УДК 62-506

БАЙЕСОВЫ МЕТОДЫ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

Ю. И. ИНГСТЕР

(Ленинград)

Изучаются свойства статистических критериев качества в задаче автоматической классификации (самообучения). Устанавливается связь статистических и вариационных методов в задаче автоматической классификации. Рассматривается случай неизвестного числа классов.

1. Введение

Пусть имеется неклассифицированная случайная выборка $X_s = (\mathbf{x}_1, \ldots, \mathbf{x}_s)$, образованная векторами признаков $\mathbf{x}_i \in E$, где E — измеримое пространство. Каждый вектор $\mathbf{x}_i$ порождается одним из $M \geqslant 1$ классов, число которых предполагается далее, если не оговорено противное, конечным. Гипотезой о распределении классов в выборке из s векторов будем называть разбиение $a_s = (A_{1s}, \ldots, A_{Ms})$ множества номеров $\{1, \ldots, s\}$ векторов выборки на M непересекающихся подмножеств (некоторые из которых, возможно, пусты). Соответствующие подмножества $\{x_i, i \in A_{ms}\}$ выборки X_s будем обозначать X_{ms}. Число различных гипотез a_s есть M^s. Они в совокупности образуют полную систему гипотез, из которых для выборки X_s одна является истинной (реализуется «природой»).

Требуется построить алгоритм выбора гипотезы $a_s^* = a_s^*(X_s)$ по выборке X_s, оптимальный в определенном смысле (т. е. алгоритм классификации векторов выборки X_s).

Эта задача относится к числу задач самообучения распознаванию образов (автоматической классификации). Обычно [1, 2] она рассматривается в рамках вариационного или геометрического подхода. В пространстве признаков E фиксируется мера расстояния или близости, на основе которой конструируется оптимизируемый функционал $\varphi(X_s, a_s)$, характеризующий разброс векторов в группах X_{ms} или разнесенность групп в пространстве признаков. Для некоторых функционалов $\varphi(X_s, a_s)$ имеются итеративные или рекуррентные [3] алгоритмы оптимизации.

В данной работе эта задача рассматривается в рамках теории статистических решений (аналогично задаче классификации одного вектора).

При заданной матрице потерь $\{W(a_s, a_s^*)\}$ (размера $M^s \times M^s$) оптимальный по критерию минимума среднего риска алгоритм решения задачи состоит в выборе гипотезы $a_s^* = a_s^*(X_s)$, на которой достигается

$$\min_{a_s^*} \sum_{a_s} W(a_s, a_s^*) p(a_s) f(X_s | a_s).$$

Здесь $W(a_s, a_s^*)$ — величина потерь при выборе гипотезы a_s^*, в то время как истинной является гипотеза a_s; $f(X_s | a_s)$ — условная плотность ве-

роятностей выборки X_s в предположении, что истинной является гипотеза a_s; $p(a_s)$ — априорная вероятность гипотезы a_s.

Далее будем рассматривать случай вырожденной матрицы потерь: $W(a_s, a_s^*)=\{0$, если $a_s \neq a_s^*$; -1, если $a_s=a_s^*\}$, который приводит к критерию

$$\max_{a_s^*} f(X_s, a_s)=\max_{a_s} p(a_s) f(X_s|a_s), \tag{1}$$

а также к критерию

$$\max_{a_s} f(X_s|a_s), \tag{2}$$

вытекающему из (1) при равных априорных вероятностях гипотез a_s.

Цель данной работы — исследование структуры и асимптотических свойств критериев (1) и (2). Оно проводится в рамках параметрического байесового подхода, основные положения которого для задач распознавания образов разработаны в [4].

2. Структура байесовых критериев

Пусть задано параметрическое семейство плотностей распределений $f(\mathbf{x}|\boldsymbol{\theta})$ $(\mathbf{x}\in E,\ \boldsymbol{\theta}\in\Theta)$. Пусть $\boldsymbol{\theta}_m$ — значение параметра, соответствующее m-му классу, p_m — априорная вероятность m-го класса $(m=1,\ldots,M)$. Обозначим совокупности параметров:

$$\boldsymbol{\xi}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M)\in\Theta^M,\quad \boldsymbol{\alpha}=(p_1,\ldots,p_M)\in\Delta_{M-1}=\left\{\boldsymbol{\alpha}\colon p_m\geqslant 0;\ \sum_m p_m=1\right\};\quad \boldsymbol{\eta}(\boldsymbol{\xi},\boldsymbol{\alpha})\in\Theta^M\times\Delta_{M-1}.$$

При выполнении предположений о независимости векторов выборки (имеющих смешанное распределение с параметром $\boldsymbol{\eta}$), используемых в статистических методах самообучения [1, 2], и при известных параметрах $\boldsymbol{\xi}$ или $\boldsymbol{\eta}$ имеем

$$f(X_s|a_s,\boldsymbol{\xi})=\prod_m\prod_{i\in A_{ms}}\{f(\mathbf{x}_i|\boldsymbol{\theta}_m)\},\ f(X_s,a_s|\boldsymbol{\eta})=\prod_m\prod_{i\in A_{ms}}\{p_m f(\mathbf{x}_i|\boldsymbol{\theta}_m)\}. \tag{3}$$

При неизвестных параметрах $\boldsymbol{\xi}$ или $\boldsymbol{\eta}$ в соответствии с байесовым подходом будем рассматривать их как случайные с априорными распределениями $p(d\boldsymbol{\xi})$, $p(d\boldsymbol{\eta})$. При этом по формуле Байеса, считая, что $p(d\boldsymbol{\xi}|a_s)=p(d\boldsymbol{\xi})$, имеем

$$f(X_s|a_s)=\int_{\Theta^M} f(X_s|a_s,\boldsymbol{\xi})\,p(d\boldsymbol{\xi}), \tag{4}$$

$$f(X_s,a_s)=\int_{\Theta^M\times\Delta_{M-1}} f(X_s,a_s|\boldsymbol{\eta})\,p(d\boldsymbol{\eta}).$$

Выражения (3), (4) определяют структуру байесовых критериев. Для частичного случая критерий (2), (4) рассматривался в [5]. В общем случае аналитические методы вычисления (4) отсутствуют.

В данной работе строятся асимптотические аппроксимации критериев (3) и (4). Рассмотрим максимальные значения логарифмических функций правдоподобия

$$L(X_s \mid a_s) = s^{-1} \sup_{\xi} \ln f(X_s \mid a_s, \xi),$$

$$L(X_s, a_s) = s^{-1} \sup_{\eta} \ln f(X_s, a_s \mid \eta).$$

При достаточно общих условиях, как показано в приложении 1,

$$|s^{-1} \ln f(X_s|a_s) - L(X_s|a_s)| \underset{s\to\infty}{\longrightarrow} 0, \tag{5}$$

$$|s^{-1} \ln f(X_s, a_s) - L(X_s, a_s)| \underset{s\to\infty}{\longrightarrow} 0. \tag{6}$$

Соотношения (5), (6) при больших s позволяют считать, что

$$s^{-1} \ln f(X_s|a_s) \approx L(X_s|a_s), \quad s^{-1} \ln f(X_s, a_s) \approx L(X_s, a_s). \tag{7}$$

Отметим, что в аппроксимации (7) не входят априорные распределения $p(d\xi)$, $p(d\eta)$, так как при выполнении весьма общих условий «невырожденности» (см. приложение 1) они асимптотически несущественны.

Аппроксимации (7) достаточны для целей данной работы. Более точные аппроксимации могут быть получены методами, аналогичными [6].

3. Экспоненциальные семейства распределений

В теоретических исследованиях в области математической статистики и в ее приложениях [7—9] широко используются экспоненциальные семейства распределений, плотности которых имеют вид

$$f(\mathbf{x}|\theta) = \exp\{-h(\theta) + \langle \mathbf{y}(\mathbf{x})\theta\rangle + c(\mathbf{x})\}, \quad \theta \in \Theta. \tag{8}$$

Здесь $\theta = (\theta_{(1)}, \ldots, \theta_{(R)})$ — вектор канонических параметров; $\mathbf{y}(\mathbf{x}) = (y^{(1)}(\mathbf{x}), \ldots, y^{(R)}(\mathbf{x}))$ — вектор достаточных статистик, $\langle \mathbf{y}\theta\rangle = \sum_{r=1}^{R} y^{(r)}\theta_{(r)}$, $h(\theta) = \ln \int_E \exp\{\langle \mathbf{y}(\mathbf{x})\theta\rangle + c(\mathbf{x})\}\mu(d\mathbf{x})$ — нормирующий множитель; $\Theta = \{\theta : h(\theta) < \infty\}$; $\mu(d\mathbf{x})$ — мера на пространстве E. В этом разделе и в приложении 2 используются определения и результаты работы [7] (глава IV).

Для экспоненциальных семейств (8) в приложении 2 доказаны свойства равномерности предельных соотношений (5) и (6), а также приведены оценки точности аппроксимаций (7).

Для семейств (8) величина $L(X_s|a_s)$ представляет собой функцию фиксированного аналитического вида. Пусть

$$g(\mathbf{y}, \theta) = -h(\theta) + \langle \mathbf{y}\theta\rangle, \quad g(\mathbf{y}) = \sup_{\theta} g(\mathbf{y}\theta). \tag{9}$$

Легко видеть, что

$$L(X_s|a_s) = \sum_m p_{ms} g(\mathbf{y}_{ms}) + C_s, \tag{10}$$

где $\mathbf{y}_{ms} = s_m^{-1} \sum_{i \in A_{ms}} \mathbf{y}(\mathbf{x}_i)$ — среднее по векторам, относимым к m-му классу; слагаемое $C_s = s^{-1} \sum_{i=1}^{s} c(\mathbf{x}_i)$ для классификации является несущественным; $s_m = s_m(a_s)$ — число элементов множества $A_{m,s}$; $p_{ms} = s_m/s$.

Практически важные примеры вычисления функции $g(\mathbf{y})$ приведены в приложении 3.

Для семейств (8) величина $L(X_s|a_s)$ допускает наглядную геометрическую интерпретацию. Рассмотрим функционал информационного уклонения

$$I(\theta_1, \theta_2) = \int_E \{\ln f(\mathbf{x}|\theta_2) - \ln f(\mathbf{x}|\theta_1)\} f(\mathbf{x}|\theta_2)\, \mu(d\mathbf{x}),$$

определенный для параметров $\theta_1 \in \Theta$, $\theta_2 \in \Theta$.

Не ограничивая общности, можно считать, что параметризация экспоненциального семейства является точной, т. е. множество внутренних точек $\operatorname{Int}\Theta$ множества Θ непусто и соответствие $\theta \leftrightarrow f(\mathbf{x}|\theta)$ взаимно-однозначно. Обозначим $\theta(\mathbf{y})$ значение параметра θ, доставляющее максимум функции $g(\mathbf{y}, \theta)$. Ясно, что $\left.\dfrac{\partial h(\theta)}{\partial \theta}\right|_{\theta=\theta(\mathbf{y})} = \mathbf{y}$.

Введем множество $Y = \left\{\mathbf{y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(R)}) : \mathbf{y}^{(r)} = \dfrac{\partial h(\theta)}{\partial \theta_{(r)}}, \quad \theta \in \operatorname{Int}\Theta\right\}$,

представляющее собой внутреннюю часть пространства натуральных параметров [7] и лежащее в пространстве достаточных статистик. При этом соответствие $\mathbf{y} \leftrightarrow \theta(\mathbf{y})$ взаимно-однозначно на множествах Y и $\operatorname{Int}\Theta$. Для регулярных экспоненциальных семейств замыкания множества выпуклых комбинаций значений $\{\mathbf{y}(\mathbf{x}), \mathbf{x} \in E\}$ и множества Y совпадают ([7], § 21). Кроме того, справедливо соотношение

$$\mathbf{y} = \int_E \mathbf{y}(\mathbf{x}) f(\mathbf{x}|\theta(\mathbf{y}))\, \mu(d\mathbf{x}).$$

Это позволяет рассматривать функционал информационного уклонения в пространстве выпуклых комбинаций векторов достаточных статистик, где его можно представить в виде

$$I(\mathbf{y}_1, \mathbf{y}_2) = g(\mathbf{y}_2) - g(\mathbf{y}_2, \theta(\mathbf{y}_1)). \tag{11}$$

Этот функционал можно рассматривать как аналог меры расстояния в пространстве Y. Он не удовлетворяет аксиомам метрики (в частности, несимметричен), но обладает многими свойствами квадрата нормы.

Практически важные примеры вычисления величины $I(\mathbf{y}_1, \mathbf{y}_2)$ приведены в приложении 3.

Если $\mathbf{y}(\mathbf{x}_i) \in Y$ (это выполняется, если плотность $f(\mathbf{x}|\theta)$ ограничена и достигает максимума по $\theta \in \Theta$), то

$$L(X_s|a_s) = -\sum_m p_{ms} \left(s_m^{-1} \sum_{i \in A_{ms}} I(\mathbf{y}_{ms}, \mathbf{y}(\mathbf{x}_i)) \right) + C_s, \tag{12}$$

где $C_s = -s^{-1} \sum_{i=1} (g(\mathbf{y}(\mathbf{x}_i)) - c(\mathbf{x}_i))$ — слагаемое, несущественное для классификации. Соотношение (12) вытекает из равенства

$$g(\mathbf{y}_{ms}) = s_m^{-1} \sum_{i \in A_{ms}} g(\mathbf{y}(\mathbf{x}_i), \theta(\mathbf{y}_{ms})).$$

Представление (12) справедливо, в частности, для семейств 1, 2 приложения 3 и семейства 3 при $\mathbf{x}_i \neq \mathbf{0}$.

Если $\mathbf{y}_{ms} \in Y$ $(m = 1, \ldots, M)$, то также можно показать, что

$$L(X_s|a_s) = \sum_m p_{ms} I(\mathbf{y}_s, \mathbf{y}_{ms}) + C_s, \tag{13}$$

где $\mathbf{y}_s=s^{-1}\sum_{i=1}^{s}\mathbf{y}(\mathbf{x}_i)$; $C_s=g(\mathbf{y}_s)+s^{-1}\sum_{i=1}^{s}c(\mathbf{x}_i)$ — несущественное слагаемое. Представление (13) является более общим и справедливо, в частности, для семейств 1, 2 приложения 3; семейства 3 при $\mathbf{x}_i\neq\mathbf{0}$, с вероятностью 1 для семейств 4, если $s_m>R$ и выборка X_s порождена непрерывным распределением.

4. Приближенные байесовы критерии и алгоритмы их оптимизации

1. Аппроксимации (7) позволяют перейти от критериев (2), (1) к приближенным байесовым критериям соответственно

$$\max_{a_s} L(X_s|a_s), \tag{14}$$

$$\max_{a_s} L(X_s, a_s). \tag{15}$$

Для семейства многомерных нормальных распределений общего вида критерий (14) рассматривался в [10].

Для экспоненциальных семейств (8), как следует из соотношений (12), (13), критерий (14) эквивалентен критерию минимума суммарного разброса векторов достаточных статистик в подмножествах, соответствующих классам, или критерию максимума разнесенности центров подмножеств в пространстве достаточных статистик относительно центра выборки, причем в качестве меры расстояния используется информационное уклонение (11).

2. Критерии (14), (15) можно представить в виде

$$\begin{aligned} &\max_{a_s}\max_{\xi} s^{-1}\sum_m\sum_{i\in A_{ms}}\ln(f(\mathbf{x}_i\,|\,\boldsymbol{\theta}_m));\\ &\max_{a_s}\max_{\eta} s^{-1}\sum_m\sum_{i\in A_{ms}}\ln(p_m f(\mathbf{x}_i\,|\,\boldsymbol{\theta}_m)). \end{aligned} \tag{16}$$

Это позволяет применить для их оптимизации итеративную процедуру, каждый шаг которой состоит из 1) максимизации по ξ или η при фиксированной гипотезе a_s, что сводится к построению оценок максимального правдоподобия $\boldsymbol{\theta}_{ms}$ по выборкам X_{ms}; 2) максимизации по a_s при фиксированных ξ или η, что сводится к классификации векторов выборки X_s при известных значениях параметров по критерию идеального наблюдателя. На каждом шаге процедуры значения максимизируемых выражений не убывают, что обеспечивает сходимость к локальному максимуму (16).

Подобные итеративные алгоритмы предлагались для оптимизации различных геометрических критериев группировки (например, [11]). Необходимо отметить, однако, что критерии (14), (15), как правило, многоэкстремальны.

Для экспоненциальных семейств (8) нахождение оценок $\boldsymbol{\theta}_{ms}$ сводится к решению векторных уравнений $\partial h/\partial\boldsymbol{\theta}=\mathbf{y}_{ms}$, а классификация на каждом шаге проводится по максимуму линейных от $\mathbf{y}$ функций $g(\mathbf{y}, \boldsymbol{\theta}_{ms})$ или по минимуму информационных уклонений $I(\mathbf{y}_{ms}, \mathbf{y})$ в пространстве достаточных статистик.

3. В случае неограниченной выборки критерии (1), (2), (14), (15) сводятся к критерию выбора разбиения $e=(E_1, \ldots, E_M)$ пространства признаков на непересекающиеся подмножества E_m, предложенному в [3], при специальном выборе функций потерь. Это позволяет применить для их оптимизации рекуррентные алгоритмы [3].

Действительно, разбиение e индуцирует последовательность гипотез $\{a_s\}_{s=1}^{\infty}$: $A_{ms}=\{i : \mathbf{x}_i \in E_m,\ i \leqslant s\}$. При выполнении достаточно общих условий (см. следствие 2 приложения 1 и следствие 3 приложения 2):

$$\lim_{s\to\infty} s^{-1} \ln f(X_s \mid a_s) = \lim_{s\to\infty} L(X_s \mid a_s) = \sup_{\xi} \sum_m \int_{E_m} \ln f(\mathbf{x} \mid \theta_m)\, H(d\mathbf{x});$$

(17)

$$\lim_{s\to\infty} s^{-1} \ln f(X_s, a_s) = \lim_{s\to\infty} L(X_s, a_s) = \sup_{\eta} \sum_m \int_{E_m} \ln (p_m f(\mathbf{x} \mid \theta_m)) \times$$
$$\times H(d\mathbf{x}),$$

откуда следует, что критерии (1), (2), (14), (15) при $s\to\infty$ сводятся к критериям [3]

$$\min_e \min_{\mathbf{c}} \sum_m \int_{E_m} F_m(\mathbf{x}, \mathbf{c}) H(d\mathbf{x})$$

при выборе функций потерь $F_m(\mathbf{x}, \mathbf{c})$ в виде

(18) $$F_m(\mathbf{x}, \xi) = -\ln f(\mathbf{x} \mid \theta_m), \quad F_m(\mathbf{x}, \eta) = -\ln [p_m f(\mathbf{x} \mid \theta_m)].$$

Таким образом, задание функций потерь при использовании методов [3] соответствует заданию параметрического семейства плотностей в рамках методов байесова подхода; априорные распределения неизвестных параметров при этом, как отмечалось, асимптотически несущественны.

Соотношения (18) можно также рассматривать как рекомендации по выбору функций потерь в [3]. Аналогичный результат на основе других соображений получен в [12].

4. Рассмотрим критерии (1), (15). Справедливо соотношение

$$L(X_s, a_s) = L(X_s \mid a_s) + E(a_s),$$

где

$$E(a_s) = \max_{\alpha \in \Delta_{M-1}} \sum_m p_{ms} \ln p_m = \sum_m p_{ms} \ln p_{ms}.$$

Величина $E(a_s)$ представляет собой один из вариантов «меры концентрации» [2]. Учет члена $E(a_s)$ вызывает тенденцию к группировке векторов, так как величина $E(a_s)$ принимает максимальное значение, равное 0, при объединении всех векторов в один класс, и минимальное значение (равное $-\ln M$ при $s=kM$, $k\geqslant 1$ — целое число) в случае, когда векторы равномерно распределены по классам.

Это дает основание для использования критериев (1), (15) при неизвестном (бесконечном) числе классов (легко видеть, что критерий (14) в этом случае неэффективен, так как он достигает максимального значения в случае, когда каждому вектору соответствует свой класс).

Выше рассматривался случай конечного числа классов. Однако определение гипотезы a_s переносится на случай $M=\infty$; при этом число гипотез a_s бесконечно, но для каждой гипотезы имеется не более s непустых подмножеств A_{ms}, и поэтому произведения в (3) содержат не более s сомножителей, отличных от 1. Поэтому при $M=\infty$ интегрирование в (4) ведется лишь по конечному числу параметров в соответствии с маргинальными распределениями параметров из бесконечных совокупностей ξ, η.

При $M=\infty$ максимумы критериев (1), (15) могут не достигаться, поэтому под оптимальными следует понимать гипотезы, обеспечивающие ε-близость к верхним граням критериев.

Априорные распределения $p(d\xi)$, $p(d\eta)$ бесконечных совокупностей ξ и η должны обеспечивать невырожденность распределений конечных совокупностей параметров, что позволит использовать соотношения (7) для

последовательностей гипотез с конечным числом классов. Этого можно добиться, например, считая параметры θ_m и совокупность α независимыми: $p(d\eta)=p(d\alpha)\prod_m p_m(d\theta_m)$, распределения $p_m(d\theta_m)$ невырожденными, а априорное распределение совокупности $\alpha=(p_1,\ldots,p_m,\ldots)$ $\left(p_m\geqslant 0, \sum_{m=1}^{\infty} p_m=1\right)$ задавать путем задания согласованного набора конечномерных распределений, выбираемых в классе распределений Дирихле [13]

$$p(dp_{m_1},\ldots,dp_{m_k})=D\left(dp_{m_1},\ldots,dp_{m_k}\,|\,\nu_{m_1},\ldots,\nu_{m_k};\nu-\sum_{i=1}^{k}\nu_{m_i}\right),$$

где $\nu_1,\ldots,\nu_m,\ldots$ — последовательность, удовлетворяющая условиям

$$\nu_m>0,\qquad \sum_{m=1}^{\infty}\nu_m<\infty.$$

Анализ свойств критериев (1), (15), проведенный в приложении 4 при сравнительно слабых ограничениях (близких к ограниченности дисперсии выборки), приводит к следующим результатам.

Пусть $K(R, s)$ — общее количество тех векторов для оптимальной (или «достаточно хорошей», см. приложение 4) гипотезы по критерию (1) или (15) при $M=\infty$, которые лежат в классах, содержащих не более R векторов. Пусть $M(s)$ — количество непустых классов для оптимальной (или «достаточно хорошей») гипотезы по критерию (1) или (15). Справедливы соотношения:

(19) $$s^{-1}K(R, s)=O((\ln s/R)^{-1}),$$

(20) $$s^{-1}M(s)=O((\ln s)^{-1}).$$

Таким образом, для оптимальных (или «достаточно хороших») гипотез подавляющее число векторов содержится в классах с большим числом векторов. Например, полагая $R=s^{\gamma}$ $(0<\gamma<1)$, получим, что доля векторов, которые содержатся в классах с числом векторов, бо́льшим R, составляет не меньше $1-O((\ln s)^{-1})$ и стремится к 1 с ростом объема выборки.

Кроме того, для этих гипотез количество непустых классов растет не быстрее $s/\ln s$, а среднее число векторов в непустых классах есть величина порядка не меньшего, чем $\ln s$.

Заключение

Полученные результаты показывают, что приближенные байесовы методы при больших выборках могут быть использованы в задаче автоматической классификации (самообучения), рассматриваемой как задача выбора наиболее вероятной гипотезы о распределении классов в неклассифицированной выборке. Априорная информация, необходимая для использования этих методов, по существу сводится к заданию параметрического семейства распределений признаков, поскольку априорные распределения параметров асимптотически несущественны.

Имеется тесная связь «геометрических» и статистических критериев качества, состоящая в том, что для экспоненциальных семейств распределений аппроксимации байесовых критериев допускают геометрическую интерпретацию, причем возникающая «мера расстояния» определяется самим семейством распределений, а не выбирается эвристически.

В рамках байесовых методов в асимптотике также удается получить вариационные критерии Я. З. Цыпкина и Г. К. Кельманса, причем вид функций потерь также определяется рассматриваемым семейством распределений.

Алгоритмы оптимизации приближенных байесовых критериев аналогичны известным алгоритмам оптимизации геометрических или вариационных критериев.

Байесов подход также позволяет рассмотреть задачу самообучения при неизвестном (бесконечном) числе классов, не привлекая дополнительных (нестатистических) предположений типа «стоимости классов». При этом обеспечивается достаточно эффективная «группировка» векторов выборки.

ПРИЛОЖЕНИЕ I

Условия справедливости аппроксимаций [5], [6]

Введем функции

$$L(X_{ms}\,|\,\theta)=s_m^{-1}\sum_{i\in A_{ms}}\ln f(\mathbf{x}_i\,|\,\theta),\qquad L(X_{ms})=\sup_{\theta\in\Theta}L(X_{ms}\,|\,\theta).$$

Ясно, что

$$L(X_s\,|\,a_s)=\sum_m p_{ms}L(X_{ms}).$$

Сформулируем следующие условия.

1. Пространство Θ локально компактно. Меры $p(d\xi)$, $p(d\eta)$ определены на борелевских множествах пространств Θ^M и $\Theta^M\times\Delta_{M-1}$. Априорные распределения не вырождены, т. е. носители мер $p(d\xi)$, $p(d\eta)$ суть пространства Θ^M и $\Theta^M\times\Delta_{M-1}$.

2. $p(d\eta)=p(d\xi)P(\alpha)\nu(d\alpha)$, где $P(\alpha)$ — непрерывная априорная плотность относительно лебеговой меры $\nu(d\alpha)$ на Δ_{M-1}; существуют c_1, c_2:

$$\text{(П.1)}\qquad \ln P(\alpha)\geqslant c_1\min_m\{\ln p_m\}+c_2$$

(т. е. плотность $P(\alpha)$ не слишком быстро убывает при приближении к границе симплекса Δ_{M-1}). Этому условию удовлетворяют, в частности, распределения Дирихле.

3. Выборка X_s есть отрезок неограниченной выборки $X=\{\mathbf{x}_i\}_{i=1}^{\infty}$. Последовательность $\{a_s\}_{i=1}^{\infty}$ такова, что векторы в выборках X_{ms} независимы с распределением $H_m(d\mathbf{x})$; существуют пределы $p_m^*=\lim p_{ms}$ при $s\to\infty$.

Условия 3, в частности, выполняются, если последовательность $\{a_s\}$ порождена измеримым разбиением $e=(E_1,\ldots,E_M)$ пространства признаков на непересекающиеся множества E_m: $A_{ms}=\{i:\mathbf{x}_i\in E_m,\ i\leqslant s\}$; $X=\{\mathbf{x}_i\}_{i=1}^{\infty}$ — выборка независимых векторов с распределением $H(d\mathbf{x})$. При этом $H_m(d\mathbf{x})=H(d\mathbf{x}\,|\,E_m)$, $p_m^*=H(E_m)$.

4. Функция $\ln f(\mathbf{x}\,|\,\theta)$ непрерывна по $\theta\in\Theta$. Для любого компактного множества $G\subset\Theta$ существует измеримая функция $h(\mathbf{x})$:

$$\text{(П.2)}\qquad \sup_{\theta\in G}|\ln f(\mathbf{x}\,|\,\theta)|\leqslant h(\mathbf{x}),\qquad \int_E h(\mathbf{x})H_m(d\mathbf{x})<\infty\quad(m=1,\ldots,M).$$

Из условий 3, 4 вытекает, что с вероятностью 1 при $s\to\infty$

$$\text{(П.3)}\qquad L(X_{ms}\,|\,\theta)\to L_m(\theta)=\int_E\ln f(\mathbf{x}\,|\,\theta)H_m(d\mathbf{x})$$

равномерно на компактных множествах $G\subset\Theta$ (см. [14], следствие 4.1).

5. С вероятностью 1 $L(X_{ms})\xrightarrow[s\to\infty]{}\sup_{\theta\in\Theta}L_m(\theta)=L_m$; если $p_m^*>0$, то $L_m<\infty$.

Условие 5 вытекает из условия 4, если пространство Θ компактно. Оно может быть выведено из условия 4 аналогично [15] (п. 5е. 2), если функции $L(X_{ms}\,|\,\theta)$ и $L_m(\theta)$ имеют единственный локальный максимум. Последнее выполняется, например, для экспоненциальных семейств.

Теорема 1. Пусть выполнены условия 1, 3—5. Тогда с вероятностью 1

$$\text{(П.4)}\qquad s^{-1}\ln f(X_s\,|\,a_s)\xrightarrow[s\to\infty]{}\sum_m p_m^*L_m.$$

Если при этом $p_m^*>0$ $(m=1, \ldots, M)$ или выполнено условие 2, то

$$s^{-1}\ln f(X_s, a_s) \underset{s\to\infty}{\longrightarrow} \sum_m [p_m^* L_m + p_m^* \ln p_m^*]. \tag{П.5}$$

Если же выполнены условия 1—5 и $A_M^{(s)}$ — множество возможных гипотез a_s с фиксированным числом M классов, то

$$\sup_{a_s\in A_M^{(s)}} |s^{-1}\ln p(a_s) - E(a_s)| = O(\ln s/s). \tag{П.6}$$

Следствие 1. При выполнении условий теоремы 1 с вероятностью 1 справедливы соотношения (5), (6).

Следствие 2. Пусть последовательность $\{a_s\}_{s=1}^{\infty}$ порождена разбиением $e=(E_1, \ldots, E_M)$ и выполнены условия 1, 4, 5. Тогда с вероятностью 1 справедливы соотношения (17).

Доказательство теоремы 1. Докажем (П.4). В силу условия 4 функция $L_m(\theta)$ непрерывна, и поэтому для любого $\varepsilon>0$ найдется компактное множество $G_m\subset\Theta$, имеющее внутренние точки и такое, что $L_m(\theta)>L_m-\varepsilon/3$ при $\theta\in G_m$. В силу (1.3) и условия 5 с вероятностью 1 найдется s_0, такое, что

$$|L(X_{ms}|\theta)-L_m(\theta)|<\varepsilon/3 \quad \text{при } s>s_0,\ \theta\in G_m,$$

а также $|L_m-L(X_{ms})|<\varepsilon/3$ при $s>s_0$. Отсюда для $\theta\in G_m$ имеем $|L(X_{ms})-L(X_{ms}|\theta)|<\varepsilon$ при $s>s_0$.

Пусть $G^{(M)}=G_1\times\ldots\times G_M\subset\Theta^M$. По условию 1 $p(G^{(M)})>0$. Соотношение (П.4) вытекает из условия 5 и цепочки неравенств

$$\begin{aligned} &L(X_s|a_s)\geqslant s^{-1}\ln f(X_s|a_s)\geqslant \\ &\geqslant s^{-1}\ln\left[\int_{G^{(M)}} \exp\left\{s\sum_m p_{ms}L(X_{ms}|\theta)\right\}p(d\xi)\right]\geqslant \\ &\geqslant \sum_m p_{ms}L(X_{ms}) + s^{-1}\ln p(G^{(M)}) - \varepsilon. \end{aligned}$$

Доказательство (П.5) при $p_m^*>0$ $(m=1, \ldots, M)$ проводится аналогично, а при выполнении условия 2 вытекает из (П.4) и (П.6).

Докажем (П.6). Из условия 2 следует, что

$$p(a_s)=\int_{\Delta_{M-1}}\left\{\prod_m p_m^{s_m}\right\}P(\alpha)\nu(d\alpha).$$

Пусть

$$G_s(\tilde\alpha)=\{\alpha\in\Delta_{M-1}: |p_m-\tilde p_m|>s^{-2};\quad p_m>s^{-3}\},\qquad \tilde\alpha=(\tilde p_1, \ldots, \tilde p_m)\in\Delta_{M-1};$$

лебегова мера множества $G_s(\tilde\alpha)$ удовлетворяет условию

$$\max_{\tilde\alpha\in\Delta_{M-1}} \nu(G_s(\tilde\alpha)) = O(s^{-2(M-1)}). \tag{П.7}$$

Пусть $\alpha_s=(p_{1s}, \ldots, p_{Ms})$. При $\alpha\in G_s(\alpha_s)$, $s>2$ для m таких, что $s_m\geqslant 1$, т. е. $p_{ms}\geqslant s^{-1}$, имеем $\ln p_m\geqslant\ln p_{ms}-2s^{-1}$;

$$\sum_{m=1}^{M} p_{ms}\ln p_m = \sum_{m:s_m>0} p_{ms}\ln p_m \geqslant \sum_m p_{ms}\ln p_{ms} - 2s^{-1}.$$

Из условия 2 и (П.7) также имеем

$$\min_{a_s}\ln p(G_s(\alpha_s))\geqslant \min_{a_s}\min_{\alpha\in G_s(\alpha_s)}\{\ln P(\alpha)+\ln\nu(G_s(\alpha_s))\} = O(\ln s).$$

Соотношение (П.6) вытекает из цепочки неравенств

$$\begin{aligned} &E(a_s)\geqslant s^{-1}\ln p(a_s)\geqslant \\ &\geqslant s^{-1}\ln\int_{G_s(\alpha_s)}\exp\left\{s\sum_m p_{ms}\ln p_m\right\}P(\alpha)\nu(d\alpha)\geqslant \\ &\geqslant E(a_s)-2s^{-1}+s^{-1}\ln p(G_s(\alpha_s)) = E(a_s)-O(s^{-1}\ln s). \end{aligned}$$

Свойства аппроксимаций [7] для экспоненциальных семейств распределений [8]

Теорема 2. а. Пусть $A_G^{(s)}$ — множество гипотез a_s таких, что $\mathbf{y}_{ms}\in G\subset Y$ при $s_m>0$ $(m=1, \ldots, M)$, где G — компактное множество, и выполнено условие 1 приложения 1. Тогда

$$\text{(П.8)}\qquad \sup_{a_s\in A_G^{(s)}} |s^{-1}\ln f(X_s|a_s)-L(X_s|a_s)| \xrightarrow[s\to\infty]{} 0.$$

б. Пусть, кроме того, выполнено условие 2 приложения 1. Тогда

$$\text{(П.9)}\qquad \sup_{a_s\in A_G^{(s)}} |s^{-1}\ln f(X_s, a_s)-L(X_s, a_s)| \xrightarrow[s\to\infty]{} 0.$$

в. Пусть $A_{G,\varepsilon}^{(s)}$ — множество гипотез a_s, таких, что $\mathbf{y}_{ms}\in G$; $p_{ms}\geqslant\varepsilon$ $(m=1, \ldots, M)$, и выполнено условие 1. Тогда

$$\text{(П.10)}\qquad \sup_{a_s\in A_{G,\varepsilon}^{(s)}} |s^{-1}\ln f(X_s, a_s)-L(X_s, a_s)| \xrightarrow[s\to\infty]{} 0.$$

Доказательство. Соотношение (П.8) означает, что при $s\to\infty$

$$\sup\left| s^{-1}\ln\int\limits_{\Theta^M}\exp\left\{s\sum_m p_m g(\mathbf{y}_m, \theta)\right\}p(d\xi)-\sum_m p_m g(\mathbf{y}_m)\right|\to 0,$$

где sup берется по $\mathbf{y}_1\in G, \ldots, \mathbf{y}_M\in G$, $\alpha=(p_1, \ldots, p_M)\in A_{M-1}$.

Пусть $\mathbf{y}\to\theta(\mathbf{y})$ — отображение, определенное в разделе 3; тогда $\Theta_1=\theta(G)$ — компактное множество, $\Theta_1\subset\operatorname{Int}\Theta$. Ясно, что для некоторого $\beta>0$ существует компактное множество $\Theta_2\subset\operatorname{Int}\Theta$, представляющее собой β-окрестность множества Θ_1.

Обозначим $D_\delta(\mathbf{y})=\{\theta:|\theta-\theta(\mathbf{y})|<\delta\}$. Из равномерной непрерывности функции $g(\mathbf{y}, \theta)$ при $\mathbf{y}\in G$, $\theta\in\Theta_2$ следует, что для любого $\varepsilon>0$ найдется δ $(\beta>\delta>0)$ такое, что

$$\sup_{\mathbf{y}\in G,\,\theta\in D_\delta(\mathbf{y})} |g(\mathbf{y})-g(\mathbf{y}, \theta)|<\varepsilon.$$

Аналогично доказательству теоремы приложения 1, ограничивая область интегрирования множеством $D_\delta(\mathbf{y}_1)\times\ldots\times D_\delta(\mathbf{y}_M)$, можно получить неравенства

$$\sum_m p_m g(\mathbf{y}_m)\geqslant s^{-1}\ln\int\limits_{\Theta^M}\exp\left\{s\sum_m p_m g(\mathbf{y}_m, \theta)\right\}p(d\xi)\geqslant$$

$$\geqslant\sum_m p_m(g(\mathbf{y}_m)-\varepsilon)+s^{-1}\ln p(D_\delta(\mathbf{y}_1)\times\ldots\times D_\delta(\mathbf{y}_M)).$$

Из компактности множества $\Theta_2{}^M=\Theta_2\times\ldots\times\Theta_2$ и условия 1 нетрудно получить, что

$$\min_{\mathbf{y}_1\in,\ldots,\mathbf{y}_M\in G} p(D_\delta(\mathbf{y}_1)\times\ldots\times D_\delta(\mathbf{y}_M))>0,$$

откуда вытекает утверждение «а».

Утверждение «б» вытекает из соотношения (П.6) и утверждения «а».

Доказательство утверждения «в» аналогично доказательству соотношения «б».

Замечание. Более детальный анализ показывает, что левые части (П.8) и (П.10) представляют собой величины порядка соответственно

$$(2s)^{-1}R\sum_{m:s_m>0}\ln s_m+O(s^{-1});$$

$$(2s)^{-1}\left[R\sum_m\ln s_m+(M-1)\ln s\right]+O(s^{-1}).$$

Следствие 1. Пусть выполнено условие 1; последовательность $\{a_s\}_{s=1}^\infty$ такова, что существуют пределы $\mathbf{y}_m=\lim\limits_{s\to\infty} y_{m,s}\in G$. Тогда справедливо (5). Если, кроме того, существуют пределы $p_m{}^*=\lim\limits_{s\to\infty} p_{ms}>0$ или выполнено условие 2, то справедливо (6).

Это утверждение вытекает из того, что множество Y открыто.

Следствие 2. Пусть выполнены условия 1, 3, причем для $r=1, \ldots, R$

$$\int_E |y^{(r)}(\mathbf{x})| H_m(d\mathbf{x}) < \infty, \qquad \mathbf{y}_m = \int_E \mathbf{y}(\mathbf{x}) H_m(d\mathbf{x}) \in Y.$$

Тогда с вероятностью 1 справедливо (5). Если, кроме того, $p_m^* > 0$ или выполнено условие 2, то с вероятностью 1 справедливо (6).

Действительно, в условиях следствия 2 с вероятностью 1 $\mathbf{y}_{ms} \to \mathbf{y}_m$.

Следствие 3. Пусть $X=\{\mathbf{x}_i\}_{i=1}^{\infty}$ — выборка независимых векторов с распределением $H(d\mathbf{x})$, $\{a_s\}_{s=1}^{\infty}$ — последовательность гипотез, порожденная разбиением $e=(E_1, \ldots, E_M)$ пространства признаков E, $H(E_m)>0$ для $r=1, \ldots, R$

$$\int_E |y^{(r)}(\mathbf{x})| H(d\mathbf{x}) < \infty; \qquad \mathbf{y}_m = (H(E_m))^{-1} \int_{E_m} \mathbf{y}(\mathbf{x}) H(d\mathbf{x}) \in Y.$$

Тогда с вероятностью 1 справедливы (5), (6), (17).

Это вытекает из следствия 2 и непрерывности функции $g(\mathbf{y})$ по $\mathbf{y} \in Y$.

Следствие 4. Пусть выборка $X=\{\mathbf{x}_i\}_{i=1}^{\infty}$ такова, что $\mathbf{y}(\mathbf{x}_i) \in G \subset Y$, где G — выпуклое компактное множество, и выполнено условие 1. Тогда

а) $\sup\limits_{a_s} |s^{-1} \ln f(X_s | a_s) - L(X_s | a_s)| \xrightarrow[s\to\infty]{} 0;$

б) если, кроме того, выполнено условие 2, то
$\sup\limits_{a_s} |s^{-1} \ln f(X_s, a_s) - L(X_s, a_s)| \xrightarrow[s\to\infty]{} 0.$

Это утверждение непосредственно вытекает из теоремы 2.

ПРИЛОЖЕНИЕ III

Приведем явный вид функций $g(\mathbf{y})$ и $I(\mathbf{y}_1, \mathbf{y}_2)$ для некоторых важных экспоненциальных семейств.

1. Семейство R-мерных нормальных распределений с неизвестным вектором средних $\boldsymbol{\mu}$ и известной невырожденной ковариационной матрицей K. В этом случае $\mathbf{y}(\mathbf{x})=\mathbf{x}$; Y — R-мерное векторное пространство, $\boldsymbol{\theta}=K^{-1}\boldsymbol{\mu}$,

$$g(\mathbf{y}) = {}^1/_2 \mathbf{y}^T K^{-1} \mathbf{y} + C, \qquad I(\mathbf{y}_1, \mathbf{y}_2) = {}^1/_2 (\mathbf{y}_1 - \mathbf{y}_2) K^{-1} (\mathbf{y}_1 - \mathbf{y}_2).$$

Здесь и далее C — слагаемое, не зависящее от $\mathbf{y}$, T обозначает транспонирование.

Мера расстояния в этом случае совпадает с расстоянием Махаланобиса в пространстве признаков.

2. Семейство мультиноминальных распределений с неизвестными вероятностями p_r $(r=1, \ldots, R+1)$ исходов в серии из N независимых испытаний, $y^{(r)}(\mathbf{x})$ — количество r-х исходов в серии $\mathbf{x}$,

$$g(\mathbf{y}) = \sum_r y^{(r)} \ln y^{(r)} + C; \qquad I(\mathbf{y}_1, \mathbf{y}_2) = \sum_r y_2^{(r)} (\ln y_2^{(r)} - \ln y_1^{(r)}).$$

Мера компактности выборки в этом случае характеризуется энтропией эмпирической гистограммы.

3. Семейство N-мерных нормальных распределений с нулевыми математическими ожиданиями и неизвестной ковариационной матрицей K. В этом случае $R=N(N+1)/2$, $\mathbf{y}(\mathbf{x})=\mathbf{x}\mathbf{x}^T$, множество Y можно отождествить с множеством $\{H\}$ положительно определенных симметрических матриц,

$$g(\mathbf{y}) = -{}^1/_2 \ln \det H + C, \qquad I(\mathbf{y}_1, \mathbf{y}_2) = {}^1/_2 (\operatorname{Tr} H_1^{-1} H_2 - \ln \det H_1^{-1} H_2 - N).$$

4. Семейство N-мерных нормальных распределений с неизвестными вектором математических ожиданий $\boldsymbol{\mu}$ и ковариационной матрицей K; $\boldsymbol{\theta}=(K^{-1}\boldsymbol{\mu}, K^{-1})$. В этом случае $\mathbf{y}(\mathbf{x})=(\mathbf{x}, \mathbf{x}\mathbf{x}^T)$; вектор $\mathbf{y} \in Y$ можно представить в виде $\mathbf{y}=(\mathbf{m}, H)$, где $\mathbf{m}$ — R-мерный вектор, H — симметрическая матрица, такая, что матрица $\Sigma = H - \mathbf{m}\mathbf{m}^T$ положительно определена,

$$g(\mathbf{y}) = -{}^1/_2 \ln \det \Sigma + C; \quad I(\mathbf{y}_1, \mathbf{y}_2) =$$
$$= {}^1/_2 [(\mathbf{m}_1 - \mathbf{m}_2)^T \Sigma_1^{-1} (\mathbf{m}_1 - \mathbf{m}_2) + \operatorname{Tr} \Sigma_1^{-1} \Sigma_2 - \ln \det \Sigma_1^{-1} \Sigma_2 - N].$$

Мера компактности выборки в этом случае определяется величиной определителя эмпирической ковариационной матрицы. Подобный критерий компактности рассматривался, например, в [2, 10].

Доказательство соотношений (19), (20) ($M = \infty$)

Пусть $a_s^{(1)}$ — гипотеза, при которой все векторы относятся к одному классу (например, к первому); $a_s^{(s)}$ — гипотеза, при которой каждый вектор относится к своему классу. Для любой гипотезы с любым числом классов, как легко видеть, выполняется соотношение

$$L(X_s|a_s^{(s)}) \geqslant L(X_s|a_s) \geqslant L(X_s|a_s^{(1)}) = L(X_s, a_s^{(1)}).$$

Пусть существует постоянная $c>0$, не зависящая от s и такая, что $L(X_s|a_s^{(s)}) - L(X_s|a_s^{(1)}) < c$. Для экспоненциального семейства (8) это означает при $\mathbf{y}(\mathbf{x}_i) \in Y$, $\mathbf{y}_s \in Y$, что информационный разброс выборки $\{\mathbf{y}(\mathbf{x}_i)\}$ ограничен: $s^{-1} \sum_{i=1}^{s} I(\mathbf{y}_s, \mathbf{y}(\mathbf{x}_i)) < c$. Пусть также для последовательности $\{a_s^{(1)}\}$ выполнено соотношение (6). Это условие фактически относится к случаю конечного числа классов, и поэтому применимы результаты приложений 1 и 2.

Обозначим $A^{(s)}$ — множество гипотез $\{a_s\}$ с любым числом классов, удовлетворяющих условию

$$s^{-1} \ln f(X_s, a_s) \geqslant s^{-1} \ln f(X_s, a_s^{(1)})$$

(такие гипотезы назовем «достаточно хорошими»).

Для $a_s \in A^{(s)}$ при достаточно больших s и при $\varepsilon>0$

$$L(X_s, a_s) \geqslant s^{-1} \ln f(X_s, a_s) \geqslant L(X_s, a_s^{(1)}) - \varepsilon.$$

Отсюда можно получить, что при $a_s \in A^{(s)}$ при достаточно больших s

(П.11) $$E(a_s) \geqslant -a, \qquad a = \varepsilon + c > 0.$$

Множество $A^{(s)}$ содержит, в частности, гипотезы, оптимизирующие (1) и (15).

Пусть $K(R, a_s)$ — общее количество векторов для гипотезы a_s, которые лежат в классах, содержащих не более R векторов; $M(a_s)$ — количество непустых классов для гипотезы a_s.

Теорема 3:

(П.12) $$\sup_{a_s \in A^{(s)}} s^{-1} K(R, a_s) \leqslant a/\ln(s/R);$$

(П.13) $$\sup_{a_s \in A^{(s)}} s^{-1} M(a_s) \leqslant a/\ln s + 2\exp(a)/s.$$

Соотношения (19), (20) непосредственно вытекают из теоремы 3.

Доказательство теоремы 3. Введем обозначения. Пусть k_{is} — количество классов для гипотезы a_s, в которые попадает по i векторов, $i=1, \ldots, s$, $\beta_{is} = ik_{is}/s$, $\boldsymbol{\beta}_s = (\beta_{1s}, \ldots, \beta_{ss})$. Ясно, что

(П.14) $$0 \leqslant \beta_{is} \leqslant 1, \qquad \sum_{i=1}^{s} \beta_{is} = 1$$

для любой гипотезы a_s.

Неравенство (П.11) эквивалентно следующему:

(П.15) $$\sum_{i=1}^{s} \beta_{is} \ln(s/i) \leqslant a.$$

При этом

$$s^{-1}K(R, a_s) = K(\beta_s, R) = \sum_{i=1}^{R} \beta_{is};$$

(П.16)

$$s^{-1}M(a_s) = M(\beta_s) = \sum_{i=1}^{s} \beta_{is}/i.$$

Докажем соотношение (П.12). Из (П.15) имеем

$$a \geqslant \sum_{i=1}^{R} \beta_{is} \ln(s/i) \geqslant \sum_{i=1}^{R} \beta_{is} \ln(s/R) = \ln(s/R) K(\beta_s, R),$$

откуда следует (П.12).

Докажем соотношение (П.13), которое вытекает из следующего:

(П.17) $$\sup_{\beta_s \in D_s} M(\beta_s) \leqslant a/\ln s + 2\exp(a)/s,$$

где множество D_s определяется условиями (П.14), (П.15) и представляет собой выпуклый многогранник, образованный пересечением симплекса (П.14) и полупространства (П.15) в векторном пространстве размерности s.

Функция $M(\beta_s)$, определяемая в (П.16), является линейной и поэтому достигает максимума в одной из вершин многогранника D_s, которые могут быть точками двух типов:

а) вершины α_i симплекса (П.14):

(П.18) $$\alpha_i = (0, \ldots, 0, 1, 0, \ldots, 0)$$

(единица на i-м месте); $1 \leqslant i \leqslant s$;

б) вершины α_{ij}, образованные пересечением гиперплоскости, соответствующей равенству в (П.15), с одномерными ребрами δ_{ij} симплекса (П.14) $(1 \leqslant i < j \leqslant s)$:

(П.19) $$\delta_{ij} = \{0, \ldots, 0, p, 0, \ldots, 0, q, 0, \ldots, 0\},$$
$$p+q=1, \quad p \geqslant 0, \quad q \geqslant 0$$

(p и q на i-м и j-м местах соответственно).

Для вершин α_i, удовлетворяющих неравенству (П.15),

(П.20) $$\ln(s/i) \leqslant a; \quad i \geqslant s/\exp(a); \quad M(\alpha_i) = s/i \leqslant \exp(a)/s.$$

Для вершин α_{ij} имеем

(П.21) $$p \ln(s/i) + q \ln(s/j) = a,$$

откуда получаем неравенства $\max\{\ln(s/i), \ln(s/j)\} = \ln(s/i) \geqslant a$, $\min\{\ln(s/i), \ln(s/j)\} = \ln(s/j) \leqslant a$, и, следовательно,

(П.22) $$i \leqslant s/\exp(a); \quad j \geqslant s/\exp(a).$$

Кроме того, из (П.12) имеем неравенство $p \ln(s/i) \leqslant a$, откуда получаем, что

(П.23) $$p \leqslant a/\ln(s/i).$$

Из (П.22) и (П.23) для вершин α_{ij} имеем

(П.24) $$M(\alpha_{ij}) = \frac{p}{i} + \frac{q}{j} \leqslant \frac{a}{i} \ln \frac{s}{i} + \frac{\exp a}{s}.$$

Заметим теперь, что функция $i \ln(s/i)$ является вогнутой функцией аргумента i, и поэтому она достигает минимума на концах интервала его изменения. В рассматриваемом случае согласно (П.19) и (П.22) величина i принимает значения в интервале $[1, s/\exp a]$, и поэтому

(П.25) $$i \ln(s/i) \geqslant \min\{\ln s, as/\exp a\}.$$

Из (П.24) и (П.25) имеем

(П.26) $$M(\alpha_{ij}) \leqslant \max\{a/\ln s, \exp a/s\} + \exp a/s.$$

Из (П.20) и (П.26) вытекает (П.17).

Поступила в редакцию
22 ноября 1976 г.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

1. *Дорофеюк А. А.* Алгоритмы автоматической классификации. Автоматика и телемеханика, № 12, стр. 78—113, 1971.
2. *Айвазян С. А., Божаева З. И., Староверов О. В.* Классификация многомерных наблюдений. «Статистика», 1964.
3. *Цыпкин Я. З., Кельманс Г. К.* Рекуррентные алгоритмы самообучения. Техническая кибернетика, № 5, стр. 78—87, 1967.
4. *Пугачев В. С.* Статистические проблемы теории распознавания образов. Тр. III Всесоюз. совещания по автоматическому управлению, стр. 192—202. «Наука», 1967.
5. *Репин В. Г.* Об одном алгоритме распознавания в условиях априорной неопределенности. Техническая кибернетика, № 6, стр. 144—150, 1970.
6. *Большаков И. А., Левин Б. Р., Репин В. Г., Тартаковский Г. П.* Некоторые вопросы статистического синтеза информационных систем. Техническая кибернетика, № 2, стр. 153—170, 1970.
7. *Ченцов Н. Н.* Статистические решающие функции и оптимальные выводы. «Наука», 1973.
8. *Стратонович Р. Л.* Принципы адаптивного приема. «Сов. радио», 1973.
9. *Добровидов А. В.* Об одном алгоритме непараметрической оценки многомерных случайных сигналов. Автоматика и телемеханика, № 2, стр. 88—99, 1971.
10. *Scott A. J., Symon M. J.* Clastering Methods Based on Likelihood Ratio Criteria. Biometrics, v. 27, p. 387, 1971.
11. *Шлезингер М. И.* О самопроизвольном различении образов. В сб. «Читающие автоматы», стр. 38—45. «Наукова думка», 1965.
12. *Ингстер Ю. И.* Рекуррентные алгоритмы самообучения распознаванию образов. В сб. «Вопросы кибернетики. Адаптивные системы», стр. 162—167. Изд. Научн. совета по комплексной проблеме «Кибернетика», АН СССР, М., 1974.
13. *Уилкс С.* Математическая статистика. «Наука», 1967.
14. *Ле Кам.* О некоторых асимптотических свойствах оценок максимального правдоподобия и соответствующих байесовых оценок. В сб. «Математика», 4:2, стр. 69—119, 1960.
15. *Рао С. Р.* Линейные статистические методы и их применения. «Наука», 1968.

BAYES METHODS IN COMPUTER-AIDED CLASSIFICATION

Yu. I. INGSTER

The paper is concerned with statistical performance criteria in computer-aided classification, or learning. The relation of statistical and variational methods in this problem is established. The case of an unknown number of classes is taken up.