Г. М. Адельсон-Вельский, П. Е. Кунин, А. А. Леман, Об одном классе обучающихся алгоритмов узнавания,
Докл. АН СССР, 1967, том 173, номер 3, 532–534

https://www.mathnet.ru/dan32967

Доклады Академии наук СССР
1967. Том 173, № 3

УДК 519.95 КИБЕРНЕТИКА И ТЕОРИЯ РЕГУЛИРОВАНИЯ

Г. М. АДЕЛЬСОН-ВЕЛЬСКИЙ, П. Е. КУНИН, А. А. ЛЕМАН

ОБ ОДНОМ КЛАССЕ ОБУЧАЮЩИХСЯ АЛГОРИТМОВ УЗНАВАНИЯ

(Представлено академиком М. В. Келдышем 29 VII 1966)

Пусть A и B — непересекающиеся подмножества множества C. Алгоритм f называется алгоритмом узнавания для множеств A и B, если про всякий элемент $X \in A \cup B$ он или выдает ответ, что $X \in A$ (или $f(X) = A$), или что $X \in B$ (или $f(X) = B$), или отказывается от ответа ($f(X) = 0$).

В работе описан класс обучающихся алгоритмов, которые по заданным конечным подмножествам $\bar{A} \subset A$ и $\bar{B} \subset B$ создают алгоритм узнавания для множеств A и B.

В дальнейшем считается, что C — множество вершин $X(x_1, x_2, \ldots, x_n)$ n-мерного единичного куба; координаты $x_1, x_2, \ldots, x_n$ называются признаками.

М. М. Бонгард ([1]) предложил отыскивать комбинации признаков и их значений, характерных для $\bar{A}$ и $\bar{B}$. Предлагаемые итерационные алгоритмы позволяют при нахождении таких комбинаций избежать полного перебора и, следовательно, находить комбинации большого числа признаков.

Определение 1. Расстоянием между точками $X \in C$ и $Y \in C$ по системе признаков $(i_1, i_2, \ldots, i_k)$ называется

$$\rho_{i_1, i_2, \ldots, i_k}(X, Y) = \sum_{l=1}^{k} |x_{i_l} - y_{i_l}|. \tag{1}$$

Определение 2. Трубкой $T\{(i_1, i_2, \ldots, i_k); X^0; r\}$ называется множество точек $X \in C$, для которых

$$\rho_{i_1, i_2, \ldots, i_k}(X, X^0) \leqslant r. \tag{2}$$

Признаки $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ называются существенными для трубки T, X^0 называется центром, а r — радиусом трубки. Очевидно, центром трубки T является любая точка гиперплоскости $\{x_{i_l} = x_{i_l}^0\}$, $(l = 1, 2, \ldots, k)$.

Определение 3. Трубка T называется q-различающей для множеств M и N и функции $\varphi(x)$, если

$$\frac{\nu[T \cap M \cap \Phi_-] + \nu[T \cap N \cap \Phi_+] + \nu[T \cap (M \cup N) \cap \Phi_0]}{\nu[T \cap (M \cup N)]} < q, \tag{3}$$

где $\nu[\Delta]$ — число элементов множества Δ; Φ_0, Φ_- и Φ_+ — соответственно множества точек $X \in C$, для которых $\varphi(X) = 0$, $\varphi(X) < 0$, $\varphi(X) > 0$. Частным случаем q-различающих трубок являются q-различающие чистые M-трубки ${}_MT$, для которых $\varphi(X) \equiv 1$, и чистые N-трубки ${}_NT$, для которых $\varphi(X) \equiv -1$.

Определение 4. Система трубок $\{T_1, T_2, \ldots, T_s\}$ называется полной для множества D, если $D \subset \bigcup_i T_i$.

Пусть для множества $A \cup B$ существует полная система q-различающих трубок $\{T_1, T_2, \ldots, T_s\}$, где $s \leqslant \nu[A \cup B]$.

Очевидно, что для любых подмножеств $\bar{A} \subset A$ и $\bar{B} \subset B$ также существует полная система q-различающих трубок.

Алгоритм построения q-различающих чистых $\bar{A}$-трубок состоит в следующем.

Пусть $T\{(i_1, i_2, \ldots, i_k);\ X^0;\ R\}$ — некоторая трубка, $0 \leqslant \delta_0 \leqslant \delta_1 \leqslant 1$,

$$\sigma_{\bar{A}, l} = \sum_{X \in \bar{A} \cap T} x_l / \nu[\bar{A} \cap T], \quad l = 1, 2, \ldots, n. \tag{4}$$

Признак x_l объявляется существенным для трубки $\tilde{T}_r\{(\tilde{i}_1, \tilde{i}_2, \ldots, \tilde{i}_k);\ \tilde{X}^0;\ r\}$, если $\sigma_{\bar{A}, l} \leqslant \delta_0$ или $\sigma_{\bar{A}, l} \geqslant \delta_1$; в первом случае $\tilde{x}_l^0 = 0$, во втором $\tilde{x}_l^0 = 1$.

Пусть

$$\Psi(\tilde{T}_r) = \Psi(\nu[\bar{A} \cap \tilde{T}_r], \nu[\bar{B} \cap \tilde{T}_r]), \tag{5}$$

где $\Psi(n_1, n_2)$ — заданная функция, монотонно возрастающая по первому переменному и монотонно убывающая по второму. Радиус $\tilde{R}$ трубки $\tilde{T}\{(\tilde{i}_1, \tilde{i}_2, \ldots, \tilde{i}_k);\ \tilde{X}^0;\ \tilde{R}\}$ выбирается так, чтобы значение $\Psi(\tilde{T}_R)$ было максимальным среди всех $\Psi(\tilde{T}_r)$ при $\nu[\bar{A} \cap \tilde{T}_r] > \gamma$. Таким образом, алгоритм является итерационным.

В качестве начального центра можно выбрать произвольную точку $X \in \bar{A}$, а в качестве существенных признаков начальной трубки — все признаки $x_1, x_2, \ldots, x_n$. Количество итераций может быть задано заранее; можно также продолжать итерации, пока качество получаемых трубок не ухудшается. Для последней итерации требуется, кроме того, чтобы $\tilde{T}$ была q-различающей чистой $\bar{A}$-трубкой, т. е. $\Psi(\tilde{T}) < q$.

Описанный выше процесс может и не привести к построению q-различающей чистой трубки; в таком случае в качестве начальной нужно выбрать новую точку $X \in \bar{A}$.

Границы δ_0 и δ_1 могут зависеть от номера признака l. В одном из вариантов алгоритма эти границы определяются формулами

$$\delta_{0, l} = \sum_{X \in \bar{B}} x_l / \nu[\bar{B}] - \Delta_0, \tag{6}$$

$$\delta_{1, l} = \sum_{X \in \bar{B}} x_l / \nu[\bar{B}] + \Delta_1, \tag{6'}$$

где эталоны Δ_0, Δ_1 заданы.

Пусть уже построена система q-различающих чистых $\bar{A}$-трубок $T_1, {}^{A} T_2, \ldots, {}_{\bar{A}} T_\alpha\}$. Если она не является полной для множества $\bar{A}$ и не все точки $X \in \bar{A}$ испробованы в качестве начальных центров, то процесс построения новых q-различающих чистых $\bar{A}$-трубок можно продолжить. В качестве начального центра выбирается, например, точка $X \in \bar{A}$, наиболее удаленная от всех уже построенных центров (в метриках соответствующих трубок). Аналогично строится система q-различающих чистых $\bar{B}$-трубок.

После того как получена система q-различающих чистых A- и B-трубок, алгоритм f строится так, что

$$f(X) = A, \quad \text{если } X \in \bigcup {}_{\bar{A}} T_i \setminus \bigcup_j {}_{\bar{B}} T_j,$$

$$f(X) = B, \quad \text{если } X \in \bigcup_j {}_{\bar{B}} T_j \setminus \bigcup_i {}_{\bar{A}} T_i,$$

$$f(X) = 0 \quad \text{в остальных случаях.}$$

Описанный выше вариант алгоритмов этого класса был осуществлен в виде программы для ЭВМ. Проверка алгоритма показала, что он успешно находит чистые A- и B-трубки, вероятность попадания в которые достаточно велика ($c > 1 / \sqrt{m}$). Вместе с тем при малом количестве элементов

$\bar{A} \cup \bar{B}$ алгоритм имеет склонность к созданию «предрассудков», т. е. к отысканию чистых $\bar{A}$- и $\bar{B}$-трубок, не являющихся A- и B-трубками.

Процесс построения не чистых q-различающих трубок также является итерационным. Функция $\varphi(X)$ определяется следующим образом. Пусть $T\{(i_1, i_2, \ldots, i\ ); X^0; r\}$ — трубка, тогда

$$\varphi(X) = \omega_{\bar{A}}(X)/\{[\omega_{\bar{A}}(X) + \omega_{\bar{B}}(X)] - 1/2\},$$

где

$$\omega_\Delta(X) = \nu[\Delta \cap T] \prod_i \frac{\nu\{X' \in \Delta \cap T : x_i' = x_i\}}{\nu[\Delta \cap T]},$$

причем произведение берется по всем признакам, несущественным для трубки T. Эти формулы следуют из предположения, что признаки, несущественные для трубки T, некоррелированы для элементов трубки, а средние значения этих признаков близки к вероятностям $P\{x_i = 1\}$.

Институт
теоретической и экспериментальной физики

Поступило
18 VI 1966

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] М. М. Бонгард, Биофизика, 6, № 2 (1961).