УДК 519.2

Т.В. Казакова, В.В. Стрижов
Вычислительный центр им. А.А. Дородницына РАН,
г. Москва
tatiana_kazakova@mail.ru

Устойчивые интегральные индикаторы с выбором опорного множества описаний объектов*

Исследуется задача построения интегрального индикатора «без учителя», устойчивого к изменениям множества описаний объектов. Объекты описаны в линейных шкалах. При построении интегрального индикатора выбирается такое опорное множество, которое доставляет максимум критерия устойчивости.

Введение

Пусть каждый объект из заданного множества описан вектором, компоненты которого являются результатами измерений соответствующих показателей. Все измерения выполнены в линейных шкалах. Интегральный индикатор – скаляр, поставленный в соответствие объекту.

Распространенным алгоритмом построения интегральных индикаторов для объектов, описанных в линейных шкалах, является линейная комбинация значений показателей. Веса назначаются экспертами или вычисляются исходя из некоторого критерия информативности описаний. Метод главных компонент, предложенный С.А. Айвазяном для получения интегрального индикатора [1], использует дисперсионный критерий информативности показателей. Веса показателей при этом совпадают с элементами первой главной компоненты, а интегральный индикатор вычисляется как проекция объектов на первую главную компоненту. Альтернативный метод вычисления интегрального индикатора «без учителя» – метод сингулярных векторов [2]. В этом случае интегральный индикатор является проекцией объектов на первый правый сингулярный вектор. Интегральные индикаторы, вычисленные методом главных компонент и методом сингулярных векторов, совпадают.

Однако если отдельные объекты имеют значения показателей, существенно отличающиеся от показателей основного числа объектов, то такие объекты – объекты-выбросы – имеют большее влияние на веса показателей, чем прочие объекты. На практике используют два способа решения этой проблемы: исключение подобных объектов из выборки и разбиение множества объектов на несколько классов, внутри которых производится сравнение. Часто эти способы неприемлемы из-за самой постановки прикладной задачи: необходимо найти такую свертку – интегральный индикатор, которая бы адекватно, с точки зрения экспертов, описывала все элементы множества объектов.

* Работа поддержана грантом РФФИ 04-01-00401.

Существует несколько алгоритмов получения устойчивых интегральных индикаторов с использованием как линейных [3], так и нелинейных [4], [5] моделей. В рамках линейной модели используется регуляризация. А.М. Шурыгин в работе [3] рассмотрел два способа регуляризации ковариационной матрицы: регуляризация посредством ридж-регрессии и диагональная регуляризация. Было показано, что второй способ дает лучшую устойчивость к выбросам. Однако подобные алгоритмы используют регуляризирующий множитель, что приводит к задаче поиска такого значения множителя, которое доставляло бы оптимальную потерю информативности. Поставим задачу так, чтобы избежать появления такого множителя.

Поиск устойчивых интегральных индикаторов

Пусть значения показателей есть независимые случайные величины с неизвестной плотностью распределения. Будем считать случайными не только значения показателей, но и сам факт попадания объектов в выборку. Пусть каждый объект попадает в выборку с вероятностью, пропорциональной числу объектов. Рассмотрим индикаторы произвольных подмножеств выборки и выберем подмножество, имеющее устойчивый индикатор и состоящее из опорных описаний.

Задано множество описаний объектов $S_0 = \{\mathrm{a}_{1.},...,\mathrm{a}_{m.}\}$. Обозначим $\mathrm{S} = \{S_1,...,S_l\}$ – множество всех подмножеств S_0 и $\mathrm{Q} = \{\mathrm{q}_1,...,\mathrm{q}_l\}$, $\mathrm{W} = \{\mathrm{w}_1,...,\mathrm{w}_l\}$ – множества соответствующих им интегральных индикаторов и весов показателей, $l = 2^m$. Алгоритм, вычисляющий наиболее информативный линейный предиктор, получает множество S_ξ, отыскивает веса $\mathrm{w}_\xi = \mathrm{w}(S_\xi)$ и возвращает индикатор $\mathrm{q}_\xi = A\mathrm{w}_\xi \in R^m$. Обозначим $S_\xi^{'}$ – дополнение S_ξ до S_0. Для простоты обозначений дальнейшие рассуждения будут проводиться для фиксированного значения ξ.

Пусть $p_1 = P(\mathrm{a}_{i.} \in S)$ обозначает вероятность принадлежности некоторого объекта множеству S, и p_2 – вероятность того, что этот объект принадлежит дополнению до S_0. Найдем в S такое множество S, для которого отношение $p_1 / p_2 \to \max$. Множество, доставляющее этому критерию максимум, называется множеством опорных векторов.

Рассмотрим суммарные дисперсии σ_1 и σ_2 проекций объектов множеств S и $S^{'}$ на первые главные компоненты, определяемые матрицей S. Обозначим n_1, n_2, n – число элементов в множествах $S, S^{'}, S_0$ соответственно. Суммарная дисперсия проекций элементов S и $S^{'}$ всей выборки $\sigma^2(x)$ равна сумме дисперсий каждой выборки, взвешенных вероятностями принадлежности вектора с проекцией x ко множествам $S, S^{'}$, $\sigma^2(S_0) = p_1^2\sigma^2(S) + p_1^2\sigma^2(S^{'}) = n_1^{-1}p_1^2\sigma_1^2 + n_2^{-1}p_2^2\sigma_2^2$. Для получения выражения отношения вероятностей минимизируем дисперсию $\sigma^2(S_0)$. Так как предыдущее выражение должно удовлетворять ограничению $n_1 + n_2 = n$, при дифференцировании используем метод множителей Лагранжа, обозначив множитель λ. Тогда $L = \sigma^2(\tilde{x}) + \lambda(n_1 + n_2 - n) = p_1^2\sigma_1^2 / n_1 + p_2^2\sigma_2^2 / n_2 + \lambda(n_1 + n_2 - n)$. Приравняв

частные производные по λ и по n_1 к нулю, получаем $\partial L/\partial n_1 = -p_1\sigma_1^2/n_1^2 + \lambda = 0$, $\partial L/\partial\lambda = n_1 + n_2 - n = 0$. Следовательно, $p_1\sigma_1 = n_1\sqrt{\lambda}$. Из двух последних выражений $n\sqrt{\lambda} = p_1\sigma_1 + p_2\sigma_2$ и $p_1 = n_1(p_1\sigma_1 + p_2\sigma_2)/n\sigma_1$. Продифференцировав лагранжиан L по n_2, получим аналогичное отношение для вероятности p_2. Искомое отношение вероятностей равно $p_1/p_2 = n_1\sigma_2/n_2\sigma_1$. Таким образом, вероятность принадлежности описания объекта опорной выборке прямо пропорциональна мощности выборки и обратно пропорциональна дисперсии выборки.

Результаты

Одним из авторов был выполнен сравнительный анализ регионов РФ по уровню загрязнения основных продуктов питания ртутью. Матрица описаний содержит информацию по 29 регионам и 3 группам продуктов питания. Это мясные продукты, молочные продукты и хлебобулочные изделия. Данные нормируются с учетом предельно допустимой концентрации ртути по каждому продукту.

Предварительный анализ показал наличие выбросов по молочным продуктам (второй показатель) в двух регионах. Кроме того, в одном из регионов зафиксирован выброс по всем трем показателям. Предложенный алгоритм выбирает опорное множество, удаляя из исходной выборки регионы с выбросами. До применения алгоритма выбросы по второму показателю приводили к неадекватному увеличению его вклада в интегральный индикатор (рис. 1). В результате сравнение данных осуществлялось по второму показателю (табл. 1, в скобках показано ранговое значение интегрального индикатора). Веса показателей рассчитывались на основе метода главных компонент.

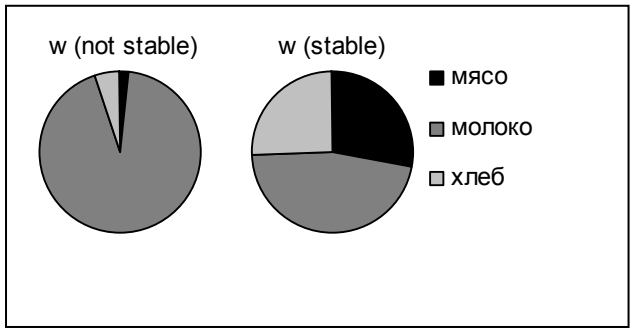

Рисунок 1 – Веса показателей до и после применения алгоритма

Таблица 1 – Исходные данные и значения интегральных индикаторов

Регион РФ \ Продукт	Мясо	Молоко	Хлеб	q (not stable)	q (stable)
Архангельская область	0,5	0,5	0,5	0,5367 (19)	0,8356 (23)
Хабаровский край	0	0,8	0	0,7986 (21)	0,6165 (19)
…	…	…	…	…	…
Владимирская область	0,3333	0	0,4667	0,0324 (12)	0,3577 (14)
Краснодарский край	0,1	0,032	0,2	0,0449 (16)	0,1578 (10)

Заключение

В работе рассматривается задача построения устойчивых интегральных индикаторов. При построении индикатора предлагается выбирать опорное множество векторов-описаний объектов из фактор-множества. Каждому набору этого множества ставится в соответствие суммарная дисперсия проекций описаний этого набора на первые главные компоненты. Опорным считается такое множество, элементы которого доставляют максимум отношению вероятностей принадлежности элементов к опорному множеству и к его дополнению. Описанный алгоритм построения интегрального индикатора является альтернативой алгоритмам, которые используют регуляризацию. В отличие от них в предложенном алгоритме влияние объектов-выбросов на интегральный индикатор исключено.

Литература

1. Айвазян С.А. Интегральные индикаторы качества жизни населения: их построение и использование в социально-экономическом управлении и межрегиональных сопоставлениях. – М.: ЦЭМИ РАН, 2000. – С. 56.
2. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. – М.: Мир, 1969. – С. 15–18.
3. Шурыгин А.М. Прикладная стохастика: робастность, оценивание, прогноз. – М.: Финансы и статистика, 2000. – С. 99.
4. Nabney I.T. – NETLAB: Algorithms for pattern recognition. – Springer, 2004. – P. 330.
5. 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 // Сб. ГИС для устойчивого развития территорий. Материалы Междунар. конф. – Петропавловск-Камчатский, 2001. – С. 84–105.

Т.В. Казакова, В.В. Стрижов

Стійкі інтегральні індикатори з вибором опорної множини описів об'єктів

Досліджується задача побудови інтегрального індикатора «без вчителя», стійкого до змін множини описів об'єктів. Об'єкти описані в лінійних шкалах. При побудові інтегрального індикатора вибирається така опорна множина, що доставляє максимум критерію стійкості.

T.V. Kazakova, V.V. Strijov

Stable Integral Indicators with the Choice of Objects Features for a Support Set

The problem of stable integral indicators for an object set is considered. The objects are featured in the linear scales. To construct a stable integral indicator one has to choose an objects features subset such that causes the maximal value to the stable criterion.

Статья поступила в редакцию 26.04.2006.