,6165	19
...	...	...	...	...
Владимирская обл.	0,0324	12	0,3577	14
Краснодарский край	0,0449	16	0,1578	10

Заключение

В работе рассмотрена задача построения устойчивых интегральных индикаторов. При построении индикаторов предлагается выбирать из заданного множества описаний объектов опорное множество, используя предложенный критерий вероятности принадлежности описаний объектов этому множеству. Алгоритм построения интегральных индикаторов с выбором опорного множества является альтернативой алгоритмам, которые используют регуляризацию. В отличие от них, в предложенном алгоритме влияние объектов-выбросов на интегральный индикатор исключено. Предложенный алгоритм был использован для получения интегральных индикаторов регионов России по уровню загрязнения основных продуктов питания.

Литература

[1] *Орлов А. И.* Современный этап развития теории экспертных оценок. Заводская лаборатория, 1996, № 1.

[2] *Шурыгин А. М.* Прикладная стохастика: робастность, оценивание, прогноз. — М.: Финансы и статистика, 2000. — С. 99.

[3] *Nabney I. T.* NETLAB: Algorithms for pattern recognition. Springer, 2004. — Pp. 330.

[4] *Зубаревич Н. В., Тикунов В. С., Крепец В. В., Стрижов В. В., Шакин В. В.* Многовариантные методы интегральной оценки развития человеческого потенциала в регионах Российской Федерации // ГИС для устойчивого развития территорий. — Петропавловск-Камчатский, 2001. — С. 84–105.

[5] *Strijov V., Shakin V.* Index construction: the expert-statistical method. Environmental research, engineering and management. 2003. — № 4(26). — Pp. 51–55.

[6] *Рао С. Р.* Линейные статистические методы и их применения. — М.: Наука, 1968. — С. 530–533.

[7] *Jolliffe I. T.* Principal Component Analysis, 2nd ed., Springer, 2002.