ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Сборник трудов
Института математики СО АН СССР

1966 г. Выпуск 22

ПОЛУЧЕНИЕ ДОСТАТОЧНЫХ ХАРАКТЕРИСТИК $P_i(x)$ ПРИ РАСПОЗНАВАНИИ ОБРАЗОВ

Б.М. Курилов

Статистическая постановка задачи распознавания образов не нова [1,2,3]. Основным недостатком существующей постановки является требование точного знания плотностей распределения вероятностей образов и отсутствие метода получения достаточных оценок этих плотностей в процессе обучения по конечной выборке. Кроме того, в классической постановке не дается ответа на вопрос , какова должна быть выборка обучающей последовательности.

Это заставляет вернуться к статистической трактовке задачи распознавания образов.

Введем некоторые понятия и обозначения, которые понадобятся в дальнейшем.

Под о б р а з о м будем понимать некоторое подмножество из множества отражений предметов или явлений материального мира, выделенное воспринимающей системой в пространстве измеряемых ею параметров в замкнутую односвязную область.

Каждое отражение предмета (или явления) из этого подмножества будем называть р е а л и з а ц и е й образа. Реализация характеризуется n-мерным вектором в пространстве измеряемых системой параметров. Значение этого вектора есть случайная величина. Тогда задачу автоматического распознавания образов можно сформулировать на языке теории статистических решений.

Имеется $\pi_1, \pi_2, \ldots \pi_m$ - m- генеральных совокупностей с плотностями распределения $P_1(x)$, ... , $P_m(x)$ и априорными

вероятностями $q_1, q_2, \ldots, q_m$ соответственно. Заданы цены ошибочной классификации $C(i/i)$.

Необходимо выбрать решающее правило так, чтобы математическое ожидание потерь

$$\mathcal{P}(R, i/i) = \sum_{i=1}^{m} q_i \Big\{ \sum_{\substack{j=1 \\ j \neq i}}^{m} C(i/i) \int_{R_j} P_i(x)\,dx \Big\}$$

в фиксированном пространстве $\{R\}$ было минимальным.

Как известно [I], метод Бейеса дает минимум математического ожидания потерь в фиксированном $\{R\}$. Он заключается в том, что реализация относится к π_i, если результат измерения x попадает в область R_i. При равных ценах ошибочной классификации $C(i/i)$ область R_i состоит из тех точек x, для которых

$$q_i P_i(x) \geq q_j P_j(x)$$
$$(j = 0, 1, \ldots, m;\ j \neq i). \qquad (I)$$

Если

$$P\Big\{ \frac{P_i(x)}{P_j(x)} = a / \pi_h \Big\} = 0,$$

то функция распределения $P_i(x)/P_j(x)$ для любого π_h и совместная функция распределения двух отношений являются непрерывными (вытекает непосредственно из определения образа). Тогда, как показано в работе [I],

I) метод Бейеса является оптимальным в смысле минимума математического ожидания потерь,

2) любой допустимый метод является методом Бейеса,

3) класс методов Бейеса является минимальным полным классом.

2. Получение достаточных оценок $P_i(x)$ по конечной выборке

Метод Бейеса предполагает, что распределение всех генеральных совокупностей известны точно. Но при решении большинства практических задач распознавания образов эти распределения являются неизвестными. В процессе обучения распознающей системе предъявляется лишь конечная выборка из генеральных совокупностей. Рассмотрим вопрос о том, как можно получить достаточные оценки $P_i(x)$, используя эту выборку.

Ввиду того, что пространство измеряемых системой параметров $\{R\}$ всегда ограничено (из-за ограниченности пределов измерения каждого параметра), а точность измерения их не может быть сколь угодно большой (из-за наличия шумов), в практических задачах распознавания образов пространство $\{R\}$ всегда можно представить в виде конечного числа дискретных точек. Так, если число градации по $x_1, x_2, \ldots, x_n$ равно соответственно $l_1, l_2, \ldots, l_n$, то $\{R\}$ содержит в себе $\prod_{i=1}^{n} l_i$ дискретных точек.

Тогда отношение числа реализаций, попавших в точку с координатами x_ξ, к объему выборки N (частота), даст наилучшую (для данной выборки) оценку $P_i(x)$ в этой точке. Проделав эту операцию для всех x, получим гистограмму $\bar{P}_i(x)$, заданную в дискретных точках. Причем, согласно теореме Бернулли, $\bar{P}_i(x)$ стремится к $P_i(x)$ при $N \to \infty$, т.е.

$$\lim_{N \to \infty} P\Big(\Big|\frac{m}{N} - p\Big| < \varepsilon\Big) = 1,$$

где $\frac{m}{N}$ - частота события, а p - вероятность (ε -любое положительное число).

Процедуру получения оценок $P_i(x)$ можно осуществить следующим ооразом: задается пространство $\{R\}$ с координатными осями $x_1, x_2, \ldots, x_n$ и с числом градаций по осям $l_1, l_2, \ldots, l_n$, соответственно. Каждой точке этого дискретного пространства ставятся в соответствие некоторые ячейки памяти (например, ячейки памяти ЭВМ), число ячеек для представления одной точки пространства равно m (по числу распознаваемых образов). В процессе обучения системе предъявляются реализации образов из выборок объемом N_1, N_2, ..., N_m, и в ячейках памяти суммируется число попавших в эту точку реализаций. Полученная сумма делится на N_i и (при достаточно большом N) получается достаточно хорошая оценка плотностей распределения $P_i(x)$. Таким образом можно в принципе решить любую задачу распознавания образов.

Здесь подчеркивается слово "в принципе", т.к., во-первых, объем выборки при такой процедуре должен быть очень большим и, более того, трудно оценить даже приблизительно его величину; во-вторых, уже при не очень больших n, ℓ и m объем необходимой памяти становится настолько большим, что задача оказывается практически не реализуемой на самых мощных ЭВМ. При $n = 20$, $\ell_i = 100$, $m = 50$ (размерность многих практических задач распознавания образов), объем необходимой памяти равен $5 \cdot 10^{22}$ ячеек.

Покажем, что этот объем можно существенно сократить, не отходя от принципиальных позиций статистических методов.

Согласно (I), область принятия решения о принадлежности к генеральной совокупности π_i содержит все те точки x, для которых

$$q_i P_i(x) \geq q_j P_j(x) \quad (j \neq i).$$

Эта область в общем случае может содержать точки, для которых $q_j P_j(x) = 0$ для всех $j \neq i$ (область уверенных ответов $R_i^{(1)}$). Чтобы принять решения о точках, попавших в эту область, нет необходимости точно знать функцию $P_i(x)$, достаточно лишь знать, что она больше нуля.

С другой стороны, пространство $\{R\}$ может содержать такую область $R_i^{(3)}$, где $P_i(x) = 0$ (хотя бы для некоторых i) и область перекрытий $R_i^{(2)}$, где одновременно $P_i(x) \neq 0$ и $P_j(x) \neq 0$ (хотя бы для одного j). Тогда для каждого i пространство $\{R\}$ можно представить в виде суммы областей:

$$\{R\} = R_i^{(1)} + R_i^{(2)} + R_i^{(3)}.$$

Для принятия решения по (I) достаточно знать функцию $P_i(x)$, заданную лишь на $R_i^{(2)}$ и границы областей $R_i^{(1)}, R_i^{(2)}, R_i^{(3)}$.

Вероятность попадания точки в $R_i^{(2)}$ для реализаций k-ой генеральной совокупности есть

$$q_k \int_{R_i^{(2)}} P_i(x)dx \quad (i = 0,1,\ldots,m),$$

а вероятность получения от системы в целом неуверенных ответов выразится как

$$\sum_{i=1}^{m} q_i \int_{R_i^{(2)}} P_i(x)dx \qquad (i = 0,1,\ldots,m), \qquad (2)$$

На основании этого определим понятие информативности пространства. Будем говорить, что пространство измеряемых системой признаков $\{R\}$ достаточно информативно для распознавания образов π_i, если

$$\sum_{i=1}^{m} q_i \int_{R_i^{(2)}} P_i(x)dx \leq \tau \qquad (i = 1,2,\ldots,m),$$

где τ - заданная положительная величина в интервале $[0,1]$; и полностью информативно, если

$$\sum_{i=1}^{m} q_i \int_{R_i^{(2)}} P_i(x)dx = 0.$$

Для получения количественного суждения об информативности пространства $\{R\}$ нужно вычислить (2). В этом случае нет неоходимости точно знать $P_i(x)$ в ооласти $R_i^{(2)}$, а достаточно лишь иметь её оценку по конечной выборке объема N_i из генеральной совокупности π_i, причем величина N_i будет зависеть от задания точности и достоверности вычисления интеграла и не будет зависеть от вида функции распределения вероятности.

Для вычисления интеграла в выражении (2) воспользуемся методом статистических испытаний. Если зафиксирована ооласть $R_i^{(2)}$ (вопрос о получении границ областей $R_i^{(1)}$ и $R_i^{(2)}$ будет рассмотрен ниже) и системе представляются реализации случайного вектора x_ξ из выборки объема N, то отношение числа точек, попавших в область $R_i^{(2)}$, ко всему объему выборки даст искомую оценку [4], причем,

$$\lim_{N \to \infty} P\left(\left| \frac{n_{R_i^{(2)}}}{N} - p \right| < \varepsilon \right) = 1,$$

здесь

$$p = \int_{R_i^{(2)}} P_i(x)\,dx,$$

$n_{R_i}^{(2)}$ - число точек, попавших в область $R_i^{(2)}$,

$\mathcal{E}$ - сколь угодно малое положительное число,

то есть

$$\frac{n_{R_i}^{(2)}}{N} = \bar{\rho} \approx \rho . \qquad (3)$$

Будем говорить, что равенство (3) имеет точность $\mathcal{E}$ с надежностью α , если для неравенства

$$|\rho - \bar{\rho}| < \mathcal{E}$$

справедливо соотношение

$$P(|\rho - \bar{\rho}| < \mathcal{E}) = \alpha . \qquad (4)$$

Свяжем точность вычисления $\mathcal{E}$ и надежность α с необходимым объемом выборки N .

Воспользуемся для этого неравенством Чебышева:

$$P(|\bar{\rho} - \rho| < \mathcal{E}) \geq 1 - \frac{\sigma_{\bar{\rho}}^2}{\mathcal{E}^2} . \qquad (5)$$

Заметим, что $\bar{\rho}$ есть случайная величина с математическим ожиданием

$$M(\bar{\rho}) = \rho$$

и дисперсией

$$D(\bar{\rho}) = \frac{\rho(1-\rho)}{N} .$$

Тогда средняя квадратическая ошибка равенства (3) $\sigma_{\bar{\rho}}$ равна:

$$\sigma_{\bar{\rho}} = \sqrt{\frac{\rho(1-\rho)}{N}} . \qquad (6)$$

Нетрудно заметить, что она достигает максимума при ρ = 0,5. Сравнивая (4) и (5), запишем:

$$\alpha \geq 1 - \frac{\sigma_{\bar{\rho}}^2}{\mathcal{E}^2} .$$

Подставляя вместо $\sigma_{\bar{\rho}}^2$ его значение из (6) получим:

$$\alpha \geq 1 - \frac{\rho(1-\rho)}{N\mathcal{E}^2}$$

или

$$N = \frac{\rho(1-\rho)}{(1-\alpha)\mathcal{E}^2} , \qquad (7)$$

то есть для того, чтобы вычислить интеграл в выражении (2) с точностью $\mathcal{E}$ равной, например 0,0I и надежностью α = 0,95 при ρ = 0,5, необходимо взять выборку объёма N = 50000 испытаний.

Формула (7), полученная на основании неравенства Чебышева, дает сильно завышенное значение для N . Более точную оценку для N можно получить в том случае, если использовать закон распределения случайной величины $\bar{\rho}$. Известно [4] , что величина $\bar{\rho}$ имеет асимтотически (при $N \to \infty$) нормальное распределение.На этом основании равенство (4) можно записать в виде:

$$P\left(\frac{|\bar{\rho} - \rho|}{\sigma_{\bar{\rho}}} < t_\alpha\right) = \alpha ,$$

где t_α - величина критического интервала, которая выбирается из таолиц нормального распределения по заданному значению α. Сравнивая это соотношение с (5) , получаем:

$$\mathcal{E} = t_\alpha \sigma_\rho$$

или

$$\mathcal{E} = t_\alpha \sqrt{\frac{\rho(1-\rho)}{N}} .$$

Тогда

$$N = \frac{\rho(1-\rho)}{\mathcal{E}^2} t_\alpha^2 . \qquad (8)$$

Для $\mathcal{E}$ = 0,0I, α = 0,95 и ρ = 0,5 получим объём необходимой выборки N = 390.

3. О представительности выборки

Рассмотрим, как выделяются границы областей $R_i^{(1)}$ и $R_i^{(2)}$ в случае одномерного и многомерного пространства. Вычислим для этого так называемые толерантные (допустимые) пределы.

В о д н о м е р н о м п р о с т р а н с т в е. Пусть задана желаемая надежность правильной классификации ρ . Это значит, что необходимо выделить область $\bar{R}_i$ в пространстве $\{R\}$, попадание в которую при единичном испытании не меньше ρ (рис.I) причем достоверность выделения области должна быть не меньше γ , то есть

$$\mathcal{P}\left\{ \int_{\bar{R}_i} p_i(x)dx \geq \rho \right\} \geq \gamma ;$$

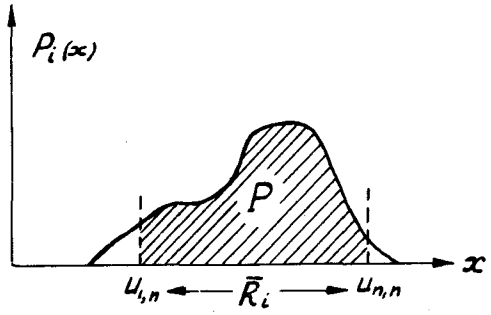

Рис. I. Область $\bar{R}_i$ в одномерном пространстве.

Тогда утверждается [5], что можно подобрать такое N, что крайние члены вариационного ряда, то есть наименьшего $u_{1,n}$ и наибольшего $u_{n,n}$ из выборки объема N, могут быть приняты в качестве границ области $\bar{R}_i$:

$$u_1(x_1, x_2, \ldots, x_n) = u_{1,n},$$

$$u_2(x_1, x_2, \ldots, x_n) = u_{n,n}.$$

При этом необходимый объем выборки N определяется из выражения:

$$N\rho^{N-1} - (N-1)\rho^N = 1 - \gamma. \qquad (9)$$

Например, для $\gamma = 0{,}95$ и $\rho = 0{,}99$ можно найти значение $N \approx 130$. Таким образом, имея выборку объема 130, взятую из произвольной совокупности с любым законом распределения вероятностей, можно утверждать, с достоверностью 0,95, что между крайними членами её лежит не менее 0,99 всей совокупности.

Следовательно, для того, чтобы получить границы области $\bar{R}_i$, нужно выбрать наибольшее и наименьшее значения из предъявленных реализаций выборки, причем объем выборки зависит от заданной величины точности и достоверности и не зависит от вида функции распределения вероятностей образов.

Вычислив границы областей $\bar{R}_i$, нетрудно получить и границы областей $R_i^{(1)}$ и $R_i^{(2)}$

В многомерном пространстве.

для выделения области $\bar{R}_i$ в случае многомерного пространства построим выпуклую оболочку, натянутую на крайние точки из реализаций выборки объёма N.

Проведем нелинейное преобразование пространства $x_k \Rightarrow y_k$, так, чтобы выпуклая оболочка области $\bar{R}_i$ отобразилась в n-мерный параллелепипед $\bar{D}_i$, грани которого перпендикулярны осям координат (рис.2).

$$y_k = a_k(\varphi)\, x_k .$$

Наложим дополнительное ограничение - требование неизменности объема области при преобразовании, выражающееся в условии:

$$\prod_{k=1}^{n} a_k(\varphi) = 1 .$$

При этом производная от $P_i(x)$ будет иметь конечное число разрывов. Следовательно,

$$\int_{\bar{R}_i} P_i(x)\,dx = \int_{\bar{D}_i} P_i(y)\,dy \qquad (10)$$

и оба интеграла существуют. Тогда можно сформулировать следующую теорему:

Теорема. Для любых наперед заданных величин P и γ, удовлетворяющих условиям $0 < P < 1$ и $0 < \gamma < 1$, и независимых параметров y_k, можно указать такой достаточный объем выборки N из генеральной совокупности $\mathfrak{X}$, что n-мерный параллелепипед, натянутый на крайние точки из реализаций этой выборки, ограничит область, вероятность попадания в которую при единичном испытании не меньше P с достоверностью не меньше γ, то есть

$$\mathcal{P}\Big\{ \int_{\bar{D}_i} P_i(y)\,dy \geq P \Big\} \geq \gamma . \qquad (11)$$

Доказательство.

Пусть $u_1^{(k)}$ и $u_N^{(k)}$ минимальная и максимальная точки выборки по k-той оси соответственно. Грани параллелепипеда $\bar{D}_i$ проходят через эти точки перпендикулярно к этим осям.

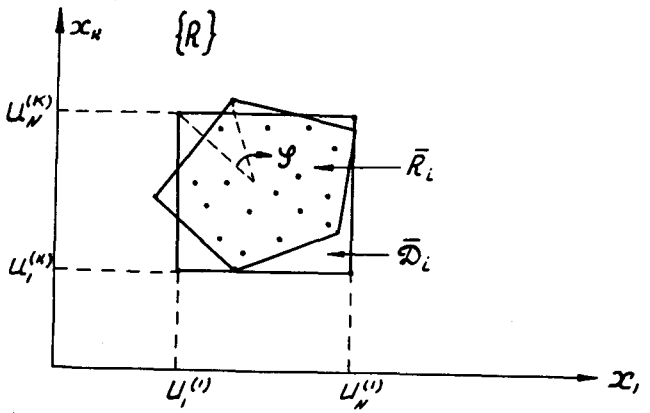

Рис. 2. Проекции областей $\bar{R}_i$ и $\bar{D}_i$ на плоскости $x_K O x_1$.

Тогда

$$\mathcal{P}\Big(\int_{\bar{D}_i} P_i(y)\,dy \geq P\Big) =$$

$$= \mathcal{P}\Big\{\prod_{K=1}^{n}\Big[F_i(u_N^{(K)}) - F_i(u_1^{(K)})\Big] \geq P\Big\} =$$

$$= \mathcal{P}(\tau_{1,N}\cdot\tau_{2,N},\ldots,\tau_{n,N} \geq P), \qquad (12)$$

где

$$F_i(u^{(i)}) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{u^{(K)}}\cdots\int_{-\infty}^{\infty} P_i(y)\,dy =$$

$$= \mathcal{F}(\infty,\infty,\ldots,u^{(K)},\ldots,\infty),$$

$$\tau_{K,N} = F_i(u_N^{(K)}) - F_i(u_1^{(K)}).$$

Известно, [5] что случайная величина $z = F(y)$ распределена по равномерному закону независимо от дифференциальной функции распределения $P_i(y)$, а разность

$$\tau_N = F(u_N) - F(u_1)$$

имеет интегральное распределение:

$$\mathcal{P}(\tau_{1,N} < t) = N\int_0^1 \big[F(u+t) - F(u)\big]^{N-1} du = N(1-t)t^{N-1} + t^N.$$

$$(0 \leq t \leq 1)$$

Отсюда плотность вероятности случайной величина $\tau_N^{(K)}$ имеет вид:

$$W(\tau_{K,N}) = N(N-1)(1-\tau_{K,N})\tau_{K,N}^{N-2}.$$

$$\tau_{K,N} \in [0,1].$$

Тогда

$$\mathcal{P}(\tau_1\cdot\ldots\cdot\tau_n \geq P) = \int_{\tau_{1,N}}\cdot\int_{\tau_{2,N}}\cdots\int_{\tau_{n,N}} W(\tau_1)\cdot$$

$$\cdot W(\tau_n)\,d\tau_1\ldots d\tau_n =$$

$$= \int_P^1 W(\tau_n)\,d\tau_n \int_{\frac{P}{\tau_n}}^1 W(\tau_{n-1})\,d\tau_{n-1}\cdots\int_{\frac{P}{\tau_n\cdot\tau_{n-1}\ldots\tau_2}} W(\tau_1)\,d\tau_1 =$$

$$= \psi(N,n,P). \qquad (13)$$

Приравнивая (13) к заданной достоверности γ, находим:

$$\psi(N,n,P) = \gamma; \qquad (14)$$

$$N = \varphi(n,P,\gamma).$$

Следовательно, обеспечив объем выборки N из (I4), мы получим

$$P\{(\prod_{k=1}^{n}[F_i(u_N^{(k)})-F_i(u_i^{(k)})])\geq P\}\geq\gamma.$$

Тогда в силу (I2) выражение (II) справедливо и теорема доказана.

З а м е ч а н и е

В случае зависимых y_k в большинстве практических задач распознавания величина интеграла в (I2) не может быть меньше. Это даёт основание полагать, что теорема будет справедлива и в случае зависимых параметров, хотя строгого доказательства этого утверждения провести не удалось.

На основании теоремы дадим определение представительности выборки. Выборку будем считать п р е д с т а в и т е л ь н о й, если её объем не меньше N, где N определяется выражением (I4).

Итак, имея представительную выборку, можно получить с заданной достоверностью границы областей $R_i^{(1)}$ и $R_i^{(2)}$, внутри которых заключена доля вероятности не меньше P. Это означает, что при дальнейших испытаниях все реализации из генеральной совокупности π_i будут с вероятностью P попадать внутрь выделенных областей. Если же специфика задачи такова, что объем выборки заранее ограничен, то можно указать достоверность результатов решения задачи распознавания. Зафиксировав область перекрытий $R_i^{(2)}$, мы можем вычислить информативность пространства $\{R\}$ (по той же выборке объема N). Продолжать решение задачи имеет смысл тогда, когда информативность пространства нас удовлетворяет; в противном случае следует искать другое, более информативное пространство признаков.

Рассмотрим вопрос, о получении достаточных оценок $P_i(x)$ по конечной выборке в области перекрытий $R_i^{(2)}$. Можно получить гистограмму $\bar{P}_i(x)$ в этой области вычислением величины $\frac{l_2\xi}{N}$, где $l_2\xi$ - число точек, попавших в элементарный гиперобъем; но для этого нужно иметь большой объем памяти и большой объем выборки, причем вопрос о достаточном объеме выборки опять остается нерешенным.

Как уже говорилось, вероятность попадания реализации из генеральной совокупности π_i в область перекрытия $R_i^{(2)}$ определяется выражением:

$$q_i\int_{R_i^{(2)}}P_i(x)dx.$$

Величину интеграла можно вычислить с заданной точностью и достоверностью, имея конечную выборку объема N. При условии достаточной информативности пространства области перекрытий, как правило, будут содержать "хвосты" функций распределения плотностей вероятности образов. Эти "хвосты" независимо от вида всей функции, могут быть достаточно хорошо аппроксимированы функцией κx^t. Тогда, положив $P_i(x)$ на границе области $R_i^{(1)}$ равной нулю и зная величину интеграла в (2), получим значение κ из уравнения:

$$\kappa\int_{R_i^{(2)}}x^t dx=c,$$

где c - значение интеграла в (2), вычисленного методом статистических испытаний. При этом величина средней квадратической ошибки отклонения $\bar{P}_i(x)$ от $P_i(x)$ будет не больше ε. Тогда значение $\bar{P}_i(x)$ в области $R_i^{(2)}$ получим из уравнения

$$\bar{P}_i(x)=\kappa x^t \qquad (I6)$$

(величина t может варьироваться).

Таким образом, в процессе обучения мы получили оценку информативности выбранного пространства параметров $\{R\}$, границы областей уверенных ответов $R_i^{(1)}$, границы областей перекрытий $R_i^{(2)}$ и приближение с заданной точностью $\bar{P}_i(x)$ в области $R_i^{(2)}$ к функции плотности распределения вероятностей образов $P_i(x)$.

В процессе распознавания в зависимости от специфики задачи могут быть выбраны различные стратегии: стратегия Бейеса, метод последовательного анализа Вальда, стратегия неуверенных ответов. Рассмотрим подробнее каждую из них.

С т р а т е г и я Б е й е с а. Предъявляя системе неизвестную реализацию для распознавания (т.е. задавая значения x_ξ в пространстве $\{R\}$, определяем, в какую из возможных областей $R_i^{(1)}$ или $R_i^{(2)}$ она попадает. Если точка попадет в область $R_i^{(1)}$, то принимается решение о принадлежности предъявленной реализации к генеральной совокупности π_i, если же она по-

падет в область $R_i^{(2)}$, то по (I6) вычисляется значение $\bar{P}_i(x)$ в этой точке для всех i и выбирается тот индекс i, для которого $q_i \bar{P}_i(x)$ максимально. При попадании точки в об - ласть $R^{(3)}$ (область, где $P_i(x) \equiv 0$) принимается гипотеза π_0. При этом математическое ожидание потерь будет:

$$\sum_{i=1}^{m} q_i \sum_{\substack{j=1 \\ j \neq i}}^{m} \int_{R_j^{(2)}} P_i(x)dx + \sum_{i=1}^{m} (1-P_i), \qquad (I7)$$

где P_i - доля вероятности, заключенная в области $\bar{R}_i$. Интеграл в (I7) вычисляется методом статистических испытаний в процессе обучения по конечной выборке объема N.

Для сокращения времени при распознавании определение принадлежности точки к той или иной области делается последовательно за n шагов методом дихатомического деления пространства [9].

Метод последовательного анализа Вальда. Системе предъявляется неизвестная реализация, определяется, какой из возможных областей $R_i^{(1)}$ или $R_i^{(2)}$ принадлежит точка. Если точка попадает в область $R_i^{(1)}$, то принимается гипотеза о принадлежности реализации к генеральной совокупности π_i, если же точка попадает в область неуверенных ответов, то вычисляется отношение правдоподобия

$$\lambda_{\kappa i,j} = \frac{\bar{P}_{i\kappa}(x)}{\bar{P}_{j\kappa}(x)} \qquad (I8)$$

для тех индексов i и j, для которых $\bar{P}_{i(j)}(x)$ в найденной области отличны от нуля.

Если $\lambda_{\kappa ij} \geq A$, то принимается гипотеза о принадлежности реализации к π_i; если $\lambda_{\kappa ij} \leq B$, то реализация относится к π_j, и если $B_{ij} < \lambda_{\kappa ij} < A_{ij}$, то принимается решение о повторном испытании (переспросе) для вычисления следующего значения $\lambda_{(\kappa+1) ij}$. Переспрос продолжается до тех пор, пока не будет принята гипотеза о принадлежности x_κ к тому или иному образу. Константы A_{ij} и B_{ij} ($A > B$) связаны с вероятностями ошибок первого и второго рода α и β следующим соотношением:

$$A \leq \frac{1-\beta}{\alpha}; \qquad B \geq \frac{\beta}{1-\alpha}. \qquad (I9)$$

Для независимых и имеющих одну и ту же функцию распределения наблюдений $\bar{P}_{i\kappa}$ определяется как:

$$\bar{P}_{i\kappa} = \prod_{n=1}^{\kappa} \bar{P}_i(x_n) \quad (n=1,2,\ldots,\kappa).$$

Значения $\bar{P}_i(x_n)$ вычисляются из (I6).

Вальд и Вольфовитц [7] показали, что метод последовательного анализа оптимален в смысле минимизации среднего риска, получаемого в случае единичного испытания; кроме того, для заданных α и β, он минимизирует среднее число переспросов. Следует заметить, что в том случае, когда на каждом κ-ом шаге точка x_κ будет попадать в область равных вероятностей конкурирующих гипотез, метод Вальда не будет сходиться. Поэтому при решении практических задач распознавания образов имеет смысл применить так называемый усеченный метод последовательных испытаний. Усечение заключается в том, что заранее назначается максимально допустимое число переспросов n. Последовательные испытания проводятся до тех пор, пока не будет принято решение или не будет проверено n-ое наблюдение. Если при n-ом наблюдении решение еще не принято, то переходят либо к стратегии Бейеса, либо к стратегии неуверенных ответов, принимая за измерение реализации значение λ_n.

Стратегия неуверенных ответов. В отличие от вышеописанных стратегий, при попадании точки x_ξ в область неуверенных ответов вычисляются значения $\bar{P}_i(x)$ по (I6), и система выдает те индексы i, со значениями их вероятностей в этой точке, для которых $\bar{P}_i(x) > 0$, [8].

Основные выводы

I) Задача распознавания образов по своей сущности является статистической и может быть решена статистическими методами независимо от того, априори известны или нет функции распределения образов. Достоверность результатов решения задачи зависит от объема выборки обучающей последовательности.

2) Задача распознавания решается в несколько этапов: задаются алфавит распознаваемых образов, желаемая надежность и достоверность решения задачи, выбирается ориентировочно пространство параметров и оценивается информативность этого пространства в процессе обучения; если информативность пространства не удовлетворяет поставленным требованиям, то ведется поиск но-

вого пространства до тех пор, пока не будут удовлетворены эти требования (вопрос о выборе пространства параметров является самостоятельным и здесь не рассматривается).

3) Алгоритм решения задачи заключается в следующем: задаются : алфавит распознаваемых образов $1,2,\ldots,m$; пространство **измеряемых** параметров $1,2,\ldots,n$; число градаций **по каждому из параметров** $l_1, l_2, \ldots, l_n$, доля вероятности P_i , заключенная в области R_i , и достоверность выделения границы области γ . Вычисляется по (I4) необходимый объем выборки N_i . В процессе обучения для всех m образов строятся границы областей $\bar{R}_i$. Определяются границы областей $R_i^{(1)}$ и $R_i^{(2)}$ и вычисляется информативность пространства по (2) В зависимости от специфики задачи выбирается одна из возможных стратегий: стратегия Бейеса, метод последовательного анализа Вальда или стратегия неуверенных ответов. Если выбрана стратегия Вальда, то дополнительно задаются значения α_{ij} и β_{ij} и тогда по (I9) вычисляются значения A_{ij} и B_{ij} . Если выбрана стратегия Бейеса, то вычисляется по (I7) математическое ожидание потерь.

После этого процесс обучения считается временно законченным и система готова к непосредственному распознаванию образов. При предъявлении системе неизвестной реализации определяется, к какой из областей $R_i^{(1)}$ или $R_i^{(2)}$ она относится, и выдается i -ый индекс, если реализация попадает в область $R_i^{(1)}$. Если реализация попадает в область $R_i^{(2)}$, то либо вычисляется i-ый индекс по (I) (в зависимости от выбранной стратегии), либо принимается решение согласно (I8), либо выдаются все конкурирующие гипотезы с их значениями вероятностей в этой области.

4) В процессе работы системы возможно её совершенствование (подучивание), при этом изменяются границы областей $R_i^{(1)}$ и $R_i^{(2)}$ и повышается достоверность результатов распознавания.

5) Задачу распознавания образов в общем случае нужно рассматривать как "многоэтажную", иерархическую систему. На первом этапе распозаются образы в пространстве **физических**, химических или каких-либо других свойств предметов и явлений; на следующем этапе распознаются образы в пространстве признаков, которые состоят из алфавита распознанных образов и т.д.

Примером такой иерархической системы может служить задача распознавания речи. На первом этапе проводится членение непрерывного речевого потока на фонемы; затем распознаются фонемы на выделенных участках, далее в пространстве фонем распознаются слова и т.д.

В статье рассмотрена элементарная система для решения задачи распознавания образов, элементарная в том смысле, что она может служить элементом построения иерархической системы любой сложности.

Автор считает своим приятным долгом выразить глубокую благодарность Н.Г. Загоруйко, Л.Я. Савельеву, Э.Х. Гимадутдинову, Г.Я. Волошину и Г.С. Лбову, оказавшим помощь в настоящей работе.

Л И Т Е Р А Т У Р А

I. Т. Андерсон. Введение в многомерный статистический анализ. М., ФМ, I963, стр. I75-2II.

2. К. ФУ. Модель последовательных решений для оптимального опознавания. - Проблемы бионики. М.,"Мир" I965., стр. 374-384.

3. В.А. Ковалевский. Задача распознавания образов с точки зрения математической статистики. - Семинар "Распознавание образов и конструирование читающих автоматов". Киев, I965.

4. Н.П. Бусленко, Ю.А. Шрейдер. Метод статистических испытаний. М., ФМ, I96I, стр. 66-72.

5. И.В. Дунин-Барковский, Н.В. Смирнов. Теория вероятностей и математическая статистика в технике. М., Гостехиздат, I955.

6. В.Н. Елкина, Н.Г. Загоруйко. Алгоритм поиска формальных элементов алфавита. (Данный сборник, стр.59).

7. Wald A.,Wolfowitz J., Optimum character of the sequential probability ratio test. Ann.Math.Stat., I9, 326 - 339, I948.

8. Streck G.P. Stochastic Model for the Browning-Bledsoe Pattern Recognition Scheme. IRE Trans. on Electr. Comp. 1962, EC-11, pp.274-282.

9. Forgie J.W. and Forgie C.D. Computer Identification of Vowel Types. GASA, vol.33, 1, January 1961, p.7-11.

Поступила в редакцию
II.У. I966.