УДК 681.3.06:621.391

МЕТОД РАСПОЗНАВАНИЯ, ОСНОВАННЫЙ НА АППРОКСИМАЦИИ ВЫБОРОК СМЕСЬЮ НОРМАЛЬНЫХ ЗАКОНОВ

Г.Я.Волошин, С.Т.Косенкова

Для реализации стратегии Байеса при распознавании необходимо иметь распределения вероятностей значений параметров (признаков) каждого образа.

Рассмотрим методику построения распределений по выборкам (обучающим последовательностям), основанную на использовании смеси нормальных законов с диагональными ковариационными матрицами

$$\rho(\bar{x}/H_j) \simeq \sum_{i=1}^{N} \alpha_i \varphi_{ij}(\bar{x}, n_{ij}), \qquad (1)$$

где $\bar{x} = \{x_1, x_2, \ldots, x_m\}$ – текущее значение координат выборочного m-мерного пространства; α_i – весовые коэффициенты, удовлетворяющие условию $\sum_{i=1}^{N} \alpha_i = 1$; $\varphi_{ij}(\bar{x}, n_{ij})$ – нормальное распределение с оценочной диагональной ковариационной матрицей и вектором средних, приходящихся на i-ю аппроксимационную компоненту в j-й их смеси; H_j – обозначение j-го образа.

Выбор для аппроксимации различных распределений смеси гауссовых законов стимулируется следующими обстоятельствами:

– для оценочных параметров гауссовых законов, являющихся случайными величинами, известны функции распределения;

– для распределения, описываемого конечной совокупностью гауссовых законов, нетрудно построить датчик случайных векторов, что весьма важно для процесса обучения и оценки вероятности ошибок распознавания;

– гауссовы законы при определенных условиях [1] образуют полную систему функций в пространстве L_2 , то есть могут быть использованы для представления достаточно широкого класса распределений.

Тот факт, что $\varphi_{ij}(x, n_{ij})$ описываются диагональными ковариационными матрицами, имеет особое значение для реализации последовательных процедур распознавания. Дело в том, что последовательные процедуры, легко реализуемые при независимых признаках (последовательных измерениях), не часто встречаются на практике.

При использовании представления (1) "независимость" последовательных измерений (точнее, их мультипликативность) обеспечивается следующим образом. Пусть имеются две гипотезы H_0 и H_1 с априорными вероятностями P_0 и P_1 . Тогда из-за диагональности ковариационных матриц законов $\varphi_{ij}(x, n_{ij})$ отношение правдоподобия после измерений вычисляется по формуле:

$$\lambda(x_1, x_2, \ldots, x_r) = \frac{P_0 \sum_{i=1}^{N_0} \alpha_i \prod_{k=1}^{r} \varphi_{i0}(x_k, n_{i0})}{P_1 \sum_{j=1}^{N_1} \beta_j \prod_{k=1}^{r} \varphi_{j1}(x_k, n_{j1})},$$

где N_0 – количество компонент в смеси, описывающей распределение H_0 ; N_1 – то же для H_1 .

Таким образом, с помощью (1) как для независимых, так и для зависимых признаков можно реализовать мультипликативную последовательность измерений без традиционного громоздкого декорреллирующего преобразования выборочного пространства.

Рассмотрим вопрос о зависимости оценочных параметров распределений $\varphi_{ij}(\bar{x}, n_{ij})$ от объема выборки n_{ij} (для простоты нижние индексы будем опускать). Пусть дан нормальный m-мерный закон с диагональной ковариационной матрицей

$$f(\bar{x}, \bar{\mu}, \bar{\sigma}) = \prod_{q=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma_q} \exp\left[-\frac{1}{2\sigma_q^2}(x_q - \bar{\mu}_q)^2\right], \qquad (2)$$

где $\bar{\mu} = \{\mu_1, \ldots, \mu_m\}$ и $\bar{\sigma} = \{\sigma_1, \ldots, \sigma_m\}$ неизвестны, но известны их оценки

$$\hat{\mu}_q = \frac{1}{n} \sum_{i=1}^{n} x_{qi}; \qquad s_q^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{qi} - \hat{\mu}_q)^2.$$

Поскольку в (2) истинные значения $\bar{\mu}$ и $\bar{\sigma}$ неизвестны, то мы можем формально говорить об их плотности вероятности.

Определим математическое ожидание функции (2)

$$\tilde{\varphi}(\bar{x},n)=\int_{\mathcal{D}} f(\bar{x},\bar{\mu},\bar{\sigma})\,\psi(\bar{\mu},\bar{\sigma}/\bar{\tilde{\mu}},\bar{\tilde{s}},n)\,d\bar{\mu}\,d\bar{\sigma},$$

где $\psi(\bar{\mu},\bar{\sigma}/\bar{\tilde{\mu}},\bar{\tilde{s}},n)$ – совместное распределение истинных $\bar{\mu}$ и $\bar{\sigma}$ при условии известных оценок $\bar{\tilde{\mu}}$ и $\bar{\tilde{s}}$; $\mathcal{D}$ – область существования $\bar{\mu}$ и $\bar{\sigma}$.

В соответствии с формулой Байеса

$$\psi(\bar{\mu},\bar{\sigma}/\bar{\tilde{\mu}},\bar{\tilde{s}},n)=\frac{\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)\,\rho(\bar{\mu},\bar{\sigma})}{\int_{\mathcal{D}}\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)\,\rho(\bar{\mu},\bar{\sigma})\,d\bar{\mu}\,d\bar{\sigma}},$$

где $\rho(\bar{\mu},\bar{\sigma})$ – априорная совместная плотность распределения неизвестных $\bar{\mu}$ и $\bar{\sigma}$. В том случае, если $\rho(\bar{\mu},\bar{\sigma})$ неизвестно, можно исходить из самых плохих предположений – равномерной плотности по всему полупространству $(\bar{\mu},\bar{\sigma})$,то есть:

$$\tilde{\varphi}(x,n)=\lim_{R,Q\to\infty}\frac{\int_{(R,Q)} f(\bar{\mu},\bar{\sigma},x)\,\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)\,\frac{1}{V(R,Q)}\,d\bar{\mu}\,d\bar{\sigma}}{\int_{(R,Q)}\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)\,\frac{1}{V(R,Q)}\,d\bar{\mu}\,d\bar{\sigma}}=$$

$$=\int_{\Phi} f(\bar{\mu},\bar{\sigma},\bar{x})\,\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)\,d\bar{\mu}\,d\bar{\sigma},$$

где $V(R,Q)$ – объем гиперпараллелограмма (R,Q), ограниченного координатами $(-r_1,\dots,-r_m,r_1,\dots,r_m;\,0,q_1,\dots,q_m)$; R – область определения $\bar{\mu}$; Q – область определения $\bar{\sigma}$; Φ – полупространство всех возможных значений $\bar{\mu}$ и $\bar{\sigma}$.

Известно [2], что для выборки из нормальной совокупности

$$\rho(\bar{\tilde{\mu}},\bar{\tilde{s}}/\bar{\mu},\bar{\sigma},n)=\rho_1(\bar{\tilde{\mu}}/\bar{\mu},\bar{\sigma},n)\,\rho_2(\bar{\tilde{s}}/\bar{\sigma},n),$$

причем величина $\Delta\mu_q=\mu_q-\tilde{\mu}_q$ распределена по нормальному закону с нулевым средним и частной дисперсией $\sigma_q/\sqrt{n}$. После усреднения по $\Delta\mu_q$ получим (при неизвестных σ_q)

$$\tilde{\varphi}_{\sigma}(\bar{x},n)=\prod_{q=1}^{m}\frac{\sqrt{n}}{2\pi\sigma_q^2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma_q^2}\left[(x_q-\tilde{\mu}_q+\Delta\mu_q)^2+n\Delta\mu_q^2\right]\right\}d\Delta\mu_q=$$

$$\prod_{q=1}^{m}\frac{\sqrt{n}}{\sqrt{2\pi(n+1)}\sigma_q}\exp\left\{-\frac{n(x_q-\tilde{\mu}_q)^2}{2(n+1)\sigma_q^2}\right\}. \qquad (3)$$

Теперь учтем тот факт, что дисперсии признаков неизвестны и их тоже приходится оценивать по выборке. Известно [2], что оценка ковариационной матрицы выборки из нормальной совокупности связана с распределением Уишарта. Для диагональных ковариационных матриц это распределение имеет вид:

$$g(\{u_{ij}\})=|u_{ij}|\left(\frac{1}{2}\right)^{n-m+2}\prod_{q=1}^{m}\frac{\exp(-u_q^2/2\sigma_q^2)}{2^{\frac{n-1}{2}}\,\pi^{\frac{m-1}{4}}\,\Gamma\left(\frac{n-q}{2}\right)\sigma_q^{n-1}},$$

где

$$u_{ij}=\sum_{\alpha=1}^{n}(x_{i\alpha}-\tilde{\mu}_i)(x_{j\alpha}-\tilde{\mu}_j),\quad u_q^2=u_{ii}=(n-1)s_q^2.$$

Если оцениваются только диагональные элементы u_q^2 ,то,воспользовавшись теоремой 7.3.4 из [3],можно записать

$$g(\{u_q^2\})=\prod_{q=1}^{m}\left\{\frac{u_q^{n-3}\exp(-u_q^2/2\sigma_q^2)}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma_q^{n-1}}\right\}.$$

При замене переменных по правилу $t_q=u_q^2/(n-1)\sigma_q^2$ $(u_q\neq 0)$, получим

$$g(\{t_q\})=\prod_{q=1}^{m}\left\{\frac{\left(\frac{n-1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}\,t_q^{\frac{n-3}{2}}\,e^{-\frac{n-1}{2}t_q}\right\}.$$

Легко видеть, что последнее выражение есть не что иное, как декартово произведение одномерных распределений χ^2 с n степенями свободы. При усреднении (3) по t_q получим следующее распределение:

$$\tilde{\varphi}(\bar{x},n)=\prod_{q=1}^{m}\left\{\frac{\sqrt{n}\left(\frac{n-1}{2}\right)^{\frac{n-1}{2}}}{\sqrt{2\pi(n+1)}\,s_q\,\Gamma\left(\frac{n-1}{2}\right)}\int_0^{\infty}t_q^{\frac{n-1}{2}}\exp\left[-\frac{n(x_q-\tilde{\mu}_q)^2}{(n^2-1)s_q^2}t_q-\frac{n-1}{2}t_q\right]dt_q\right\}=$$

$$\prod_{q=1}^{m}\left\{\sqrt{\frac{n}{\pi(n^2-1)}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)s_q}\left[\frac{n(x_q-\tilde{\mu}_q)^2}{(n^2-1)s_q^2}+1\right]^{-\frac{n}{2}}\right\}. \qquad (4)$$

Практически пользоваться распределением (4) неудобно из-за его громоздкости. Поэтому целесообразно рассмотреть вопрос о приемлемой его аппроксимации. Для этого прежде всего изучим асимптотические свойства $\tilde{\varphi}(\bar{x}, n)$.

Справедлива следующая

ТЕОРЕМА. При $n \to \infty$ распределение (4) сходится к нормальному с математическим ожиданием $\tilde{\mu}$ и дисперсией $\tilde{s}$.

ДОКАЗАТЕЛЬСТВО. Обозначим через d_q величину $\frac{(x_q - \tilde{\mu}_q)^2}{s_q^2}$ и рассмотрим предел

$$\lim_{n\to\infty} \sqrt{\frac{n}{\pi(n^2-1)}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{s_q\,\Gamma\left(\frac{n-1}{2}\right)}\left[1+\frac{d_q}{n-\frac{1}{n}}\right]^{-\frac{n}{2}} = \tilde{\varphi}_\infty(x_q).$$

Умножив и разделив это выражение на величину $\frac{\sqrt{n-1}}{2}$ и проделав элементарные преобразования, получим

$$\tilde{\varphi}_\infty(x_q) = \frac{1}{\sqrt{2\pi}\,s_q}\lim_{n\to\infty}\sqrt{\frac{n}{n+1}}\lim_{n\to\infty}\frac{\Gamma\left(\frac{n-1}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}e^{-\frac{1}{2}\ln\left(\frac{n-1}{2}\right)}\lim_{n\to\infty}\left(1+\frac{d_q}{n-\frac{1}{n}}\right)^{-\frac{n}{2}}.$$

Используя известное соотношение [4]

$$\lim_{|z|\to\infty}\frac{\Gamma(z+\alpha)}{\Gamma(z)}\,e^{-\alpha\ln z} = 1,$$

можно записать

$$\tilde{\varphi}_\infty(x_q) = \frac{1}{\sqrt{2\pi}\,s_q}\lim_{n\to\infty}\left(1+\frac{d_q}{n-\frac{1}{n}}\right)^{-\frac{n}{2}}.$$

Для определения оставшегося предела используем следующее известное свойство. Пусть $\varphi_n^1(x) \ge \tilde{\varphi}_n(x) \ge \varphi_n^2(x)$ и $\lim_{n\to\infty}\varphi_n^1(x) = \lim_{n\to\infty}\varphi_n^2(x) = \alpha$. Тогда $\lim_{n\to\infty}\tilde{\varphi}_n(x) = \alpha$.

Выберем

$$\varphi_n^1(x_q) = \left(1+\frac{d_q}{n}\right)^{-\frac{n}{2}};\quad \varphi_n^2(x_q) = \left(1+\frac{d_q}{n-1}\right)^{-\frac{n}{2}}.$$

При любом фиксированном значении d_q для $\varphi_n^1(x_q)$ предел при $n \to \infty$ равен $\exp(-d_q/2)$.

Перейдем к $\varphi_n^2(x_q)$.

$$\lim_{n\to\infty}\left(1+\frac{d_q}{n-1}\right)^{-\frac{n}{2}} = \lim_{n\to\infty}\left(1+\frac{d_q}{n-1}\right)^{-\frac{n}{2}}\cdot\lim_{n\to\infty}\left(1+\frac{d_q}{n-1}\right)^{-\frac{1}{2}} = \exp\left(-\frac{d_q}{2}\right)$$

Таким образом, $\lim_{n\to\infty}\tilde{\varphi}_n(x_q) = \exp\left(-\frac{d_q}{2}\right)$ и

$$\tilde{\varphi}_\infty(x_q) = \frac{1}{\sqrt{2\pi}\,s_q}\,e^{-\frac{(x_q-\tilde{\mu}_q)^2}{2s_q^2}}$$

Следовательно,

$$\tilde{\varphi}_\infty(x) = \prod_{q=1}^{m}\left\{\frac{1}{\sqrt{2\pi}\,s_q}\,e^{-\frac{(x_q-\tilde{\mu}_q)^2}{2s_q^2}}\right\}$$

Теорема доказана.

Анализ распределения (4) показывает, что математическое ожидание имеет координаты $\tilde{\mu}_q$, распределение симметрично по всем осям относительно математического ожидания, частные дисперсии равны $s_q\sqrt{\frac{n^2-1}{n(n-3)}}$. Аппроксимация распределения (4) нормальным законом с математическим ожиданием $\tilde{\mu}$ и дисперсиями по осям $s_q\sqrt{\frac{n^2-1}{n(n-3)}}$ оказывается хорошей при $n > 6 \div 8$. При $2 \le n \le 8$ более подходящим является нормальный закон с частными дисперсиями $s_q\frac{n+1}{n-1}$. На рис. I кривая I показывает зависимость от объема выборки множителя к s_q гипотетического нормального закона, имеющего такую дифференциальную энтропию, что и распределение (4). Кривая 2 – зависимость множителя $\sqrt{\frac{n^2-1}{n(n-3)}}$ от n, кривая – 3 то же для множителя $\frac{n+1}{n-1}$.

Таким образом, распределение (4) целесообразно аппроксимировать нормальным законом с математическим ожиданием $\tilde{\mu}$ и частными дисперсиями:

$$\sigma_q^* = s_q\cdot f(n),\ \text{где}\ f(n) = \begin{cases}\sqrt{\frac{n^2-1}{n(n-3)}} & \text{при}\quad n > 6,\\ \frac{n+1}{n-1} & \text{при}\quad 2 \le n \le 6.\end{cases} \qquad (5)$$

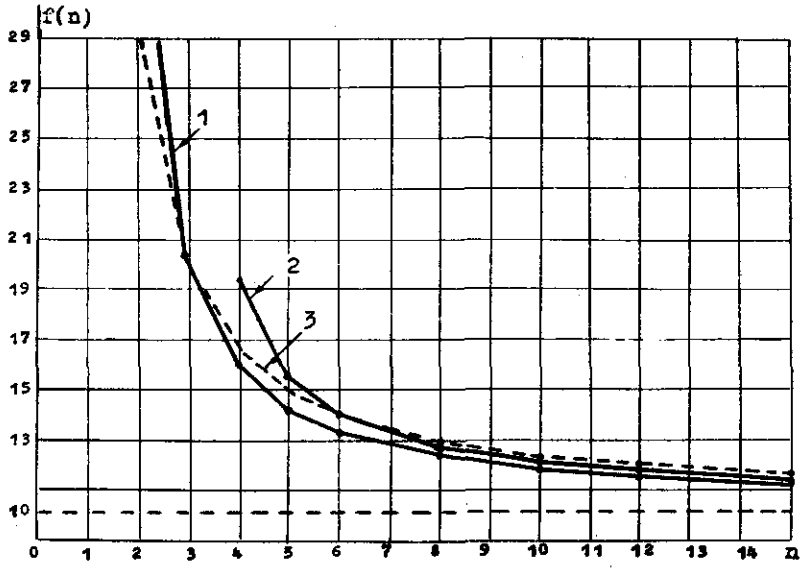

Рис.I. Зависимость поправочного множителя $f(n)$ к оценочным дисперсиям от объема выборки

При выработке этой рекомендации нами использована в качестве критерия "похожести" распределений их дифференциальная энтропия. Такой выбор обусловлен тем, что минимум дифференциальной энтропии использован для определения некоторых параметров смеси (I), описывающей обучающую последовательность (см.ниже).

Теперь вернемся к описанию обучающей последовательности смесью (I) нормальных законов с диагональными ковариационными матрицами. Для определения параметров смеси α_i, $\vec{\mu}_i$ и $\vec{\sigma}_i$ по выборке могут быть использованы алгоритмы стохастической аппроксимации (при фиксированном N-числе компонент в смеси (I)). Нами использован алгоритм, предложенный Шлезингером [5], критерием оценивания в котором является достижение максимума выражения

$$L(A) = \sum_{j=1}^{n} \log P(\vec{x}_j / A),$$

где A – оцениваемый векторный параметр $(\alpha_i, \vec{\mu}_i, \vec{\sigma}_i; i = 1,2,\dots,N)$.

Полученные в результате работы алгоритма значения частных дисперсий корректируются в соответствии с (5), причем

$$f_i(n) = \begin{cases} \sqrt{\dfrac{\alpha_i^2 n^2 - 1}{\alpha_i n(\alpha_i n - 3)}} & \text{при} \quad \alpha_i n > 6, \\ \dfrac{\alpha_i n + 1}{\alpha_i n - 1} & \text{при} \quad 2 < \alpha_i n \le 6, \end{cases}$$

где n – общий объем выборки.

Здесь величина $\alpha_i n$ соответствует объему подвыборки, участвующей в определении параметров i-й компоненты смеси (I) (условие принадлежности смеси j-му образу для простоты опускаем).

Особо следует остановится на определении N-числа компонент в смеси. В задачах самообучения, к которым фактически относится рассматриваемый алгоритм оценки векторного параметра A, число таксонов N определяется субъективно потребителем резуль-

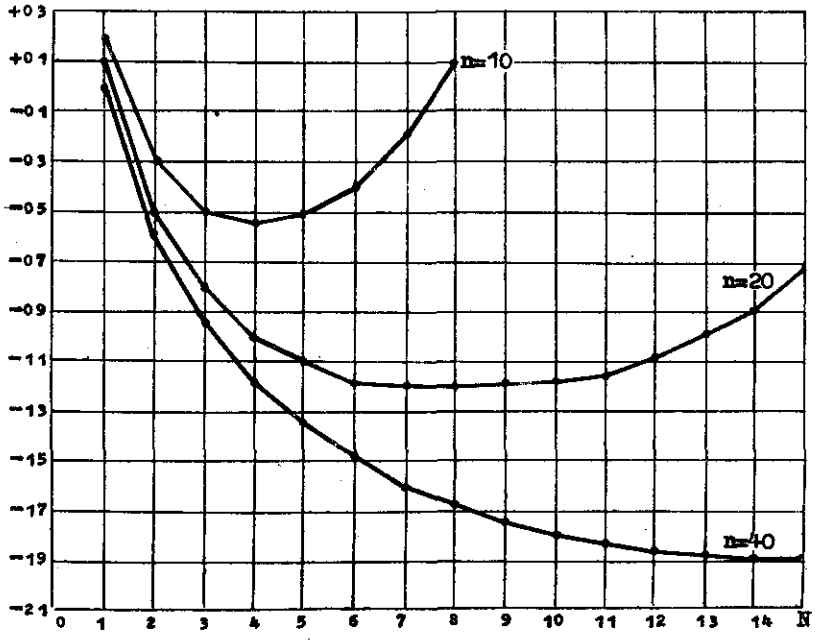

Рис.2. Зависимость дифференциальной энтропии от числа компонент в смеси нормальных законов

татов. Ему предлагается набор решений с разными N, из которых выбирается (часто на интуитивном уровне) наиболее подходящий. Мы предлагаем формализовать процесс выбора N_{opt} следующим образом. Проводится последовательное оценивание при N=I,2,3... При этом осуществляется вычисление (например, методом Монте - Карло) дифференциальной энтропии полученного распределения. При последовательном увеличении N имеют место вообще говоря, две тенденции:

- уменьшение энтропии за счет рационального дробления выборки на части с уменьшающимися дисперсиями;

- увеличение энтропии за счет уменьшения объемов подвыборок и связанного с этим увеличения множителей $f_i(n)$ (5) к оценочным дисперсиям.

Наличие этих двух тенденций обусловливает существование N_{opt} по критерию минимума дифференциальной энтропии. Для иллюстрации этих соображений приведем один из примеров работы алгоритма аппроксимации выборки смесью нормальных законов.

В качестве исходного материала брались равномерно распределенные в интервале 0-I случайные числа. Эта выборка аппроксимировалась смесью нормальных законов, и вычислялась дифференциальная энтропия. Математические ожидания значений энтропии как функции N представлены на рис.2 (при разных объемах выборки). Совершенно отчетливо проявляются наличие N_{opt}, а также возможность все более детального (точного) представления распределения выборки при увеличении n.

Литература

I. ДОРОДНОВ А.А. Ортонормированная система Гаусса.-Сб.аспирантских работ. Точные науки. Казань, Изд-во КГУ, I969.

2. УИЛКС С. Математическая статистика. М., "Наука", I967.

3. АНДЕРСОН Т. Введение в многомерный статистический анализ. М., Физматгиз, I963.

4. ГРАДШТЕЙН И.С., РЫЖИК И.М. Таблицы интегралов, сумм, рядов и произведений. М., Физматгиз, I963.

5. ШЛЕЗИНГЕР М.И. Взаимосвязь обучения и самообучения в распознавании образов. -"Кибернетика", Киев, I968, № 2.

Поступила в ред.-изд.отд.
27 апреля I973 года