ВОПРОСЫ ОБРАБОТКИ ИНФОРМАЦИИ ПРИ ПРОЕКТИРОВАНИИ СИСТЕМ
(Вычислительные системы)
1977 год Выпуск 69

УДК 51:155.001.57

ОБ АДЕКВАТНЫХ ПАРНЫХ МЕРАХ СХОДСТВА В ЗАДАЧАХ РАСПОЗНАВАНИЯ ОБРАЗОВ С РАЗНОРОДНЫМИ ПРИЗНАКАМИ

А.М.Шусторович

В настоящее время практически во всех работах по распознаванию образов, как теоретических, так и прикладных, используется понятие "сходства" между объектами. Вне зависимости от постановки задачи в образ объединяются сходные объекты; решение о принадлежности нового объекта к одному из образов также чаще всего основывается на сходстве с уже изученными объектами или образами.

Анализ свойств мер сходства наиболее важен для широкого круга прикладных задач, в которых исходным материалом служит набор разнородных показателей, чаще всего с участием количественных характеристик. Такая ситуация обычна для задач классификации объектов по косвенным данным, возникающих в биологии, геологии, географии и других науках и часто формулируемых в терминах распознавания образов.

Практически всегда формализация "сходства" происходит через парные меры сходства, заданные на парах объектов (см., например, [I]). В данной работе рассматриваются свойства одного класса парных мер сходства между объектами, охарактеризованными разнородными признаками, измеренными в шкалах порядка, интервалов или отношений.

Пусть $\{f_i\}$, $i = 1,\dots,M$, – набор признаков, измеренных в фиксированных шкалах m_i на эмпирическом множестве A, $m=\{m_i\}$, $m(A)$ – совокупность шкальных значений по набору признаков, $x,y \in m(A) \subset R^m$. Для того чтобы числовую функцию $\Phi(x,y)$ можно было рассматривать в качестве меры сходства, на нее накладываются три общепринятых условия:

I) Непрерывность. Величина $\Phi(x,y)$ непрерывна, как функция $m(A)\ m(A) \to R$ в естественной топологии R^{2M}.

2) **Симметричность.** $\Phi(x,y) = \Phi(y,x)$.

3) **Диапазон значений и свойства тождества.** $0 < \Phi(x,y) \le 1$ (или у других авторов $0 \le \Phi(x,y) \le 1$), причем $\Phi(x,y) = 1 \Leftrightarrow x = y$.

Как известно [2], шкалы интервалов и отношений определяются с точностью до группы положительных линейных преобразований $\Gamma_p = \{\gamma_{\alpha,\beta}: x \to \alpha x + \beta,\ \alpha \in R^+,\ \beta \in R\}$ и ее подгруппы растяжений $\Gamma_d = \{\gamma_{\alpha,0}: \alpha \in R^+\}$ соответственно, что отражает возможность произвольного выбора единицы измерения и начала координат; шкала порядка определена с точностью до более широкого множества Γ_n всех монотонно возрастающих непрерывных отображений совокупности своих шкальных значений в R.

Пусть $\gamma = (\gamma_1, \dots, \gamma_M)$, где $\gamma_i \in \Gamma_d$ или Γ_p, или Γ_n. Преобразования $x' = \gamma(x)$ допустимы для всего набора с учетом типа составляющих его шкал m_i, причем разнородность признаков понимается как возможность независимого выбора γ_i для формирования общего преобразования γ.

Рассмотрим сейчас класс непараметрических мер сходства, когда $\Phi = \Phi(x,y)$ независимо от выбора шкал из множества эквивалентных. От таких мер сходства в задачах с разнородными признаками необходимо потребовать как минимум сохранения порядка на парах объектов при всех допустимых преобразованиях γ: $m \to m'$. Отсюда имеем условие

4) **Адекватность.**

$$\Phi(x,y)\{\le\}\Phi(z,t) \Leftrightarrow \Phi(x',y')\{\le\}\Phi(z',t')$$

для произвольных допустимых γ и $x,y,z,t \in m(A)$.

Функции, удовлетворяющие условиям I–4, назовем адекватными (парными) мерами сходства. Отметим, что область определения исследуемых функций – произведение 2M экземпляров R или R^+. Заметим также, что требование $\Phi(x,y) \le \Phi(z,t) \to \Phi(x',y') \le \Phi(z',y')$ эквивалентно условию 4, хотя и выглядит слабее.

Описание возможного вида адекватных мер сходства в зависимости от типа используемых шкал дают теоремы I и 2.

ТЕОРЕМА I. Класс адекватных мер сходства для шкал порядка пуст.

ДОКАЗАТЕЛЬСТВО. Предположим противное. Пусть x,y измерены в одномерной шкале порядка, Φ – искомая функция, тогда для произвольного монотонного непрерывного отображения γ: R→R верно

$$\Phi(x,y)\{\le\}\Phi(z,t) \to \Phi(\gamma(x),\gamma(y))\{\le\}\Phi(\gamma(z),\gamma(t)).$$

Подберем четверку $x < y < z < t$ такую, что без ограничения общности $\Phi(x,y) \le \Phi(z,t) \ne 1$. Рассмотрим семейство допустимых преобразований γ_n таких, что $\gamma_n(y) = y$, $\gamma_n(z) = z$, $\gamma_n(t) = t$, а $\gamma_n(x) = y - 1/n$. Из непрерывности Φ и $\Phi(y,y) = 1$ получаем $\Phi(\gamma_n(x),y) > \Phi(x,y)$ при некотором n, а, по определению адекватности, должно быть $\Phi(\gamma_n(x), y) \le \Phi(z,t)$. Противоречие.

ЛЕММА I[*]. **Класс адекватных мер сходства для одномерной шкалы интервалов состоит из функций вида $\xi(|x-y|)$, где ξ – произвольное убывающее непрерывное отображение R в $[0,1]$, $\xi(0) = 1$.**

Сделаем замену переменных $\{a = x - y, b = x + y\}$. Тогда условия I-4 для функции $\Psi(a,b) = \Phi(x,y) = \Phi((a+b)/2, (b-a)/2)$ будут иметь вид:

1') $\Psi(a,b)$ непрерывна;

2') $\Psi(a,b) = \Psi(-a,b) = \Psi(|a|,b)$

3') $0 \le \Psi(a,b) \le 1$; $\Psi(a,b) = 1 \Leftrightarrow a = 0$;

4') $\Psi(a,b)\{\le\}\Psi(c,d) \Leftrightarrow \Psi(a',b')\{\le\}\Psi(c',d')$, где для шкалы отношений допустимы преобразования $\{a' = \gamma \cdot a,\ b' = \gamma \cdot b\}$, а для шкалы интервалов – $\{a' = \gamma a,\ b' = \gamma b + \eta\}$.

ДОКАЗАТЕЛЬСТВО. Рассмотрим произвольные $a \ne 0$, $\lambda > 1$, b_0, b_1 и предположим, что $\Psi(\lambda \cdot a, b_0) \ge \Psi(a, b_1)$. Рассмотрим допустимое преобразование $\{c' = c,\ d' = d + \eta\}$, где $\eta = (b_0 - \lambda \cdot b_1)/(\lambda - 1)$. При этом $b_1' = (b_0 - b_1)/(\lambda - 1)$, $b_0' = \lambda(b_0 - b_1)/(\lambda - 1)$. Обозначим $t = (b_0 - b_1)/(\lambda - 1)$, тогда, по условию 4, должно быть $\Psi(\lambda \cdot a, \lambda \cdot t) \ge \Psi(a, t)$. Рассмотрим другое допустимое преобразование: $\{c' = c/\lambda,\ d' = d/\lambda\}$. Применяя его многократно, получаем

$$\Psi(a,t) \ge \Psi(a/\lambda,\ t/\lambda),$$

$$\Psi(a/\lambda, t/\lambda) \ge \Psi(a/\lambda^2,\ t/\lambda^2),$$

$$\dots\dots\dots\dots\dots\dots\dots\dots$$

$$\Psi(a/\lambda^{n-1}, t/\lambda^{n-1}) \ge \Psi(a/\lambda^n,\ t/\lambda^n).$$

Отсюда следует $\Psi(a,t) \ge \Psi(a/\lambda^n, t/\lambda^n)$ при всех n. По непрерывности, $\Psi(a,t) \ge \Psi(0,0) = 1$, откуда следует, что $a = 0$. Получено противоречие. Таким образом, можно утверждать, что $|x-y| < |z-t| \to$

*) Доказательство леммы I и теоремы 2 для шкал интервалов получено совместно с М.Я. Финкельштейном.

$\to \Phi(x,y) > \Phi(z,t)$, а это и означает строгое убывание меры сходства как функции от $|x-y|$.

Для использования в дальнейшем обозначим внутренность полосы $|a| < C$, $C = const$, где $\Psi(a,b) > S$, через $I(S)$; ее внешность – через $E(S)$.

ЛЕММА 2. Всякая адекватная мера сходства для одномерной шкалы отношений обладает тем свойством, что ее линии уровня $\Phi(x,y) = S$ разоивают (при $S \neq 1$) область ее определения (плоскость или первую четверть) на три области: внутреннюю, содержащую прямую или луч $y = x$, и две внешние, симметричные относительно этой прямой.

Поскольку условие 4 для шкал отношений слабее, чем для шкал интервалов, то меры сходства, описываемые леммой 1, должны удовлетворять условиям леммы 2. Соответствующие области для функций $\Phi(x,y) = \xi(|x-y|)$ имеют вид полос: $I(S)$ – внутренняя полоса, $E(S)$ – две внешних.

ДОКАЗАТЕЛЬСТВО. Рассмотрим произвольные $a \neq 0$, $\lambda > 1$, b. Тогда справедливо $\Psi(a,b) > \Psi(\lambda \cdot a, \lambda \cdot b)$. Доказательство повторяет вторую часть доказательства леммы 1. Из этого следует, что на каждом луче $\{(a,b) = r \cdot (a_0, b_0),\ r > 0\}$ найдется не более одной точки, где $\Psi(a,b) = S$, причем внутренний полуинтервал соответствует точкам, где $\Psi(a,b) > S$, а внешняя часть – где $\Psi(a,b) < S$. Область, заметаемую соответствующими полуинтервалами (лучами), обозначим $I(S)$, ее внешность – $E(S)$; первая содержит прямую (или луч) $y = x$ ($a = 0$), вторая распадается на две симметричные области при $x > y$ и $x < y$ соответственно.

Предположим теперь, что в задаче рассматриваются одновременно два или более признака, измеренные в шкалах отношений или интервалов. Тогда уместна

ТЕОРЕМА 2. Класс адекватных мер сходства для шкал отношений или интервалов пуст при $M \geq 2$.

ДОКАЗАТЕЛЬСТВО. Допустим существование адекватной меры сходства $\Psi(a,b)$, где a и b – двумерные переменные, $a = (a_1, a_2)$, $b = (b_1, b_2)$. Тогда $\Psi(a,b) = \Psi(a_1, b_1; a_2, b_2)$. Рассмотрим сужение функции Ψ на плоскости $a_1 = b_1 = 0$ и $a_2 = b_2 = 0$, обозначим $\Psi_1(a_1, b_1) = \Psi(a_1, b_1; 0, 0)$, $\Psi_2(a_2, b_2) = \Psi(0, 0; a_2, b_2)$; Ψ_1 и Ψ_2 можно рассматривать как адекватные меры сходства на соответствующих плоскостях. Рассмотрим S_1, S_2, $0 < S_1, S_2 < 1$. Подберем такие $a_1, b_1; a_2, b_2; c_1, d_1; c_2, d_2$, что $(a_1, b_1) \in I_1(S_1)$, $(a_2, b_2) \in E_2(S_2)$, $(c_1, d_1) \in E_1(S_1)$, $(c_2, d_2) \in I_2(S_2)$. Попробуем сравнить $\Psi(a_1, b_1; a_2, b_2)$ и $\Psi(c_1, d_1; c_2, d_2)$. Предположим, что $\Psi(a_1, b_1; a_2, b_2) \geq \Psi(c_1, d_1; c_2, d_2)$. Применим многократно допустимое преобразование $\{a' = \gamma \cdot a,\ b' = \gamma \cdot b\}$, где $\gamma = (1/2, 1)$. Получим набор неравенств:

$$\Psi(a_1/2^n, b_1/2^n; a_2, b_2) \geq \Psi(c_1/2^n, d_1/2^n; c_2, d_2)$$

при всех n. В силу непрерывности, $\Psi(0,0; a_2, b_2) \geq \Psi(0,0; c_2, d_2)$, т.е. $\Psi_2(a_2, b_2) \geq \Psi_2(c_2, d_2)$, что противоречит $\Psi_2(c_2, d_2) > S_2 > \Psi_2(a_2, b_2)$ из-за $(a_2, b_2) \in E_2(S_2)$, $(c_2, d_2) \in I_2(S_2)$. Такое же противоречие получим, предположив $\Psi(c_1, d_1; c_2, d_2) \geq \Psi(a_1, b_1; a_2, b_2)$, если последовательно применить $\gamma = (1, 1/2)$. Доказательство проходит практически без изменения и при $M > 2$.

Непосредственным следствием доказанных теорем является вывод о невозможности корректного введения мер рассматриваемого класса в практически интересных задачах с двумя и более признаками; последнее ведет к необходимости использовать в качестве мер сходства параметризованные семейства функций или даже межгрупповые адекватные меры сходства, не сводимые к парным; на их основе также можно строить разнообразные по своим возможностям алгоритмы распознавания образов [3].

Как правило, на практике используются именно параметризованные меры сходства, что позволяет сохранять их значение неизменными при замене шкал на эквивалентные. Чаще всего при этом используются нормировки, как например, при делении на среднеквадратичное уклонение или на максимальный разброс значений по каждому из признаков [4]. С вычислительной точки зрения нормировки эквивалентны введению специальных единиц измерения (например, соответственно единиц среднеквадратичного или долей максимального разбросов).

В таких случаях никакое пополнение выборки, при котором изучается старое эмпирическое множество, не должно приводить к изменению этих специальных единиц из-за изменений в нормировках; из полученных выше результатов следует, что в противном случае неизбежно обновятся выводы о сходстве некоторых ранее рассмотренных объектов. Последнее заставляет с недоверием относиться и к параметризованным мерам сходства и вынуждает каждый раз предварительно проверять представительность выборки относительно способа параметризации: любое допустимое (не выводящее за пределы изучаемого эмпирического множества) пополнение выборки не должно менять значений параметров для каждого фиксированного набора шкал.

Для примера отметим, что этому требованию, как правило, не удовлетворяют статистические нормировки, а также и нормировка на выборочный максимальный разброс значений признака. С другой стороны, с указанной точки зрения непротиворечива нормировка на теоретический максимальный разброс, если, конечно, он поддается определению.

Литература

1. ДОРОФЕЮК А.А. Алгоритмы автоматической классификации. (Обзор литературы.) -В кн.: Проблемы расширения возможностей автоматов. Под ред. Айзермана М.А. Вып.I. М., 1976.

2. ПФАНЦАГЛЬ И. Теория измерений. М., 1976.

3. ШУСТОРОВИЧ А.М. Таксономия и распознавание образов в задачах с разнородными признаками. -В кн.: Математическое программирование и смежные вопросы. (Труды УП зимней школы, Дрогобыч, 1974.) Выпуклое программирование. М., 1976, с.176-181.

4. ВОРОНИН Ю.А. Введение мер сходства и связи для решения геолого-геофизических задач.-"Докл. АН СССР", 1971, т. 199, № 5, с. 1011-1015.

Поступила в ред.-изд.отд.
8 января 1977 года