Доклады Академии наук СССР
1971. Том 199, № 5

МАТЕМАТИКА

Ю. А. ВОРОНИН

ВВЕДЕНИЕ МЕР СХОДСТВА И СВЯЗИ ДЛЯ РЕШЕНИЯ ГЕОЛОГО-ГЕОФИЗИЧЕСКИХ ЗАДАЧ

(Представлено академиком Г. И. Марчуком 8 II 1971)

При решении задач выделения, описания, группирования и распознавания геолого-геофизических объектов прямое использование теоретико-вероятностной меры встречается с рядом трудностей, обусловленных малыми объемами выборок, разнородностью и многомерностью описания объектов. Большие трудности возникают и при интерпретации результатов, полученных таким путем. По этим причинам многие предпочитают использовать при решении упомянутых задач другие меры, которые обычно называют мерами сходства и связи. Во многих случаях использование таких мер, традиционных для геологии, как представляется, дает очевидные преимущества. Однако эти меры сходства и связи, в отличие от теоретико-вероятностной меры, не имеют аксиоматического обоснования: в одних и тех же конкретных ситуациях разными авторами они вводятся различно, часто недостаточно четко. Представляется желательным разработать аксиоматический подход к построению таких мер.

Пусть задана таблица — объекты $(a_1, a_2, \ldots, a_n)$, свойства $(f_0, f_1, \ldots, f_m)$:

$$\{f_\alpha^i\}, \quad f_\alpha^i \in (f_\alpha), \quad i = 1, 2, \ldots, n, \quad \alpha = 0, 1, \ldots, m, \qquad (1)_1$$

где через $f_\alpha{}^i$ обозначено значение свойства f_α на объекте a_i, через (f_α) — множество всех возможных значений свойства f_α. Будем говорить, что свойство f_α является: арифметическим, если на (f_α) определены арифметические операции; логическим 1-го рода, если на (f_α) определены только отношения порядка и эквивалентности; логическим 2-го рода, если на (f_α) определено только отношение эквивалентности.

Принято под мерой сходства между объектами a_i и a_j по свойствам $F = (f_1, f_2, \ldots, f_m)$ понимать функцию $\Lambda(F^i, F^j)$, отвечающую аксиомам (1): 1) $0 \leqslant \Lambda(F^i, F^j) \leqslant 1$; 2) $\Lambda(F^i, F^j) = \Lambda(F^j, F^i)$; 3) $F^i = F^j \Rightarrow \Rightarrow \Lambda(F^i, F^j) = 1$. Эта система аксиом неполна, она не позволяет однозначно определить вид $\Lambda(F^i, F^j)$. Дополним ее так, чтобы можно было однозначно определить вид $\Lambda(F^i, F^j)$ и указать способ определения ее параметров с учетом конкретной ситуации.

Пусть f_α — арифметическое свойство. Потребуем, чтобы мера сходства между объектами a_i и a_j по свойству f_α удовлетворяла аксиомам: 1) $0 \leqslant \Lambda(f_\alpha{}^i, f_\alpha{}^j) \leqslant 1$; 2) $\Lambda(f_\alpha{}^i, f_\alpha{}^j) = \Lambda(f_\alpha{}^j, f_\alpha{}^i)$; 3) $f_\alpha{}^i = f_\alpha{}^j \Rightarrow \Lambda(f_\alpha{}^i, f_\alpha{}^j) = 1$; 4) $\Lambda(f_\alpha{}^i, f_\alpha{}^j) = 1 \Rightarrow f_\alpha{}^i = f_\alpha{}^j$; 5) $\Lambda(f_\alpha{}^i, f_\alpha{}^j)$ *имела производную по* $f_\alpha{}^j$, *линейно зависящую от* $f_\alpha{}^j$.

Тогда получим

$$\Lambda(f_\alpha^i, f_\alpha^j) = [1 - ((f_\alpha^i - f_\alpha^j)/\Delta f_\alpha)^2], \qquad (2$$

$$\Delta f_\alpha = f_\alpha^{**} - f_\alpha^*, \quad f_\alpha^{**} = \max f_\alpha, \quad f_\alpha^* = \min f_\alpha.$$

Если f_α — свойство логическое 1-го рода, то $\Lambda(f_\alpha{}^i, f_\alpha{}^j)$ будем определять тоже формулой $(2)_1$, если же f_α — свойство логическое 2-го рода, $\Lambda(f_\alpha{}^i,$

$f_\alpha{}^j$) определим так:

$$\Lambda(f_\alpha^i, f^j{}_\alpha = \begin{cases} 1, & f_\alpha^i = f_\alpha^j, \\ 0, & f_\alpha^i \neq f_\alpha^j. \end{cases} \tag{2}$$

Если принять еще одну аксиому: 6) *мера сходства* $\Lambda(F^i, F^j)$ *должна определяться формулой*

$$\Lambda(F^i, F^j) = \sum_{\alpha=1}^{m} \delta_\alpha \Lambda(f_\alpha^i, f_\alpha^j), \quad \sum_{\alpha=1}^{m} \delta_\alpha = 1, \tag{3}$$

то получим ряд интересных возможностей. Параметры δ_α можно определять стандартно из условия

$$\sum_{ij} [\Lambda(f_0^i, f_0^j) - \Lambda(F^i, F^j)]^2 = \min, \tag{4}$$

$$i, j = 1, 2, \ldots, n, \quad i < j,$$

где f_0 — распознаваемое свойство.

Таблицу (1) можно описать так, чтобы ее описание не зависело от нумерации строк и столбцов, линейного преобразования столбцов $f_\alpha{}' = \varepsilon_\alpha{}^0 + \varepsilon_\alpha{}^1 f_\alpha$, числа строк и столбцов. Такое описание можно построить, например, следующим образом. Приведем таблице $(1)_1$ в соответствие таблицу:

$$\{\Lambda(F^i, F^j)\}, \quad i, j = 1, 2, \ldots, n. \tag{5}$$

Найдем такие a_h, для которых обращается в максимум выражение

$$T(i) = \Lambda(i)\left\{1 - \frac{1}{n}\sum_{j=1}^{n} [\Lambda(i) - \Lambda(F^i, F^j)]^2\right\}, \tag{6}$$

$$\Lambda(i) = \frac{1}{n}\sum_{j=1}^{n} \Lambda(F^i, F^j).$$

Возьмем любой a_k и назовем его голотипом. Зафиксируем μ — константу для разбиения объектов $a_1, a_2, \ldots, a_n$ на компоненты связности, $0 \leqslant \mu \leqslant 1$*. Пусть $A_1, A_2, \ldots, A_{n(\mu)}$ — компоненты связности, а $a_{k1}, a_{k2}, \ldots, a_{kn(\mu)}$ — их голотипы. Определим

$$\Lambda_s(\mu) = \min \Lambda(F^{ks}, F^j), \quad a_j \in A_s, \quad s = 1, 2, \ldots, n(\mu); \tag{7}$$

$$\widetilde{\Lambda}_s(\mu) = \Lambda(F^{ks}, F^s), \quad s = 1, 2, \ldots, n(\mu), \tag{8}$$

и, кроме того, найдем

$$\Lambda(\mu) = \frac{1}{n(\mu)} \sum_{s=1}^{n(\mu)} \Lambda_s(\mu), \tag{9}$$

$$\widetilde{\Lambda}(\mu) = \frac{1}{n(\mu)} \sum_{s=1}^{n(\mu)} \widetilde{\Lambda}_s(\mu). \tag{10}$$

Тогда функции

$$\theta(\mu) = (n(\mu) - 1) / (n - 1), \tag{11}_1$$

$$\varphi(\mu) = (\Lambda(\mu) - \Lambda(0)) / (1 - \Lambda(0)), \tag{11}_2$$

$$\psi(\mu) = \widetilde{\Lambda}(\mu) \tag{11}_3$$

---

* Совокупность объектов $A = (a_1, a_2, \ldots, a_l)$ является компонентой связности, если для любых a_i и $a_j \in A$ можно найти такие $a_{k_p} \in A$, что $\Lambda(F^i, F^{k_1}) \geqslant \mu$, $\Lambda(F^{k_1}, F^{k_2}) \geqslant \mu, \ldots, \Lambda(F^{k_\alpha}, F^j) \geqslant \mu$.

будут давать нужное описание таблицы $(1)_1$. Описание (11) является более грубым, чем описание через закон распределения. Закон распределения имеет смысл только тогда, когда все свойства f_α являются арифметическими, и он зависит от числа столбцов таблицы $(1)_1$. Функции (11) могут быть описаны через некоторые параметры $p_1(z)$, $p_2(z), \ldots, p_i(z)$, $z = 0, \varphi, \psi$. Это позволяет использовать формулу (3) для определения мер сходства между таблицами $(1)_1$, произвольными множествами объектов $a_{i_1}, a_{i_2}, \ldots, a_{i_h}$.

Обратимся к мерам связи между свойствами f_α и f_β. Потребуем, чтобы эта мера $\sigma(f_\alpha, f_\beta)$ являлась функцией от мер (2) и удовлетворяла аксиомам: 1) $0 \leqslant \sigma(f_\alpha, f_\beta) \leqslant 1$; 2) $\sigma(f_\alpha, f_\beta) = \sigma(f_\beta, f_\alpha)$; 3) $f_\alpha = \varepsilon_\beta^0 + \varepsilon_\beta^1 f_\beta \Rightarrow$ $\Rightarrow \sigma(f_\alpha, f_\beta) = 1$; 4) $\sigma(f_\alpha, f_\beta) = 1 \Rightarrow f_\alpha = \varepsilon_\beta^0 + \varepsilon_\beta^1 f_\beta$.

Тогда получим

$$\circ(f_\alpha, f_\beta) = \left(\sum_{ij} \Lambda(f_\alpha^i, f_\alpha^j)\,\Lambda(f_\beta^i, f_\beta^j)\right) \Big/ \left(\sqrt{\sum_{ij} \Lambda^2(f_\alpha^i, f_\alpha^j)}\ \sqrt{\sum_{ij} \Lambda^2(f_\beta^i, f_\beta^j)}\right), \quad (12)$$

$$i, j = 1, 2, \ldots, n, \quad i < j.$$

Если положить

$$\Lambda(f_\gamma^i, f_\gamma^j) = \begin{cases} f_\gamma^i - \bar{f}_\gamma, & f_\gamma^i = f_\gamma^j, \\ 0, & f_\gamma^i \neq f_\gamma^j, \end{cases} \qquad \gamma = \alpha, \beta, \quad (2)$$

то (12) перейдет в выборочный коэффициент корреляции. Аналогично можно получить коэффициенты Спирмэна, Кендалла, Чупрова и др. ([2]).

Использование (12) тоже открывает ряд интересных возможностей. Можно найти меру связи между двумя любыми совокупностями свойств $(f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_\lambda})$ и $(f_{\beta_1}, f_{\beta_2}, \ldots, f_{\beta_l})$, фиксируя распознаваемое свойство f_0. Если f_0 — распознаваемое свойство, «идеально хорошее», то легко построить таблицу

$$\{\Lambda(\tilde{f}_0^i, \tilde{f}_0^j)\}, \quad i, j, = 1, 2, \ldots, n, \quad i < j, \quad (13)$$

для «идеально плохого» свойства $\tilde{f}_0$ из условия $\sigma(f_0, \tilde{f}_0) = 0$. Тогда каждому свойству f_α можно приписать коэффициент информативности

$$\varkappa(f_\alpha / f_0) = \sigma(f_\alpha, f_0) - \sigma(f_\alpha, \tilde{f}_0). \quad (14)$$

Аналогично можно получить $\varkappa(f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_k} / f_0)$.

Из предыдущего следует, что предлагаемые меры позволяют ставить и решать, минуя введение теоретико-вероятностной меры, целый круг задач, важных для обработки любых данных, например, таких, как задача об оптимальном разбиении промежутка изменения свойства на интервалы, задача выбора оптимальной совокупности свойств, задача представительности выборки. Если $(f_\alpha^{**}, f_\alpha^*)$ разбит на интервалы $f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_l}$, то меру (2) можно записать так:

$$\Lambda(f_\alpha^i, f_\alpha^j) = [1 - ((p - q)/(l - 1))^2], \quad (2)$$

$$f_\alpha^i \in f_{\alpha_p}, \quad f_\alpha^j \in f_{\alpha_q}, \quad p, q = 1, 2, \ldots, l.$$

Можно построить алгоритм, позволяющий найти такое разбиение $(f_\alpha^{**}, f_\alpha^*)$ на $f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_l}$, которое при минимальном l обращает в максимум (14). Используя (12), можно получить таблицу

$$\{\sigma(f_\alpha, f_\beta)\}, \quad \alpha, \beta = 1, 2, \ldots, m', \quad \alpha < \beta. \quad (15)$$

Опираясь на (14) и (15), можно построить алгоритм, позволяющий при минимальном k найти такую $f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_k}$, которой отвечает максимальное значение $\varkappa(f_{\alpha_1}, f_{\alpha_2}, \ldots, f_{\alpha_k} / f_0)$. Таблица

$$\{f_\alpha^i\}, \quad i = 1, 2, \ldots, n', \quad n' \leqslant n, \quad \alpha = 1, 2, \ldots, m, \quad m' \leqslant m, \quad (1_2)$$

будет отвечать «представительной» выборке из таблицы $(1)_1$, если функции (11) для таблиц $(1)_1$ и $(1)_2$ оказываются, в некотором смысле, одинаковыми.

Имеется возможность получить ряд полезных теорем, учитывая, что меры (2) могут быть обобщены на случай, когда множество объектов несчетно.

Таблицу (5) можно интерпретировать как граф со «взвешенными» ребрами. Однако таблицы (1), по-видимому, можно рассматривать как новые математические объекты, поскольку для них иначе формулируются проблемы изоморфизма, изоморфного вхождения и пр. (*). По-видимому, предлагаемые меры следует рассматривать как основу для грубого, детерминированного, решения задач, упомянутых в самом начале. Опираясь на такие решения, в некоторых случаях можно будет ставить вопрос о более тонком, теоретико-вероятностном, решении этих задач.

Вычислительный центр
Сибирского отделения Академии наук СССР
Новосибирск

Поступило
25 I 1971

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] В. И. Васильев, Распознающие системы (справочник), Киев, 1969. [2] В. Ю. Урбах, Математическая статистика для медиков и биологов, Изд. АН СССР, М., 1963. [3] А. А. Зыков, Теория конечных графов, Новосибирск, 1969.