ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Сборник трудов
Института математики СО АН СССР

1966 г. Выпуск 22

О ПРЕДСТАВИТЕЛЬНОСТИ ВЫБОРКИ ПРИ ВЫБОРЕ ЭФФЕКТИВНОЙ СИСТЕМЫ ПРИЗНАКОВ

Г.С. Лбов

§ I. Постановка задачи

Как показано в работе [I] , эффект, вызываемый ограниченностью экспериментального материала, может оказать весьма существенное влияние на точность определения вероятности неправильного распознавания, что, в свою очередь, значительно усложняет выбор эффективной системы признаков [2] .

В настоящей работе даны некоторые рекомендации по выбору эффективной системы признаков в условиях ограниченной выборки. В работе также приводятся результаты эксперимента на ЭВМ по установлению достаточного объема выборки для определения вероятности неправильной классификации с заданной точностью (случай нормального распределения).

Рассмотрим один из возможных подходов к решению задачи выбора эффективной системы признаков в случае неизвестных законов распределения. Делаем произвольное предположение о типе распределения генеральных совокупностей $P_1'(x),\dots,P_\kappa'(x)$, соответствующих κ образам. Здесь через x обозначен вектор в пространстве признаков, на которых строится рассматриваемое про-

странство. Разбиваем пространство признаков на непересекающиеся области $Y'_1,\dots,Y'_K$ (по числу образов). Область Y'_i определим как совокупность тех точек x, для которых

$$q_i P'_i(x)=\sup_j \{q_j P'_j(x)\}_{j=1,\dots,K}. \qquad (1)$$

Положим, что $P'_1(x),\dots,P'_K(x)$ – нормальные распределения. Это означает, что мы будем использовать решающие функции либо из класса линейных, либо квадратичных функций. Выборку, на основе которой оцениваем параметры распределений $P'_1(x),\dots,P'_K(x)$, назовем обучающей выборкой. При этом, очевидно, решающее правило (I) будет отличаться от оптимального решающего правила, во-первых, из-за возможно неправильного предположения о типе закона распределения и, во-вторых, из-за использования оценок параметров распределения по ограниченной обучающей выборке вместо самих параметров распределения.

Обозначим через $\mathcal{P}$ вероятность неправильного распознавания, получаемую при условии оптимального разбиения пространства [6].

В качестве оценки вероятности неправильной классификации можем использовать либо функционал $\mathcal{P}^*$ [2], либо, если существует возможность получить контрольную выборку, величину $\mathcal{P}'$, определяемую следующим образом:

$$\mathcal{P}'=\sum_{i=1}^{K} q_i \sum_{\substack{j=1\\ j\neq i}}^{K} \mathcal{P}'(j/i)=\sum_{i=1}^{K} q_i \mathcal{P}'_i, \qquad (2)$$

где

$$\mathcal{P}'(j/i)=\int_{Y'_j} P_i(x)\,dx. \qquad (3)$$

Заметим, что под интегралом используется истинное распределение $P_i(x)$, а не $P'_i(x)$. Другими словами, значение интеграла (3) оцениваем по контрольной выборке после того, как определены области $Y'_1,\dots,Y'_K$ согласно решающему правилу (I). Априорные вероятности $q_1,\dots,q_K$ будем считать известными. Замечания о необходимом объеме контрольной выборки будут приведены ниже.

Обозначим через N_i число реализаций, составляющих обучающую выборку для i – того образа. Если увеличивать объем

обучающей выборки ($N_1\to\infty,\dots,N_K\to\infty$), то величина $\mathcal{P}'$ будет стремиться по вероятности к некоторому значению $\mathcal{P}_1$, вообще говоря, отличному от $\mathcal{P}$. Равенство величин $\mathcal{P}_1$ и $\mathcal{P}$ достигается лишь при условии правильного предположения о нормальном законе распределения. Величина отклонения $\mathcal{P}_1$ от $\mathcal{P}$ зависит от того, насколько хорошо неизвестные распределения $P_1(x),\dots,P_K(x)$ аппроксимируются распределениями $P'_1(x),\dots,P'_K(x)$. Решающие функции, при которых вероятность неправильной классификации равна $\mathcal{P}_1$, будут наилучшими, дающими минимальную вероятность ошибки для выбранного класса решающих функций. Заметим, что при ограниченном объёме обучающей выборки будет всегда выполняться соотношение $\mathcal{P}'\geq\mathcal{P}_1$. Величина $\mathcal{P}'$ является случайной величиной, так как представляет собой функцию случайных наблюдений, составляющих обучающую выборку. Обозначим плотность распределения $\mathcal{P}'$ через $\varphi(\mathcal{P}')$.

Поясним изложенное с помощью рис.I.

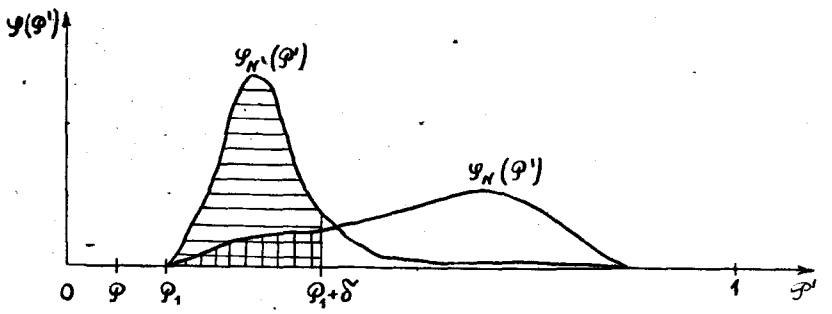

Рис. I. Плотности распределения величины $\mathcal{P}'$ для двух различных объемов обучающей выборки.

Пусть $\varphi_N(\mathcal{P}')$ отображает плотность распределения $\mathcal{P}'$ при некотором фиксированном объёме выборки $N_1=N_2=\dots=N_K=N$. Функция $\varphi_{N'}(\mathcal{P}')$ отображает плотность распределения $\mathcal{P}'$ при объеме выборки $N'_1=N'_2=\dots=N'_K=N'$, причем $N'>N$. Здесь δ – погрешность в определении $\mathcal{P}_1$, вызванная ограниченностью обучающей выборки.

При увеличении объёма обучающей выборки величина $\mathcal{P}'$ стремится к $\mathcal{P}_1$ и увеличивается вероятность получения неравенства $\mathcal{P}'<\mathcal{P}_1+\delta$ при фиксированном значении δ (см. заштрихованные области на рис.I для плотностей $\varphi_N(\mathcal{P}')$ и $\varphi_{N'}(\mathcal{P}')$). Кроме того, оказывается, что при фиксированном объёме обучающей выборки и при данном значении $\mathcal{P}_1$ вероятность равного отклонения $\mathcal{P}'$

от $\mathcal{P}_1$ тем больше, чем больше размерность пространства (§3). На практике часто встречается случай, когда при высокой размерности исходной системы признаков имеем ограниченный по тем или иным причинам объём выборки. Это может привести к существенному отклонению величины $\mathcal{P}'$ от $\mathcal{P}_1$. В этих условиях мы можем либо "забраковать" систему признаков, в то время как для этой системы $\mathcal{P}_1 < \mathcal{P}_0$ ($\mathcal{P}_0$ – допустимая вероятность неправильной классификации) , либо при сравнении двух систем X и $У$ получить $\mathcal{P}'_X > \mathcal{P}'_У$, тогда как на самом деле $\mathcal{P}_X < \mathcal{P}_У$. Решение задачи выбора эффективной системы признаков в этих условиях значительно усложняется. Для её решения потребуется понятие представительности выборки.

Для фиксированной системы признаков X и данного комплекса условий $\mathcal{D}$ получения реализаций признаков назовём обучающую выборку представительной, если выполняется следующее неравенство:

$$P_r\{(\mathcal{P}'-\mathcal{P}_1)<\delta_0\} \geqslant \beta , \qquad (4)$$

где δ_0 – допустимая погрешность в определении вероятности ошибки $\mathcal{P}_1$ (значение β близко к единице).

Величину $\mathcal{P}'$ (см.выражение (2)) определяем по контрольной выборке. Объём контрольной выборки [3] для определения $\mathcal{P}'$ с погрешностью ε задаётся выражением:

$$N_0 = \frac{\sum_{i=1}^{K} q_i \mathcal{P}'_i (1-\mathcal{P}'_i)}{\varepsilon^2} t_\alpha^2 , \qquad (5)$$

где t_α – величина критического интервала, которая выбирается из таблиц нормального распределения по заданной достоверности α . При этом число контрольных реализаций N_{0i} , принадлежащих i –тому образу, равно $q_i N_0$. Получаем

$$P_r\{|\bar{\mathcal{P}}'-\mathcal{P}'|<\varepsilon\} \geqslant \alpha , \qquad (6)$$

где $\bar{\mathcal{P}}'$ – оценка $\mathcal{P}'$ по ограниченной контрольной выборке объёма N_0. Так как вероятности $\mathcal{P}'_1,\dots,\mathcal{P}'_K$ неизвестны, то для определения N_0 поступают следующим образом: задаются либо максимальным значением N_0 , которое получается при $\mathcal{P}'_1=\mathcal{P}'_2,\dots,\mathcal{P}'_K = 0{,}5$, либо задаются некоторым первоначальным значением N_0 , а затем, найдя оценки $\bar{\mathcal{P}}'_1,\dots,\bar{\mathcal{P}}'_K$, уточняют значение

числа контрольных реализаций. При этом

$$\bar{\mathcal{P}}'_i = \frac{\bar{N}_{oi}}{N_{oi}} ,$$

где $\bar{N}_{oi}$ – число неправильно распознанных реализаций i –того образа.

В следующем параграфе рассмотрим решение задачи выбора эффективной системы признаков в условиях ограниченной обучающей выборки.

§ 2. Выбор эффективной системы признаков.

Пусть задана некоторая исходная система признаков $x_1,\dots,x_n$. Строим на указанных признаках подпространства m –мерной размерности. При каждом $m (m=1,2,\dots,n)$ будем рассматривать только самое информативное подпространство *) из всех m –мерных подпространств, число которых равно C_n^m . Число всех различных подпространств равно $2^n - 1$.

Допустим, что при каждом значении m для самого информативного подпространства известны вероятности $\mathcal{P}_1$ и $\mathcal{P}_1+\delta$, причем величина δ выбрана такой, что

$$P_r\{(\mathcal{P}'-\mathcal{P}_1)<\delta\} \geqslant \beta . \qquad (7)$$

Обозначим $\mathcal{P}_1+\delta$ через $\mathcal{P}_\delta$. На рис. 2 функции $\mathcal{P}_1(m)$ и $\mathcal{P}_\delta(m)$ заданы в точках, но для наглядности их значения для разных m соединены сплошными линиями.

Через $\mathcal{P}(0)$ обозначена вероятность ошибки , получаемая при использовании лишь априорных сведений об образах. Величина $\mathcal{P}(0)$ равна $1-q_i$, если $q_i > q_j$ $(j=1,\dots,K;\ j\neq i)$.

Функции $\mathcal{P}_1(m)$ и $\mathcal{P}_\delta(m)$ для каждого m задают интервал, внутри которого с большой вероятностью β окажется выборочное значение $\mathcal{P}(m)$ при данном объёме обучающей выборки. Такой интервал указан только для самого информативного подпространства при каждом значении m . Для других подпространств m –мерной размерности величина этого интервала тем больше, чем больше вероятность ошибочной классификации $\mathcal{P}$ (§3). Если, однако, значение вероятности $\mathcal{P}$ приближается к $\mathcal{P}(0)$, то величина интервала δ с некоторого момента будет меньше допусти-

*) Из двух подпространств будем считать более информативным то подпространство, использование которого приводит к меньшей вероятности ошибки.

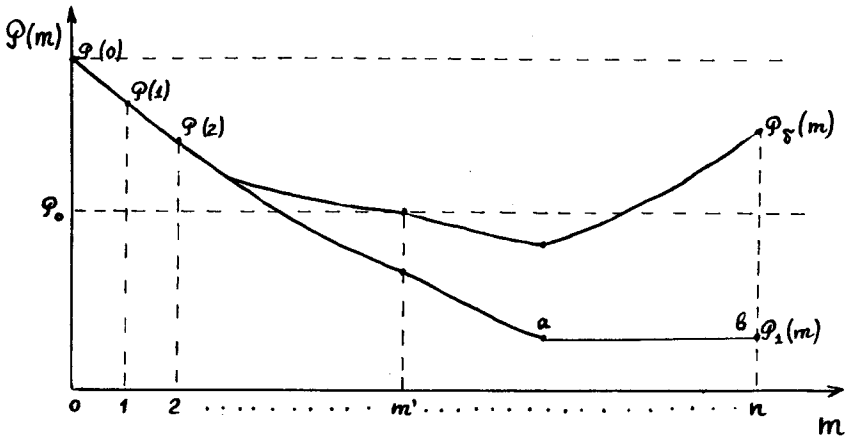

Рис. 2. Доверительный интервал величины $\mathcal{P}'$ для самых информативных подпространств m -мерной размерности.

мого значения δ_0 . Для такого подпространства в соответствии с (4) выборка любого объёма будет представительной. Это объясняется тем, что при $\mathcal{P} = \mathcal{P}(0)$ (пространство используемых признаков не несёт информации) вероятность ошибки не меняется от способа разбиения пространства на области $y'_1, \ldots, y'_K$.

Если функция $\mathcal{P}_1(m)$ остается постоянной при увеличении m (область ab рис.2), то функция $\mathcal{P}_\delta(m)$ будет возрастающей. Это может привести к тому, что в некоторых конкретных задачах использование всей системы признаков при фиксированном объёме обучающей выборки может привести к вероятности неправильной классификации большей , чем вероятность неправильной классификации в случае использования лишь части информативных признаков этой системы. При данном объёме выборки N найдется такое m' , что для всех подпространств m -мерной размерности $(m \leqslant m')$ величина интервала δ не превзойдёт некоторой наперёд заданной δ_0 . Тогда как для $m > m'$ величина интервала δ может быть больше δ_0 . В этом случае выбор эффективного m -мерного подпространства для $m > m'$ из множества m -мерных подпространств может быть неудачным: для $m > m'$ невозможно установить, насколько предпочтительнее одно подпространство перед другим, так как разница между информативностями, получаемыми при использовании выбранных двух подпространств, может оказаться меньше, чем погрешность при определении самих информативностей этих подпространств. В описанных условиях для решения задачи выбора эффективной системы признаков можно предложить следующий подход. Сначала выбираем наиболее информативное подпространство m'- мерной размерности. Если окажется, что для него $\mathcal{P}'(m') \leqslant \mathcal{P}_0$, то задача может считаться решенной. Если $\mathcal{P}'(m') > \mathcal{P}_0$, то можно увеличивать число признаков исходной системы каждый раз на некоторую величину, оставляя постоянным число реализаций $(N_1, \ldots, N_K)$. При каждом таком увеличении выбирать наиболее информативное подпространство m'- мерной размерности. Стремление к увеличению размерности исходного пространства может быть оправдано тем, что при увеличении n увеличивается общее число m'-мерных подпространств, из которых выбирается самое информативное подпространство. Это приводит к увеличению вероятности получения такого подпространства, при котором достигается решение задачи. Однако может получиться так, что все известные признаки исчерпаны, но не оказалось такого m'-мерного подпространства, при котором получаем решение задачи, то есть $\mathcal{P}'(m') \leqslant \mathcal{P}_0$. Поэтому, по-видимому, лучше всего, увеличивая n , одновременно увеличивать число реализаций ($N_1, \ldots, N_K$). Задача может быть не решена только в случае, когда исчерпаны все известные признаки и m' стало равным n , но при этом $\mathcal{P}'(n) > \mathcal{P}_0$.

Можно, конечно, с самого начала составить исходную систему из всех известных признаков и обеспечить представительность обучающей выборки для выбранной системы признаков. Однако такой подход менее приемлем, поскольку решение задачи при вышеописанном подходе может быть достигнуто раньше , чем выборка станет представительной для исходной системы из всех известных признаков.

При предлагаемом подходе необходимо использовать алгоритм сокращения числа признаков с n до m', минуя промежуточные значения m $(m > m')$. В описанных условиях алгоритм, предлагаемый в работе [4] , применять нежелательно, так как этот алгоритм предполагает на первом этапе выбор наиболее информативного подпространства среди $(n-1)$ - мерных подпространств, на втором этапе - среди $(n-2)$ - мерных подпространств и т. д. В этом случае можно использовать алгоритм случайного поиска с адаптацией, приведенной в работе [2].

§ 3. Представительность выборки из нормального распределения (случай дихотомии)

В настоящее время при построении решающего правила по выборке руководствуются следующим принципом: чем больше объем обучающей выборки, тем лучше. Такой подход дает лишь интуитивную уверенность в том, что объем используемой выборки достаточен для оценки вероятности неправильной классификации.

Задача об установлении представительности выборки заданного объема достаточно сложная.

В работе [5] в качестве меры представительности выборки введена величина $\Delta = M\mathcal{P}' - \mathcal{P}$ ($M\mathcal{P}'$ – математическое ожидание случайной величины $\mathcal{P}'$) и установлена зависимость величины Δ от объема выборки для одномерного нормального распределения с известными параметрами. Попытка выяснить эту зависимость для многомерного распределения приводит к интегралу, который не может быть выраженным через элементарные функции.

В случае неизвестного распределения можно предложить следующий способ определения представительности выборки заданного объема N ($N_1 = N_2 = \ldots = N_k = N$). Получаем выборку объема, достаточного для того, чтобы образовать из неё τ различных групп по N реализаций в каждой группе.

Так как значение $\mathcal{P}_1$ неизвестно, мы не можем воспользоваться выражением (4) для определения представительности выборки. В этом случае поступим следующим образом. Каждую группу реализаций рассматриваем как отдельную обучающую выборку. С помощью выражения (2) вычисляем вероятность $\mathcal{P}'$ для каждой группы ($\mathcal{P}'_1, \ldots, \mathcal{P}'_\tau$) и определяем из $\mathcal{P}'_1, \ldots, \mathcal{P}'_\tau$ минимальное $\mathcal{P}'_{min}$ и максимальное $\mathcal{P}'_{max}$ значения, которые образуют некоторый интервал ($\mathcal{P}'_{min}, \mathcal{P}'_{max}$). Число τ должно быть таким, чтобы

$$P_\tau\{\mathcal{P}' \in (\mathcal{P}'_{min}, \mathcal{P}'_{max})\} \geq \beta. \qquad (8)$$

В данном случае в качестве меры представительности выборки можно выбрать величину $\Delta\mathcal{P}' = \mathcal{P}'_{max} - \mathcal{P}'_{min}$. Если при этом величина интервала $\Delta\mathcal{P}'$ меньше некоторой допустимой величины $\Delta\mathcal{P}_0'$, то выборку будем считать представительной. Выбор $\Delta\mathcal{P}'$ в качестве меры представительности выборки может быть оправдан тем, что при увеличении объёма обучающей выборки N выборочные значения $\mathcal{P}'$ все теснее группируются около $\mathcal{P}_1$ и величина интервала $\Delta\mathcal{P}'$ по вероятности стремится к нулю при фиксированном значении τ.

Был поставлен следующий эксперимент для определения представительности выборки объёма N из многомерных нормальных распределений. С помощью датчика [3] вырабатывались случайные векторы из двух генеральных совокупностей π_1 и π_2 с некоторыми плотностями распределения $P_1(x), P_2(x)$, приближенно описываемыми нормальными законами распределения $N_1(\mu^{(1)}, E)$ и $N_2(\mu^{(2)}, E)$, где

$$\mu^{(1)} = \begin{Bmatrix} 0 \\ \vdots \\ 0 \end{Bmatrix} \quad \text{и} \quad \mu^{(2)} = \begin{Bmatrix} \frac{\rho}{\sqrt{n}} \\ \vdots \\ \frac{\rho}{\sqrt{n}} \end{Bmatrix}.$$

Величина ρ – евклидово расстояние между центрами нормальных распределений; E – единичная матрица ковариации. Дивергенция для нормальных распределений с указанными параметрами

$$(\mu^{(1)} - \mu^{(2)})' E (\mu^{(1)} - \mu^{(2)}) = \rho^2$$

и вероятность неправильной классификации [6]

$$\bar{\mathcal{P}} = \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{2}\rho}^{\infty} e^{-\frac{1}{2}y^2} dy. \qquad (9)$$

Для выяснения зависимости величины интервала $\Delta\mathcal{P}'$ от информативности и размерности пространства, а также объёма обучающей выборки необходимо эту величину определить для некоторых значений $\bar{\mathcal{P}}, n, N$ (в случае многомерных нормальных распределений с указанными параметрами). Величина интервала $\Delta\mathcal{P}'$ была вычислена при следующих значениях $\bar{\mathcal{P}}$: $\bar{\mathcal{P}}_1 = 0{,}0099 \approx 0{,}01$ ($\rho_1 = 4{,}66$, см. таблицу 1 приложения), $\bar{\mathcal{P}}_2 = 0{,}10083 \approx 0{,}1$ ($\rho_2 = 2{,}56$, см. таблицу 2 приложения), $\bar{\mathcal{P}}_3 = 0{,}29811 \approx 0{,}3$ ($\rho_3 = 1{,}06$, см. таблицу 3 приложения), $\bar{\mathcal{P}}_4 = 0{,}44883 \approx 0{,}45$ ($\rho_4 = 0{,}13$, см. таблицу 4 приложения). При каждом значении $\bar{\mathcal{P}}$ величина $\Delta\mathcal{P}'$ определялась для двух -шести- и десятимерного пространств и для объёма выборки $N = 20, 50, 100, 150$. Таким образом, значение $\Delta\mathcal{P}'$ было определено в 48 случаях.

В каждом таком случае находились выборочные значения $\mathcal{P}'_1, \ldots \mathcal{P}'_\tau$

случайной величины $\mathcal{P}'$ ($r=15$). При этом, например, величина $\mathcal{P}_j'$ ($j=1,\dots,15$) определялась так: для данной размерности n пространства признаков случайным образом получали N реализаций из генеральной совокупности π_1 и N реализаций из генеральной совокупности π_2. Определялись оценки $\mu^{(1)}$ и $\mu^{(2)}$

$$\bar{\mu}^{(1)}=\begin{Bmatrix}\bar{x}_1^{(1)}\\ \vdots\\ \bar{x}_n^{(1)}\end{Bmatrix}, \quad \bar{\mu}^{(2)}=\begin{Bmatrix}\bar{x}_1^{(2)}\\ \vdots\\ \bar{x}_n^{(2)}\end{Bmatrix}.$$

После этого считалось, что обучение закончено, и в соответствии с (I) устанавливалось следующее решающее правило: область y_i' ($i=1,2$) определялась как совокупность тех точек x, для которых

$$-2\sum_{p=1}^{n}\bar{x}_p^{(i)}x_p+\sum_{p=1}^{n}(\bar{x}_p^{(i)})^2=\inf_{\ell}\left\{-2\sum_{p=1}^{n}\bar{x}_p^{(\ell)}x_p+\sum_{p=1}^{n}(\bar{x}_p^{(\ell)})^2\right\}_{\ell=1,2}.$$

Для определения $\mathcal{P}_j'$ использовалась контрольная выборка объёма N_{01} = 3000 из генеральной совокупности π_1 и N_{02} = 3000 из π_2. Если положить $q_1=q_2=\frac{1}{2}$, t_α = I,96 и $\mathcal{P}_j'\approx\mathcal{P}$, то точность вычисления $\mathcal{P}_j'$ согласно (5) будет:

$$\varepsilon = 1{,}96\sqrt{\frac{\bar{\mathcal{P}}(1-\bar{\mathcal{P}})}{2\cdot 3000}}.$$

Значения ε и $\bar{\mathcal{P}}$ приводятся в нижеследующей таблице:

$\bar{\mathcal{P}}$	0,0I	0,I	0,3	0,45
ε	0,0002	0,0074	0,0II0	0,0I30

Число r выборочных значений величины $\mathcal{P}'$ задавалось следующим образом. Известно [7], что для произвольного закона распределения достоверность γ получения неравенства (8) и значения β и r связаны соотношением

$$\gamma = 1-r\beta^{r-1}+(r-1)\beta^{r}.$$

Например, для r = I5 и γ = 0,98 получаем β = 0,9.

На основе таблиц I, 2, 3, 4 приложения составлена таблица. Если для данных $\bar{\mathcal{P}}$, n, N величина полученного интервала $\Delta\mathcal{P}'$ превышала некоторую заданную величину $\Delta\mathcal{P}_0$, то считалось, что выборка объёма N непредставительна. Этим значениям $\bar{\mathcal{P}}$, n, N в таблице соответствуют нули. Если при некоторых значениях $\bar{\mathcal{P}}$, n, N величина $\Delta\mathcal{P}'$ меньше $\Delta\mathcal{P}_0$, то этим значениям в таблице соответствуют единицы (выборка объёма N считалась представительной).

Из таблицы видно, что выборка одного и того же объёма становится более представительной при увеличении информативности пространства. Выборка становится менее представительной с увеличением размерности пространства при фиксированных значениях N и $\bar{\mathcal{P}}$. Те случаи, которые не соответствуют этому правилу, обозначены звёздочкой. Это несоответствие, по-видимому, объясняется ограниченностью числа r выборочных значений случайной величины $\mathcal{P}'$.

Полученные результаты можно использовать, когда априори известно, что генеральные совокупности π_1 и π_2 имеют распределения, хорошо аппроксимируемыми нормальными распределениями с единичными матрицами ковариации.

На основе полученных данных можно дать некоторые рекомендации. Если, например, даны допустимая вероятность неправильной классификации $\mathcal{P}_0$ и размерность n пространства, то можно оценить приближенный объем выборки, необходимый для того, чтобы не "забраковать" систему признаков X при $\mathcal{P}_X<\mathcal{P}_0$.

Пусть, например, $\mathcal{P}_0$ = 0,I2 и n = I0. Если при этом $\mathcal{P}_X$ = 0,I0, то необходимо использовать выборку N = 50. Действительно, для N = 50 $Bep\{\mathcal{P}_X'>\mathcal{P}_0\}_{N=50}\approx 0$, в то время как для $N=20$ $Bep\{\mathcal{P}_X'>\mathcal{P}_0\}_{N=20}>0$.

Далее, если для двух m-мерных подпространств X и Y ($m\leqslant 10$) будет соблюдаться неравенство $\mathcal{P}_Y-\mathcal{P}_X>0{,}04$, то при N = 20 $Bep\{\mathcal{P}_X'>\mathcal{P}_Y'\}_{N=20}\approx 0$.

В заключение сделаем ряд следующих замечаний.

I. Величина $\bar{\mathcal{P}}$, определяемая выражением (9), не равна вероятности неправильной классификации $\mathcal{P}_1$, поскольку истинные математические ожидания распределений случайных чисел, вырабатываемых датчиком, вообще говоря, отличны от предполагаемых значений $\mu^{(1)}$ и $\mu^{(2)}$. По этой причине выборочные значения величи-

ны $\mathcal{P}'$ (см, таблицы I, 2, 3, 4 из приложения) могут получиться иногда меньше вероятности $\bar{\mathcal{P}}$.

2. Если предположить, что $\mathcal{P}'_{min} - \mathcal{P}_1 << \Delta\mathcal{P}'$, то интервал $\Delta\mathcal{P}'$ можно считать равным интервалу δ.

3. Если вместо вероятности $\mathcal{P}'$, определяемой по формуле (2), использовать функционал $\mathcal{P}^*$[2], то значение $\mathcal{P}^*$, как правило, будет меньше вероятности неправильной классификации (в то время, как величина $\mathcal{P}'$ всегда больше или равна $\mathcal{P}_1$). желательно было бы определить величину интервала δ для $\mathcal{P}^*$.

4. Если имеем обучающую выборку объёма N и контрольную выборку объёма N_0, то $Bep\{\bar{\mathcal{P}}' > \mathcal{P}_1 + \delta + \varepsilon\} \approx 0$ при значениях α и β, близких к единице (см. выражения (6) и (7)).

5. Проведенный эксперимент можно продолжить для произвольных значений $\mathcal{P}$, n, N и числа образов $\kappa > 2$. Причем при $\kappa = 2$ для определения выборочного значения $\mathcal{P}'_j$ ($j = 1, \ldots, \tau$) можно использовать следующее выражение:

$$\mathcal{P}'_j = q_1 \int_{|\alpha_1|}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy + q_2 \int_{|\alpha_2|}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy,$$

где $|\alpha_1|$ и $|\alpha_2|$ - расстояния от центров многомерных нормальных распределений $\mu^{(1)}$ и $\mu^{(2)}$ до гиперплоскости:

$$(\bar{\mu}^{(1)} - \bar{\mu}^{(2)})' x - \frac{1}{2}(\bar{\mu}^{(1)} - \bar{\mu}^{(2)})'(\bar{\mu}^{(1)} + \bar{\mu}^{(2)}) = 0,$$

где $\bar{\mu}^{(1)}$ и $\bar{\mu}^{(2)}$ - оценки величин $\mu^{(1)}$ и $\mu^{(2)}$, определяемые на основе обучающей выборки объёма N j-ой группы. Такой подход к вычислению $\mathcal{P}'_j$ не связан с получением контрольной выборки, что сокращает машинное время проведения эксперимента.

Автор благодарен доктору ф.м.н. С.В. Нагаеву и к.т.н. Н.Г. Загоруйко за ряд критических замечаний.

ЛИТЕРАТУРА

I. А.Е. Заславский, Н.М. Сычева. Об одной задаче оптимального распознавания образов.–Вычислительные системы. Новосибирск, 1965, вып. 19., стр. 35

Таблица

		$n=2$								$n=6$								$n=10$							
	$\Delta\bar{\mathcal{P}}_0$	$\bar{\mathcal{P}} \approx 0{,}45$			$\bar{\mathcal{P}} \approx 0{,}3$			$\bar{\mathcal{P}} \approx 0{,}1$			$\bar{\mathcal{P}} \approx 0{,}01$			$\bar{\mathcal{P}} \approx 0{,}45$			$\bar{\mathcal{P}} \approx 0{,}3$			$\bar{\mathcal{P}} \approx 0{,}1$			$\bar{\mathcal{P}} \approx 0{,}01$		
		$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$
	0,005	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,01	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,02	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
$n=2$	0,03	0	1*	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,04	1	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,005	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,01	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,02	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
$n=6$	0,03	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,04	1	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,005	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,01	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,02	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
$n=10$	0,03	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
	0,04	1	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0

2. Г.С. Лбов. Выбор эффективной системы зависимых признаков,– Вычислительные системы. Новосибирск, 1965.вып.19, стр. 21.

3. Г.С. Лбов. Алгоритмы и программы для распознавания образов (диагностики заболеваний сердца по баллистокардиограммам). Отчет ИМ СО АН СССР, 1966.

4. T.Marill and D.M.Green. On the Effectiveness of Receptors in Recognitions Systems. IREE Trans. on information theory, v. IT-9, January, 1963, pp 11-17.

5. Г. Фридман. О математическом ожидании ошибки определения вероятности ошибочной классификации. - Труды Института инженеров по электротехнике и радиотехнике (русский перевод) . М. , 1965., том 53, № 6, стр. 760.

6. Т. Андерсон. Введение в многомерный статистический анализ. М, ФМ. 1963.

7. Н.В. Дунин-Барковский, Н.В. Смирнов. Теория вероятностей и математическая статистика в технике (общая часть). М. 1955.

Поступила в редакцию
24. У. 1966 г.

П Р И Л О Ж Е Н И Е

Таблица I

$\overline{P_1} = 0{,}0099$

$\rho_1 = 4{,}66$

	$n = 2$				$n = 6$				$n = 10$			
	$N = 20$	$N = 50$	$N = 100$	$N = 150$	$N = 20$	$N = 50$	$N = 100$	$N = 150$	$N = 20$	$N = 50$	$N = 100$	$N = 150$
P'_1	0,0050	0,0060	0,0045	0,0045	0,0090	0,0080	0,0060	0,0065	0,0120	0,0075	0,0100	0,0100
P'_2	0,0085	0,0050	0,0045	0,0045	0,0090	0,0085	0,0055	0,0045	0,0125	0,0120	0,0085	0,0075
P'_3	0,0065	0,0045	0,0045	0,0045	0,0110	0,0065	0,0055	0,0075	0,0100	0,0075	0,0075	0,0100
P'_4	0,0050	0,0055	0,0045	0,0045	0,0065	0,0060	0,0075	0,0060	0,0165	0,0085	0,0085	0,0075
P'_5	0,0050	0,0055	0,0050	0,0045	0,0120	0,0060	0,0065	0,0045	0,0125	0,0110	0,0095	0,0090
P'_6	0,0050	0,0045	0,0045	0,0045	0,0075	0,0075	0,0060	0,0055	0,0115	0,0095	0,0095	0,0085
P'_7	0,0045	0,0045	0,0045	0,0050	0,0080	0,0090	0,0065	0,0070	0,0095	0,0110	0,0075	0,0080
P'_8	0,0055	0,0060	0,0050	0,0045	0,0065	0,0080	0,0065	0,0070	0,0085	0,0095	0,0080	0,0085
P'_9	0,0055	0,0050	0,0045	0,0050	0,0075	0,0065	0,0055	0,0065	0,0105	0,0115	0,0090	0,0090
P'_{10}	0,0055	0,0045	0,0050	0,0045	0,0080	0,0055	0,0065	0,0055	0,0085	0,0085	0,0085	0,0090
P'_{11}	0,0060	0,0050	0,0075	0,0045	0,0080	0,0085	0,0060	0,0070	0,0145	0,0085	0,0095	0,0090
P'_{12}	0,0080	0,0045	0,0045	0,0045	0,0105	0,0055	0,0065	0,0060	0,0100	0,0100	0,0080	0,0100
P'_{13}	0,0050	0,0050	0,0050	0,0045	0,0095	0,0060	0,0065	0,0060	0,0085	0,0075	0,0085	0,0080
P'_{14}	0,0045	0,0045	0,0050	0,0045	0,0065	0,0065	0,0060	0,0060	0,0115	0,0095	0,0095	0,0070
P'_{15}	0,0050	0,0045	0,0045	0,0045	0,0065	0,0060	0,0070	0,0060	0,0165	0,0090	0,0100	0,0085
P'_{max}	0,0085	0,0060	0,0075	0,0050	0,0120	0,0090	0,0075	0,0075	0,0165	0,0120	0,0100	0,0100
P'_{min}	0,0045	0,0045	0,0045	0,0045	0,0065	0,0055	0,0055	0,0045	0,0085	0,0075	0,0075	0,0070
$\Delta P'$	0,0040	0,0005	0,0030	0,0005	0,0055	0,0035	0,0020	0,0030	0,0080	0,0045	0,0025	0,0030

Таблица 2

$\bar{\mathcal{P}}_2 = 0{,}1003$

$\rho_2 = 2{,}56$

	$n=2$				$n=6$				$n=10$			
	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$
$\mathcal{P}_1'$	0,0890	0,0930	0,0890	0,0800	0,1080	0,1010	0,1005	0,1030	0,1225	0,1185	0,1160	0,1060
$\mathcal{P}_2'$	0,0970	0,0910	0,0930	0,0920	0,1055	0,1060	0,0990	0,1020	0,1225	0,1130	0,1075	0,1075
$\mathcal{P}_3'$	0,0920	0,0880	0,0905	0,0870	0,1205	0,1023	0,1000	0,0950	0,1215	0,1155	0,1100	0,1095
$\mathcal{P}_4'$	0,0885	0,0925	0,0895	0,0890	0,1040	0,1035	0,0975	0,0990	0,1218	0,1110	0,1065	0,1030
$\mathcal{P}_5'$	0,0895	0,0930	0,0915	0,0895	0,1110	0,1090	0,1055	0,1065	0,1115	0,1195	0,1120	0,1120
$\mathcal{P}_6'$	0,0935	0,0920	0,0900	0,0905	0,0980	0,1000	0,0985	0,0990	0,1315	0,1050	0,1090	0,1120
$\mathcal{P}_7'$	0,0920	0,0905	0,0895	0,0940	0,1010	0,1070	0,0985	0,1015	0,1190	0,1130	0,1050	0,1080
$\mathcal{P}_8'$	0,0955	0,0905	0,0875	0,0915	0,1065	0,1000	0,0980	0,0990	0,1200	0,1130	0,1175	0,1075
$\mathcal{P}_9'$	0,0925	0,0935	0,0900	0,0930	0,1295	0,1010	0,1045	0,1025	0,1260	0,1115	0,1105	0,1095
$\mathcal{P}_{10}'$	0,0930	0,0920	0,0925	0,0895	0,1060	0,1015	0,0985	0,1025	0,1325	0,1110	0,1000	0,1070
$\mathcal{P}_{11}'$	0,0925	0,0915	0,0925	0,0915	0,1165	0,1000	0,1020	0,0990	0,1230	0,1070	0,1065	0,1090
$\mathcal{P}_{12}'$	0,0950	0,0890	0,0930	0,0870	0,1195	0,1040	0,1005	0,1005	0,1160	0,1035	0,1070	0,1120
$\mathcal{P}_{13}'$	0,0920	0,0905	0,0925	0,0890	0,0995	0,1030	0,0955	0,0980	0,1330	0,1160	0,1060	0,1120
$\mathcal{P}_{14}'$	0,0900	0,0890	0,0875	0,0915	0,1025	0,1065	0,0995	0,0995	0,1160	0,1195	0,1055	0,1095
$\mathcal{P}_{15}'$	0,0985	0,0915	0,0925	0,0915	0,1100	0,1010	0,1065	0,1010	0,1335	0,1090	0,1065	0,1160
$\mathcal{P}'_{max}$	0,0985	0,0935	0,0930	0,0940	0,1295	0,1090	0,1065	0,1030	0,1335	0,1195	0,1175	0,1160
$\mathcal{P}'_{min}$	0,0885	0,0880	0,0875	0,0800	0,0980	0,1000	0,0955	0,0950	0,1115	0,1035	0,1050	0,1030
$\Delta\mathcal{P}'$	0,0100	0,0055	0,0055	0,0140	0,0315	0,0090	0,0110	0,0080	0,0220	0,0160	0,0125	0,0130

Таблица 3

$\bar{\mathcal{P}}_3 = 0{,}2981$

$\rho_3 = 1{,}06$

	$n=2$				$n=6$				$n=10$			
	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$
$\mathcal{P}_1'$	0,3055	0,3060	0,3075	0,3030	0,3410	0,3160	0,3210	0,3285	0,3820	0,3380	0,3145	0,3105
$\mathcal{P}_2'$	0,3055	0,3130	0,3065	0,3105	0,3300	0,3290	0,3065	0,3095	0,3485	0,3480	0,3050	0,3155
$\mathcal{P}_3'$	0,3045	0,3075	0,3034	0,3050	0,3385	0,3135	0,3185	0,3095	0,3225	0,3205	0,3110	0,3150
$\mathcal{P}_4'$	0,3030	0,3115	0,3085	0,3080	0,3435	0,3260	0,3115	0,3190	0,3530	0,3285	0,3050	0,3240
$\mathcal{P}_5'$	0,3060	0,3075	0,3060	0,3065	0,3185	0,3195	0,3175	0,3080	0,3295	0,3170	0,3335	0,3180
$\mathcal{P}_6'$	0,3090	0,3075	0,3060	0,3065	0,3105	0,3145	0,3170	0,3145	0,3660	0,3185	0,3155	0,3105
$\mathcal{P}_7'$	0,3090	0,3070	0,3030	0,3065	0,3195	0,3245	0,3180	0,3170	0,3090	0,3155	0,3165	0,3034
$\mathcal{P}_8'$	0,3075	0,3034	0,3080	0,3055	0,3225	0,3135	0,3190	0,3180	0,3515	0,3265	0,3115	0,2995
$\mathcal{P}_9'$	0,3155	0,3145	0,3085	0,3080	0,3360	0,3205	0,3200	0,3165	0,3485	0,3350	0,3125	0,3050
$\mathcal{P}_{10}'$	0,3120	0,3090	0,3070	0,3020	0,3160	0,3200	0,3245	0,3225	0,3645	0,3265	0,3070	0,3135
$\mathcal{P}_{11}'$	0,3055	0,3150	0,3070	0,3050	0,3420	0,3165	0,3300	0,3140	0,3325	0,3345	0,3145	0,3055
$\mathcal{P}_{12}'$	0,3045	0,3070	0,3075	0,3060	0,3235	0,3130	0,3110	0,3090	0,3120	0,3195	0,3230	0,3115
$\mathcal{P}_{13}'$	0,3080	0,3055	0,3055	0,3065	0,3175	0,3185	0,3060	0,3155	0,3510	0,3255	0,3220	0,3135
$\mathcal{P}_{14}'$	0,3155	0,3034	0,3060	0,3060	0,3135	0,3210	0,3085	0,3125	0,3350	0,3515	0,3110	0,3065
$\mathcal{P}_{15}'$	0,3185	0,3070	0,3100	0,3065	0,3405	0,3240	0,3125	0,3205	0,3670	0,3210	0,3140	0,3265
$\mathcal{P}'_{max}$	0,3185	0,3150	0,3100	0,3105	0,3435	0,3290	0,3300	0,3285	0,3820	0,3515	0,3335	0,3265
$\mathcal{P}'_{min}$	0,3050	0,3034	0,3030	0,3020	0,3105	0,3130	0,3060	0,3080	0,3090	0,3135	0,3050	0,3034
$\Delta\mathcal{P}'$	0,0155	0,0126	0,0070	0,0085	0,0330	0,0160	0,0240	0,0205	0,0730	0,0380	0,0285	0,0231

Таблица 4

$\mathcal{P}_4 = 0{,}4483$

$\rho_4 = 0{,}13$

	$n = 2$				$n = 6$				$n = 10$			
	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$	$N=20$	$N=50$	$N=100$	$N=150$
$\mathcal{P}'_1$	0,4820	0,4915	0,4745	0,4760	0,4960	0,4885	0,4805	0,4925	0,5070	0,5000	0,4865	0,4975
$\mathcal{P}'_2$	0,4745	0,4855	0,4750	0,4940	0,4900	0,4915	0,4840	0,4785	0,4935	0,4855	0,4905	0,4985
$\mathcal{P}'_3$	0,5020	0,4740	0,4735	0,4765	0,4895	0,4855	0,4850	0,4845	0,4895	0,4945	0,4840	0,4865
$\mathcal{P}'_4$	0,4770	0,4800	0,4755	0,5045	0,5095	0,5095	0,4940	0,4855	0,4995	0,4935	0,4985	0,4885
$\mathcal{P}'_5$	0,4715	0,4775	0,4785	0,4715	0,4790	0,4885	0,4940	0,4750	0,4915	0,4905	0,5015	0,4925
$\mathcal{P}'_6$	0,4745	0,4745	0,4705	0,4715	0,4945	0,4870	0,4980	0,4895	0,4945	0,4795	0,4955	0,4820
$\mathcal{P}'_7$	0,4775	0,4790	0,4785	0,5045	0,4810	0,4990	0,4820	0,4940	0,4775	0,4755	0,4770	0,4880
$\mathcal{P}'_8$	0,4775	0,4865	0,4840	0,4690	0,5060	0,4805	0,5000	0,4935	0,4965	0,4870	0,5005	0,5050
$\mathcal{P}'_9$	0,5005	0,4845	0,4765	0,4800	0,5100	0,4810	0,4810	0,4955	0,4890	0,4805	0,4845	0,5015
$\mathcal{P}'_{10}$	0,4990	0,4755	0,5075	0,4725	0,4855	0,4995	0,5120	0,4890	0,4990	0,4975	0,4810	0,5000
$\mathcal{P}'_{11}$	0,4855	0,5055	0,4755	0,4925	0,4980	0,5025	0,4810	0,4855	0,4925	0,4850	0,4985	0,5025
$\mathcal{P}'_{12}$	0,4770	0,4810	0,4760	0,4825	0,5120	0,4905	0,4890	0,4895	0,4925	0,4905	0,4935	0,4810
$\mathcal{P}'_{13}$	0,5010	0,5010	0,4765	0,4840	0,5060	0,5095	0,4875	0,4795	0,4900	0,5050	0,4825	0,4960
$\mathcal{P}'_{14}$	0,4860	0,4750	0,4795	0,4790	0,4880	0,4910	0,4805	0,4780	0,4850	0,4945	0,4975	0,5065
$\mathcal{P}'_{15}$	0,5075	0,4800	0,4750	0,4500	0,4995	0,5010	0,4850	0,4980	0,4950	0,4975	0,4825	0,4900
$\mathcal{P}'max$	0,5020	0,5010	0,5075	0,5045	0,5120	0,5095	0,5120	0,4955	0,5070	0,5050	0,5015	0,5065
$\mathcal{P}'min$	0,4715	0,4740	0,4735	0,4690	0,4790	0,4805	0,4810	0,4750	0,4775	0,4755	0,4770	0,4810
$\Delta\mathcal{P}'$	0,0305	0,0270	0,0340	0,0355	0,0330	0,0290	0,0310	0,0205	0,0295	0,0295	0,0245	0,0255