ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)
1984 год Выпуск 102

УДК 519.8

К ПРОБЛЕМЕ КЛАССИФИКАЦИИ МЕР СХОДСТВА

Ю.А.Воронин, А.И.Жигарловский, Н.Г.Шевченко

I. Об актуальности построения классификации мер сходства. Разработка эффективного сценария постановки и решения задач распознавания, наличие которого является необходимым условием проектирования соответствующих пакетов прикладных программ, сталкивается с многими трудностями, обусловленными прежде всего отсутствием необходимых классификаций алгоритмов распознавания [2]. Попытки построения необходимой классификации алгоритмов распознавания даже в том простейшем случае, когда эти алгоритмы интерпретируются только как решающие правила [I,2], заставляют думать, что предварительно необходимо ставить вопрос о классификации мер сходства, на которые так или иначе опираются алгоритмы распознавания. Как можно выяснить, вопрос о классификации мер сходства является весьма актуальным и в связи с другими аспектами разработки упомянутого выше сценария. Например, систематизацией всех известных сейчас мер сходства, порождением новых мер, выбором мер сходства с целью оптимизации распознавания [I,2,4].

2. О первоочередных целях построения классификации мер сходства и ее оценке. Будем считать, что пока цели построения классификации мер сходства следует связывать только с систематизацией всех известных сейчас мер сходства и порождением новых мер. Имея в виду оценку нашей классификации мер, будем считать, что об ее оценке можно говорить только в том случае, когда будет показано, что эта классификация удовлетворяет общим требованиям теории классификаций [2] и имеется по крайней мере еще одна другая классификация мер сходства, удовлетворяющая тем же требованиям и целям. При этом можно будет сравнивать эти классификации мер сходства с точки зрения полноты охвата известных мер, детальности, ра-

зумности их различения, оригинальности и эффективности новых мер [3].

3. О поле сходства. Говоря о мерах сходства, мы имеем в виду сравнение уже перенумерованных объектов некоторого фиксированного множества объектов, описанных с точки зрения фиксированной сово - купности унарных градуированных свойств [3]:

$$A = (a^n),\ n = 1,2,\dots;\ A = (a_i),\ i = 1 \div N_\Psi;$$
$$\Psi = (\phi_l),\ l = 1 \div L,\ \phi_l = (\phi_l^m),\ m = 1 \div M_l. \qquad (1)$$

Каждому объекту a_i из A ставится в однозначное соответствие вектор значений свойств и его вероятность:

$$a_i:\ \Psi_i = (\phi_l^i);\ P_i = p(\Psi_i),\ i = 1 \div N_\Psi. \qquad (2)$$

Фактически мы имеем дело с формально нефиксированным множеством объектов A и формально заданной совокупностью свойств Ψ, а также эмпирическим материалом

$$(a^{n'}),\ n' = 1 \div N';\ (a_{i'}, \Psi_{i'}),\ i' = 1 \div N'_\Psi. \qquad (3)$$

При переходе от (2) к (3) мы говорим о задании поля сходства [2]. Выше ϕ_l из Ψ могут толковаться и как прямые (задающие), и как косвенные (распознающие) свойства, определенные и измеренные с одинаковой точностью на всех a^n из A, независимых между собой [3].

4. О допустимых мерах сходства. Условимся иметь ввиду только допустимые меры сходства, т.е. такие, которые могут быть представлены в виде некоторых функций $\lambda_\Psi(a_i, a_j)$, имеющих явные предельные представления на случай любого одного свойства ϕ_l из Ψ и на случай бесконечного числа свойств ϕ_l из Ψ, или $\lambda_{\phi_l}(a_i, a_j)$ и $\lambda_{\phi_\infty}(a_i, a_j)$ (ограничения по форме представления), удовлетворяют некоторым требованиям (ограничения по свойствам). Эти требования сформулируем так [2]:

I) ограниченности: $\lambda^* \le \lambda_\Psi(a_i, a_j) \le \lambda^{**}$;

2) симметрии: $\lambda_\Psi(a_i, a_j) = \lambda_\Psi(a_j, a_i)$;

3) максимальной похожести: $\lambda_\Psi(a_i, a_j) = \lambda^{**} \Leftrightarrow \Psi_i = \Psi_j$;

4) минимальной похожести: $\lambda_\Psi(a_i, a_j) = \lambda^* \Leftrightarrow \Psi_i = (\phi_l^{m'}),\ \Psi_j = (\phi_l^{m_l})$

или $\phi_l^i,\ \phi_l^j = \phi_l^{m'},\ \phi_l^{m_l},\ l = 1 \div L$;

5) инвариантности по допустимым шкальным преобразованиям: $\lambda_{\Psi}(a_i,a_j) = \lambda_{\Psi'}(a_i,a_j)$ $\Psi' = H(\Psi)$;

6) согласования по порядку:
$\lambda_{\Psi}(a_i,a_i)$, $\lambda_{\Psi}(a_i,a_j) \geq \lambda_{\Psi}(a_i,a_j)$, $a_i \leq a_i \leq a_j$;

7) предельной определенности: $\lambda_{\Psi}(a_i, a_j) = \lambda(\Delta_{\phi_i}(i, j))$, $\lambda_{\phi_\infty}(a_i,a_j) = \lambda_\infty(i,j)$.

Существенно, что требования 3 и 4 можно конкретно формулировать различным образом. Мы здесь предусматриваем только одну конкретную формулировку для требования 3 и две конкретные формулировки для требования 4. Среди свойств ϕ_i из Ψ могут быть свойства, отвечающие любой шкале измерения. Однако без ущерба для общности и ради краткости мы далее предполагаем, что все ϕ_i из Ψ измерены в абсолютной шкале. Порядок в A может задаваться различным образом, не зависящим от $\lambda_{\Psi}(a_i,a_j)$, может задаваться лишь частично, может вообще не задаваться. Мы считаем, что $\Delta_{\phi_i}(i,j)$ является допустимым взвешенным оператором сравнения

$$\Delta_{\phi_i}(i,j) \equiv \left(\frac{|\phi_i^i-\phi_i^j|}{\phi_i^1-\phi_i^1}\right)^{\alpha_i}\left(\delta_i^1\,\frac{\phi_i^i-\phi_i^j}{\phi_i^1-\phi_i^1} + \delta_j^1\,\frac{\phi_i^i-\phi_i^j}{\phi_i^1-\phi_i^1}\right)^{\beta_i}, \qquad (4)$$

а $\lambda_\infty(i, j)$ - некая функция от значков i и j. Заметим, что вопрос о классификации допустимых взвешенных операторов сравнения (4) так же, как и других допустимых взвешенных операторов, отвечающих ϕ_i, измеренных в других шкалах, рассматривался в [2].

5. <u>О категориях мер сходства.</u> Введем представления о категориях мер сходства $\lambda_{\Psi}(a_i, a_j)$, интерпретируя их как функции $\lambda(\Psi_i, \Psi_j, \ldots)$. Рассматривая область прибытия функции $\lambda(\Psi_i,\Psi_j,\ldots)$, учтем, что ей может отвечать либо множество вещественных чисел, либо множество номеров. В этом смысле можно говорить о сильных и слабых мерах сходства $\lambda_{\Psi}(a_i, a_j)$. Рассматривая область отправления функции $\lambda(\Psi_i,\Psi_j,\ldots)$, учтем, что она может определяться с привлечением или без привлечения вероятностей P_i и P_j, без привлечения или с привлечением произвольных параметров (различных по числу, относительному к L), с линейной или нелинейной зависимостью от этих параметров. Это позволяет говорить о вероятностных и детерминированных, непараметрических и параметрических, линейных и нелинейных мерах сходства. Сейчас можно ограничиться грубым выделением категорий мер сходства, фиксированных в табл. I.

Таблица I

Категории мер сходства

Меры сходства		Сильные		Слабые	
		детерминированные	вероятностные	детерминированные	вероятностные
Непараметрические		I	2	3	4
Параметрические	Линейные	5	6	7	8
	Нелинейные	9	I0	II	I2

Не составляет труда сконструировать на основе уже известных мер сходства $\lambda_{\psi}(a_i, a_j)$ меру любой из указанных категорий [3].

6. О видах мер сходства. Имея в виду какую-либо фиксированную категорию мер сходства $\lambda_{\phi}(a_i, a_j)$, можно предложить для них, например, следующую классификацию видов. Выделим два типа мер сходства так: если все $\lambda_{\phi_l}(a_i, a_j)$ оказываются для всех ϕ_l (отвечающих одной и той же шкале) одинаковыми, то будем говорить об изотропных мерах сходства, в иных случаях будем говорить об анизотропных мерах сходства. Два рода мер сходства выделим следующим образом [5]: если при взаимной замене в ϕ_i и ϕ_j значений ϕ_i^i и ϕ_i^j мера сходства $\lambda_{\phi}(a_i, a_j)$ не изменится, то будем говорить о сильно симметричных мерах сходства; в иных случаях будем говорить о слабо симметричных мерах сходства. Учитывая, что условие (4) может быть конкретно сформулировано двумя способами, получим классификацию видов мер сходства, заданную в табл. 2. Как легко убедиться все до сих пор известные меры сходства $\lambda_{\psi}(a_i, a_j)$ относятся к виду 5. Можно догадываться, что возможно построение мер сходства и других видов [2].

7. О новых видах мер сходства. Для доказательства возможности построения новых мер сходства $\lambda_{\psi}(a_i, a_j)$ достаточно построить новые меры сходства $\lambda(z_i, z_j)$ между комплексными числами z_i и z_j в предположении, что $z = z(x,y)$, $0 \le x \le 1$ и $0 \le y \le 1$. Н.Г.Шевченко была предложена следующая формула:

Таблица 2

Классификация видов мер сходства

Меры сходства	Изотропные		Анизотропные	
	Сильно симметричные	Слабо симметричные	Сильно симметричные	Слабо симметричные
С двумя непохожими объектами	I	2	3	4
С многими непохожими объектами	5	6	7	8

$$\lambda(z_i, z_j) = \{\delta_x(1-|x_i-x_j|) + \delta y(1-|y_i-y_j|)\}\cdot$$

$$\cdot \frac{x_i x_j + y_i y_j}{\sqrt{(x_i)^2 + (y_i)^2}\sqrt{(x_j)^2+(y_j)^2}}. \qquad (5)$$

Е.Д.Москаленским была построена формула

$$\lambda(z_i,z_j) = 1 - \frac{1}{\delta_{12}}\sqrt{\delta_1^2[|x_i-x_j|+(y_i-y_j)]^2+\delta_2^2[|x_i-x_j|-(y_i-y_j)]^2}. \qquad (6)$$

Ю.А.Ворониным были предложены формулы:

$$\lambda(z_i,z_j) = 1-\max\{\Delta_k^+(z_i,z_j), \Delta_k^-(z_i,z_j)\}, \qquad (7)$$

$$\left.\begin{aligned}\Delta_k^+(z_i,z_j) &= \alpha_x^+\Delta_k(x_i,x_j,R_x) + \alpha_y^+\Delta_k(y_i,y_j,R_y)\\ \Delta_k^-(z_i,z_j) &= \alpha_x^-\Delta_k(x_i,x_j,R_x) + \alpha_y^-\Delta_k(y_i,y_j,R_y)\end{aligned}\right\}, \qquad (8)$$

$$\alpha_x^+ = \begin{cases}\alpha_x, & x_i > x_j,\\ 0, & x_i \le x_j,\end{cases} \quad \alpha_y^+ = \begin{cases}\alpha_y, & y_i > y_j,\\ 0, & y_i \le y_j,\end{cases}$$

$$\alpha_x^- = \begin{cases}0, & x_i > x_j,\\ \alpha_x, & x_i \le x_j,\end{cases} \quad \alpha_y^- = \begin{cases}0, & y_i > y_j,\\ \alpha_y, & y_i \le y_j,\end{cases} \qquad (9)$$

$$\Delta_k(x_i, x_j, R_x) = |x_i - x_j|^{r^k_{2x}} (x_i + x_j)^{r^k_{2x}},$$
$$\Delta_k(y_i, y_j, R_y) = |y_i - y_j|^{r^k_{2y}} (y_i + y_j)^{r^k_{2y}}. \qquad (10)$$

Легко убедиться, что (5),(6) и (7) действительно могут интерпретироваться как допустимые меры сходства $\lambda(z_i, z_j)$ видов 2,4,6,7 и 8. На этом основании можно утверждать, что возможно построение новых мер сходства $\lambda_\Psi(a_i, a_j)$, исключая, быть может, меры сходства видов I и 3. Больше того, из (5)-(I0) в общих чертах видно, как получать для мер сходства $\lambda_\Psi(a_i, a_j)$ новых видов 2,4,6,7 и 8 общие формулы [2].

8. Заключение. Как можно показать [3], в случае, когда мы имеем дело с зависимыми свойствами ϕ_i из Ψ, когда допустимы не все логически возможные векторы значений свойств ϕ_i [2], введение мер сходства $\lambda_\Psi(a_i, a_j)$ приобретает некие особенности. Например, в случае компонентного состава, когда $\sum_i \phi_i^i = 100$, для $\lambda_\Psi(a_i, a_j)$ нельзя воспользоваться, положим формулой

$$\sum_{i=1}^{L} \delta_i \left(1 - \frac{|\phi_i^i - \phi_i^j|}{\phi_i^{m_i} - \phi_i^1}\right), \qquad (11)$$

но можно воспользоваться формулой

$$\prod_{i=1}^{L} \delta_i \left(1 - \frac{\phi_i^i - \phi_i^j}{\phi_i^{m_i} - \phi_i^1}\right). \qquad (12)$$

Однако и в этом случае предложенные выше классификационные представления для $\lambda_\Psi(a_i, a_j)$ сохраняют смысл.

Можно считать, что предложенная выше классификация мер сходства отвечает целям, сформулированным в п.2. Наличие такой клас - сификации позволяет выделить различные сознательные подходы к выбору мер сходства при распознавании: когда категория, тип, род и вид меры выбираются эвристически (или задачно [3]), а оценка параметров в мере производится задачно (или эвристически). Это заставляет по-новому подходить к построению и оценке сценария постановки и решения задач распознавания. Любой сознательный выбор - это прежде всего классифицирование [2].

9. О частных мерах близости между химическими объектами.

Когда говорят о близости объектов [2], имеют в виду их сходство с учетом взаимоотношений между собой или другими объектами. Рассмотрим один из возможных подходов к введению мер близости. Пусть $\mathcal{A} = (a_i)$, $i=1 \div J$, - некоторое множество объектов, а $B(b_q)$, $q=1 \div Q$, - некоторое соотносительное к нему множество объектов. Положим, что любой a_i из $\mathcal{A}$ может соотноситься с любым b_q из B каким-либо одним способом из $R = (r_h)$, $h = 1 \div H$. Например, $\mathcal{A}$ - множество видов минералов, B - множество видов пород, R - множество возможных соотношений видов минералов и пород (данный вид минерала в данном виде пород может отсутствовать или присутствовать в относительно малых, средних или больших количествах, равномерно или неравномерно размещаясь). Например, $\mathcal{A}$ - множество химических элементов, B - множество химических соединений, R - множество возможных реакций между элементом и соединением (при некоторой заданной классификации реакций, которая может быть фиксирована различным образом [2,6]).

Фиксируем a_i и a_j из $\mathcal{A}$. Обозначим через $B(i,h)$ и $B(j,h)$ подмножества тех b_q из B, которые соотносятся с a_i и a_j способом r_h, или находятся с каждым из них в отношении r_h. Положим $B(i,j,h) = B(i,h) \cap B(j,h)$. Пусть $n(i,h)$, $n(j,h)$ и $n(i,j,h)$ - число b_q в $B(i,h)$, $B(j,h)$ и $B(i,j,h)$. Обозначим $\max n(i,h) = n^{**}(h)$ и $\min n(i,h) = n^{*}(h)$. Найдем:

$$\mu_h(a_i,a_j) = \left(\frac{2n(i,j,h)}{n(i,h) + n(j,h)}\right)^{\alpha}\left(1 - \frac{|n(i,h)-n(j,h)|}{n^{**}(h) - n^{*}(h)}\right)^{\beta}. \quad (13)$$

Используя (13), определим интересующую нас меру близости так:

$$\mu_R(a_i,a_j) = \sum_{h=1}^{H} \gamma_h \mu_h(a_i,a_j), \quad \gamma_h > 0, \quad \sum_{h=1}^{H} \gamma_h = 1. \quad (14)$$

Иногда способ соотношения a_i с b_q, отношение r_h зависят от внешних условий PT, т.е. $r_h = r_h(PT)$ [2]. Тогда вместо (14) следует записать

$$\mu_R(a_i,a_j) = \sum_{PT} \delta(PT)\mu_{R(PT)}(a_i,a_j), \quad \delta(PT) > 0, \quad \sum_{PT} \delta(PT) = 1. \quad (15)$$

Если воспользоваться результатами [2,7], то формулу (15) можно обобщить, заменив

$$\left(\frac{2n(i,j,h)}{n(i,h)+n(j,h)}\right) \quad \text{и} \quad \frac{|n(i,h)-n(j,h)|}{n^{**}(h)-n^{*}(h)},$$

например, на

$$\frac{2n(i,j,h)}{(1+U)[n(i,h)+n(j,h)]-2Un(i,j,h)}$$

и

$$\left[\frac{|n(i,h)-n(j,h)|}{n^{**}(h)-n^{*}(h)}\right]^{\varphi}\left(\frac{n(i,h)+n(j,h)}{n^{**}(h)+n^{*}(h)}\right)^{\psi}.$$

На основании такого обобщения можно считать, что так называемое химическое сходство или родство [2,8], вообще говоря, не имеет смысла вне фиксированных целевых установок.

Литература

1. ВАСИЛЬЕВ В.И. Распознающие системы. Справочник. – Киев: Наукова думка, 1983. – 262 с.

2. ВОРОНИН Ю.А. Введение в теорию классификаций. – Новосибирск, 1982. – 194 с.

3. ВОРОНИН Ю.А., ВЫСОКОС Г.Н., ГРАДОВА Т.А. Меры сходства в геологии. Отчет, 1978. – Фонды НЦ СО АН СССР, 240 с.

4. ЗАГОРУЙКО Н.Г. Методы обнаружения закономерностей. –Новое в жизни науки, техники. Сер. Математика и кибернетика. М., Знание. 1981. – 64 с.

5. МОСКАЛЕНСКИЙ Е.Д. О построении мер сходства, не обладающих избыточной симметрией. –В сб.: Вычислительные методы в геологоразведке. Новосибирск, 1984, с. 66–81.

6. БРАУН Т., ЛЕМЕЙ Г. Химия в центре наук. Ч.1. –М.: Мир, 1983. – 435 с.

7. АНДРЕЕВ В.Л. Классификационные построения в экологии и систематике. –М.: Наука, 1980. – 142 с.

8. ШТРЕКЕР Э. Атомистическое обоснование химии и ее развитие как системной науки. –В кн.: Философские проблемы современной химии. М., 1971, с. 148–159.

Поступила в ред.-изд.отд.
20 апреля 1984 года