ИНФОРМАТИКА

2008 октябрь-декабрь № 4

УДК 519.237: 681.3

В.И. Малюгин

СТАТИСТИЧЕСКИЙ АНАЛИЗ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ РЕГРЕССИОННЫХ НАБЛЮДЕНИЙ

Предлагается подстановочное байесовское решающее правило классификации выборки из смеси распределений регрессионных наблюдений. В основе решающего правила лежит EM-алгоритм, предназначенный для классификации наблюдений и вычисления оценок параметров смеси с помощью метода максимального правдоподобия. Доказывается сходимость предлагаемого алгоритма и исследуется вероятность ошибочной классификации на модельных данных.

Введение

Для вероятностного описания реальных процессов широко используются регрессионные модели [1, 2]. Обычно целью регрессионного анализа является оценивание функции регрессии по имеющейся выборке наблюдений. В то же время наличие неконтролируемых и неучтенных в регрессионных моделях факторов может привести к их структурной неоднородности, означающей существование не одной, а нескольких моделей зависимостей и соответственно классов наблюдений. Если каждое наблюдение относится с определенной вероятностью к одному из заданных классов, а номер класса при регистрации наблюдений не фиксируется, то исследователь имеет неоднородную и неклассифицированную выборку регрессионных наблюдений. Задачи оценивания параметров моделей и классификации наблюдений в таких условиях приходится решать одновременно.

Обычно предполагается, что вероятностной моделью наблюдений из одного класса выступает некоторое многомерное распределение вероятностей. Для классификации наблюдений при этом используются методы дискриминантного анализа смесей распределений (в случае параметрического задания моделей наблюдений) и кластерного анализа (в условиях непараметрической неопределенности) [3–6].

Особенностью рассматриваемой задачи является то, что для описания наблюдений используется модель статистической зависимости типа многомерной линейной регрессии. С учетом параметрического задания модели и наличия неклассифицированной выборки наблюдений рассматриваемую задачу можно отнести к задачам дискриминантного анализа смесей распределений регрессионных наблюдений. От традиционных задач дискриминантного анализа смесей распределений она отличается нарушением предположения относительно одинаковой распределенности наблюдений из одного класса, а также возможностью разбиения вектора наблюдений на подвекторы эндогенных и экзогенных переменных. Эндогенные переменные характеризуют свойства самих объектов исследования, а экзогенные переменные – контролируемые воздействия со стороны внешнего окружения. Актуальность алгоритмов анализа моделей статистических зависимостей в условиях структурной параметрической неоднородности в последнее время связана с разработкой автоматизированных систем ранжирования объектов в пространстве зависимых признаков, которые имеют место в экономических приложениях. Примером таких систем могут служить системы кредитного скоринга для классификации заемщиков коммерческих банков по степени надежности [7], системы оценки устойчивости коммерческих банков [8] и раннего предупреждения банковских кризисов [9].

В работе [10] были предложены и исследованы подстановочные байесовские решающие правила (ПБРП) классификации многомерных независимых регрессионных наблюдений, использующие статистические оценки параметров по классифицированной и неклассифицированной выборке. В статье [11] с помощью метода асимптотического разложения риска ПБРП найдены количественные оценки приращения риска (вероятности ошибки) классификации, обусловленные отсутствием классификации обучающей выборки.

1. Математическая модель наблюдений и постановка задач исследования

Пусть t-й (t =1,…,T) наблюдаемый объект описывается случайным вектором наблюдений $\mathbf{y}_t \in \Re^{N+M}$, который допускает разбиение на подвекторы:

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{pmatrix} \in \Re^{N+M},\ \mathbf{x}_t \in \Re^N,\ \mathbf{z}_t \in \mathbf{Z} \subset \Re^M, \tag{1}$$

где $\mathbf{x}_t = (x_{t1},...,x_{tN})'$ – вектор эндогенных переменных (признаков), характеризующих состояние анализируемого объекта; $\mathbf{z}_t = (z_{t1},...,z_{tM})'$ – вектор экзогенных переменных (факторов), описывающих внешние воздействия на состояние объектов.

Наблюдаемые объекты принадлежат к одному из двух классов объектов Ω_0 или Ω_1. Номер класса описывается ненаблюдаемой дискретной случайной величиной $d_t \in S = \{0,1\}$ (t =1,...,T) с распределением вероятностей:

$$P\{d_t = \alpha\} = \pi_\alpha > 0 (\alpha \in S), \pi_0 + \pi_1 = 1, \tag{2}$$

параметры $\{\pi_\alpha\}(\alpha \in S)$ соответствуют априорным вероятностям классов объектов.

Между подвекторами составного вектора наблюдений (1) существует статистическая зависимость, которая для класса d_t =α ($\alpha \in S$) описывается моделью многомерной линейной регрессии вида

$$\mathbf{x}_t = B_\alpha \mathbf{z}_t + \mathbf{v}_t,\ \alpha \in S = \{0,1\}, \tag{3}$$

где $B_\alpha = (b_{\alpha ij})$ – $(N \times M)$-матрица коэффициентов регрессии (причем $(B_0 - B_1)\mathbf{z} \neq \mathbf{0}$, $\forall \mathbf{z} \in \mathbf{Z},\ \mathbf{Z} \subset \Re^M$ – ограниченная замкнутая область); $\mathbf{v}_t = (v_{t1},\ldots,v_{tN}) \in \Re^N$ – независимые в совокупности случайные векторы (ошибки наблюдения), распределенные по нормальному закону с нулевым математическим ожиданием и невырожденной ковариационной $(N \times N)$-матрицей $\Sigma = (\sigma_{ij})$, т. е. $E\{\mathbf{v}_t\} = 0$, $E\{\mathbf{v}_t \mathbf{v}_\tau'\} = \delta_{t\tau}\Sigma$, где $\delta_{t\tau}$ – символ Кронекера.

Истинные значения параметров $\{\pi_\alpha\}, \{B_\alpha\}, \Sigma$ модели (2), (3) неизвестны. Имеется неклассифицированная обучающая выборка значений признаков $X = \{\mathbf{x}_1,...,\mathbf{x}_T\}$ объема T из классов $\Omega_0 \cup \Omega_1$, соответствующая последовательности значений факторов $Z = \{\mathbf{z}_1,...,\mathbf{z}_T\}$, где значение вектора признаков $\mathbf{x}_t \in \Re^N$ определяется на основании (3) для заданного значения вектора факторов $\mathbf{z}_t \in \mathbf{Z}$ (t =1,...,T).

В рамках рассматриваемой модели выборку регрессионных наблюдений $X = \{\mathbf{x}_1,...,\mathbf{x}_T\}$, соответствующую последовательности значений факторов $Z = \{\mathbf{z}_1,...,\mathbf{z}_T\}$, можно рассматривать как случайную выборку из смеси распределений регрессионных наблюдений, плотность распределения вероятностей для которой имеет вид

$$p_\pi(\mathbf{x};\mathbf{z},\boldsymbol{\theta}) = \pi_0 \varphi_N(\mathbf{x}|B_0\mathbf{z},\Sigma) + \pi_1 \varphi_N(\mathbf{x}|B_1\mathbf{z},\Sigma),\ \mathbf{x} \in \Re^N,\ \mathbf{z} \in \mathbf{Z}, \tag{4}$$

где $\boldsymbol{\theta} \in \Theta \subset \Re^m$ (m = 2MN + N(N+1)/2 + 1) – составной вектор параметров, образованный из независимых элементов матриц $\{B_\alpha\}$, Σ и априорной вероятности π_0 (π_1 =1–π_0);

$$\varphi_N(\mathbf{x}|B_\alpha\mathbf{z},\Sigma) = (2\pi)^{-N/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_t - B_\alpha\mathbf{z}_T)' \Sigma^{-1} (\mathbf{x}_t - B_\alpha\mathbf{z}_T)\right\} - \tag{5}$$

плотность распределения *N*-мерного нормального закона с математическим ожиданием $B_\alpha \mathbf{z}$ и невырожденной ковариационной матрицей Σ.

Имеют место следующие задачи анализа смеси (4) по неклассифицированной обучающей выборке $\{X, Z\}$:

1) статистическое оценивание вектора параметров $\boldsymbol{\theta} \in \Theta \subset \Re^m$ (вычисление оценок максимального правдоподобия (МП-оценок) $\{\hat{\pi}_\alpha\}, \{\hat{B}_\alpha\}, \hat{\Sigma}$);

2) классификация обучающей выборки $\{X, Z\}$, т. е. оценивание вектора классификации выборки $\mathbf{d} = (d_t) \in S^T$;

3) классификация вновь поступающих наблюдений $(\mathbf{x}_\tau, \mathbf{z}_\tau)$, $\tau = T+1, ..., T+n,\ n \geq 1$.

При решении указанных задач следует учитывать следующие особенности модели наблюдений:

– наблюдения $\mathbf{x}_t \in \Re^N$ над объектами из одного и того же класса, соответствующие различным значениям факторов $\mathbf{z}_t \in \mathbf{Z}$ $(t = 1, ..., T)$, являются неоднородными по среднему значению, зависящему от $\mathbf{z}_t$;

– если факторы являются управляемыми, то существует возможность минимизации вероятности ошибки классификации за счет оптимального задания значений факторов $\mathbf{z}_t \in \mathbf{Z}$ $(t = 1, ..., T)$.

Для решения задач 1, 2 предлагается алгоритм расщепления смеси распределений регрессионных наблюдений (4) из класса *EM*-алгоритмов, позволяющий одновременно вычислять МП-оценки параметров $\{\hat{\pi}_\alpha\}$, $\{\hat{B}_\alpha\}$, $\hat{\Sigma}$ и осуществлять классификацию обучающей выборки $\{X, Z\}$. Приводятся также результаты экспериментального исследования вероятности ошибки ПБРП классификации регрессионных наблюдений, использующего найденные МП-оценки $\{\hat{\pi}_\alpha\}, \{\hat{B}_\alpha\}, \hat{\Sigma}$.

2. Представления для оценок максимального правдоподобия параметров смеси

Получим представления для МП-оценок параметров смеси распределений регрессионных наблюдений (4), зависящие от апостериорных вероятностей $p_{\alpha t}(\mathbf{x}_t; \mathbf{z}_t)$ отнесения регрессионного наблюдения $(\mathbf{x}_t, \mathbf{z}_t)$ к классу Ω_α:

$$p_{\alpha t} \equiv p_{\alpha t}(\mathbf{x}_t, \mathbf{z}_t) = \frac{\pi_\alpha \varphi_N(\mathbf{x}_t \,|\, B_\alpha \mathbf{z}_t,\ \Sigma)}{p_\pi(\mathbf{x}_t; \mathbf{z}_t, \boldsymbol{\theta})},\ \mathbf{x}_t \in \Re^N,\ \mathbf{z}_t \in \mathbf{Z},\ t = 1, ..., T,\ \alpha \in S\ . \tag{6}$$

Очевидно, что для апостериорных вероятностей (6) выполняются следующие условия:

$$p_{\alpha t} \geq 0 \ \text{ и } \sum_{\alpha \in S} p_{\alpha t} = 1\,,\ t = 1, ..., T,\ \alpha \in S. \tag{7}$$

Для вычисления апостериорных вероятностей (6) по заданным значениям параметров смеси $\{\pi_\alpha\}, \{B_\alpha\}, \Sigma$ с учетом (4), (5) справедлива формула

$$p_{\alpha t}(\mathbf{x}_t, \mathbf{z}_t) = \frac{\exp\left(g'_\alpha \mathbf{x}_t + h_\alpha\right)}{\sum_{\alpha \in S} \exp\left(g'_\alpha \mathbf{x}_t + h_\alpha\right)},\ \mathbf{x}_t \in \Re^N, \tag{8}$$

$$g_\alpha = \Sigma^{-1} B_\alpha \mathbf{z}_t,\ h_\alpha = \frac{1}{2}\mathbf{z}_t' B_\alpha' \Sigma^{-1} B_\alpha \mathbf{z}_t + \ln(\pi_\alpha),\ \mathbf{z}_t \in \mathbf{Z},\ t = 1,...,T,\ \alpha \in S.$$

Теорема 1. *Если апостериорные вероятности* $\{p_{\alpha t}\}$ $(t = 1,...,T,\ \alpha \in S)$ *известны, то МП-оценки* $\{\hat{\pi}_\alpha\}, \{\hat{B}_\alpha\}, \hat{\Sigma}$ *параметров смеси (4) по неклассифицированной выборке регрессионных наблюдений* $X = \{\mathbf{x}_1,...,\mathbf{x}_T\}$, $Z = \{\mathbf{z}_1,...,\mathbf{z}_T\}$ *(*$\mathbf{x}_t \in \Re^N$, $\mathbf{z}_t \in \mathbf{Z}$*) из классов* $\Omega_0 \cup \Omega_1$ *допускают представления*

$$\hat{B}_\alpha = \sum_{t=1}^{T} p_{\alpha,t} \mathbf{x}_t \mathbf{z}_t' \left(\sum_{t=1}^{T} p_{\alpha,t} \mathbf{z}_t \mathbf{z}_t' \right)^{-1}; \tag{9}$$

$$\hat{\Sigma} = \sum_{\alpha=0}^{1} \sum_{t=1}^{T} p_{\alpha,t} \left(\mathbf{x}_t - \hat{B}_\alpha \mathbf{z}_T \right) \left(\mathbf{x}_t - \hat{B}_\alpha \mathbf{z}_T \right)'; \tag{10}$$

$$\hat{\pi}_0 = \frac{1}{T} \sum_{t=1}^{T} p_{0,t} \left(\hat{\pi}_1 = 1 - \hat{\pi}_0 \right). \tag{11}$$

Доказательство. Из независимости $\mathbf{x}_1,...,\mathbf{x}_T$ при заданных значениях $\mathbf{z}_1,...,\mathbf{z}_T$, а также формулы (4) следует выражение для логарифмической функции правдоподобия выборки регрессионных наблюдений $X = \{\mathbf{x}_1,...,\mathbf{x}_T\} \in \Re^{NT}$, $Z = \{\mathbf{z}_1,...,\mathbf{z}_T\} \in \mathbf{Z}^T$:

$$l_N(\boldsymbol{\theta}; X, Z) = \sum_{t=1}^{T} \ln \left(\sum_{\alpha \in S} \pi_\alpha \varphi_N(\mathbf{x}_t | B_\alpha \mathbf{z}_t, \Sigma) \right). \tag{12}$$

С учетом (6), (7) для логарифмической функции правдоподобия (12) справедливо

$$\begin{aligned} l_N(\boldsymbol{\theta}; X, Z) &= \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \sum_{t=1}^{T} \ln \left(\frac{\pi_\alpha \varphi_N(\mathbf{x}_t | B_\alpha \mathbf{z}_t, \Sigma)}{p_{\alpha t}} \right) = \\ &= \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \ln(\pi_\alpha) + \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \ln\left(\varphi_N(\mathbf{x}_t | B_\alpha \mathbf{z}_t, \Sigma) \right) - \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \ln(p_{\alpha t}). \end{aligned} \tag{13}$$

Искомые представления для параметров смеси $\{\pi_\alpha\}, \{B_\alpha\}, \Sigma$ найдем из условия максимума целевой функции (13). Очевидно, что задача максимизации функции (13) по $\{\pi_\alpha\}$ и $\{B_\alpha\}, \Sigma$ распадается на две независимые задачи максимизации непрерывных по параметрам функций: задачу максимизации первого слагаемого по $\{\pi_\alpha\}$ вида

$$\sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \ln(\pi_\alpha) \Rightarrow \max \tag{14}$$

и задачу максимизации второго слагаемого в (13) по $\{B_\alpha\}, \Sigma$ вида

$$\sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha t} \ln\left(\varphi_N(\mathbf{x}_t | B_\alpha \mathbf{z}_t, \Sigma) \right) \Rightarrow \max. \tag{15}$$

Дифференцируя целевую функцию в задаче (14) по π_α и приравнивая производную к нулю, находим представление (11). С учетом (5) и (7) задача максимизации по $\{B_\alpha\}, \Sigma$ целевой функции в (15) эквивалентна задаче максимизации по $\{B_\alpha\}, \Sigma^{-1}$ функции

$$l\left(\{B_\alpha\},\Sigma^{-1}\right)=\frac{1}{2}\ln\left|\Sigma^{-1}\right|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{\alpha\in S}\sum_{t=1}^{T}p_{\alpha t}\left(\mathbf{x}_t-B_\alpha\mathbf{z}_t\right)\left(\mathbf{x}_t-B_\alpha\mathbf{z}_t\right)'\right). \qquad (16)$$

Накладывая на матричные производные функции $l\left(\{B_\alpha\},\Sigma^{-1}\right)$ по $\{B_\alpha\},\Sigma^{-1}$ нулевые ограничения, получаем систему уравнений правдоподобия относительно $\{B_\alpha\},\Sigma^{-1}$:

$$\frac{\partial l\left(\{B_\alpha\},\Sigma^{-1}\right)}{\partial B_\alpha}=\Sigma^{-1}\sum_{\alpha\in S}\sum_{t=1}^{T}p_{\alpha t}\left(\mathbf{x}_t-\hat{B}_\alpha\mathbf{z}_t\right)\mathbf{z}_t'=\mathbf{O}_{N\times M}\ ;$$

$$\frac{\partial l\left(\{\hat{B}_\alpha\},\Sigma^{-1}\right)}{\partial\Sigma^{-1}}=\frac{T}{2}\Sigma-\frac{1}{2}\sum_{\alpha\in S}\sum_{t=1}^{T}p_{\alpha t}\left(\mathbf{x}_t-\hat{B}_\alpha\mathbf{z}_t\right)\left(\mathbf{x}_t-\hat{B}_\alpha\mathbf{z}_t\right)'=\mathbf{O}_{N\times N}\ ,$$

из которых находим представления (9), (10) (здесь $\mathbf{O}_{N\times M}$ – нулевая ($N\times M$)-матрица). ■

3. Алгоритм расщепления смеси распределений регрессионных наблюдений

Для решения задач 1, 2 дискриминантного анализа смеси (4) по неклассифицированной выборке наблюдений $\{X,Z\}$ объема T предлагается использовать итерационный алгоритм последовательного уточнения оценок вектора параметров $\boldsymbol{\theta}\in\Theta\subset\Re^m$ смеси (4), образованного из $\{\pi_\alpha\},\{B_\alpha\},\Sigma$, и вектора классификации $\mathbf{d}=(d_t)\in S^T$ выборки $\{X,Z\}$. Данный алгоритм относится к классу *EM*-алгоритмов, широко применяемых в задачах статистического оценивания параметров в условиях априорной неопределенности [3]. Каждая итерация предлагаемого *EM*-алгоритма включает два последовательно выполняемых этапа:

– *E (Estimation*): оценивание при заданных начальных значениях параметров модели $\boldsymbol{\theta}$ векторов апостериорных вероятностей классов $\left\{\mathbf{p}_\alpha=\left(p_{\alpha,t}(\mathbf{x}_t,\mathbf{z}_T)\right)\right\}$ (очевидно, $\mathbf{p}_1=\mathbf{1}_N-\mathbf{p}_0$, где $\mathbf{1}_T$ – единичный T-вектор), знание которых позволяет оценить вектор классификации выборки $\mathbf{d}=(d_t)\in S^T$;

– *M (Maximization)*: нахождение оценок параметров смеси $\boldsymbol{\theta}$ из условия максимума логарифмической функции правдоподобия на основании полученных ранее апостериорных вероятностей классов $\{\mathbf{p}_\alpha\}$.

Если задаются начальные значения вектора классификации выборки $\mathbf{d}$, то меняется последовательность выполнения этапов. Работа алгоритма продолжается до достижения заданного условия остановки.

3.1. Описание алгоритма

Для обозначения номера итерации будем использовать верхний индекс k в скобках при соответствующих параметрах и переменных. Опишем действия, которые выполняются на начальном шаге и k (k = 1, 2, …) последующих итерациях алгоритма.

Шаг 0. Задаются начальные значения оценок параметров $\{\mathbf{p}_\alpha\}$, такие, что $p_\pi(\mathbf{x}_t;\mathbf{z}_t,\boldsymbol{\theta})>0$ (t=1,…, T).

Шаг k.1. По формуле (8) с использованием оценок $\{\pi_\alpha^{(k-1)}\},\{B_\alpha^{(k-1)}\},\Sigma^{(k-1)}$ вычисляются апостериорные вероятности классов $\mathbf{p}_\alpha^{(k)}=\left(p_{\alpha,t}^{(k)}(\mathbf{x}_t,\mathbf{z}_T)\right)$, $\alpha\in S$, и при необходимости осуществляется классификация выборки регрессионных наблюдений $X=\{\mathbf{x}_1,...,\mathbf{x}_T\}\in\Re^{NT}$, $Z=\{\mathbf{z}_1,...,\mathbf{z}_T\}\in\mathbf{Z}^T$, т. е. находится оценка вектора классификации $\mathbf{d}^{(k)}=(d_t^{(k)})\in S^T$ по формуле

$$d_t^{(k)} = \operatorname{argmax}\left\{ p_{\alpha,t}^{(k)}(\mathbf{x}_t, \mathbf{z}_T) \right\} \text{ по } \alpha \in S. \tag{17}$$

Шаг k.2. По формулам (9)–(11) вычисляются оценки параметров $\{\pi_\alpha^k\}, \{B_\alpha^k\}, \Sigma^{(k)}$.

Шаг k.3. Проверяется условие остановки алгоритма. Если компоненты составного вектора параметров удовлетворяют условию близости значений параметров на соседних итерациях: $\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\right\| \le \varepsilon$, $0 < \varepsilon \ll 1$, то полагается, что $\hat{\pi}_\alpha = \pi_\alpha^{(k-1)}$, $\hat{B}_\alpha = B_\alpha^{(k-1)}$, $\hat{\Sigma} = \Sigma^{(k-1)}$, и работа алгоритма прекращается. В противном случае полагается *k*:= *k*+1 и осуществляется переход к шагу *k*.1 на новой итерации.

3.2. Доказательство сходимости алгоритма

Существует значительное число публикаций, посвященное исследованию сходимости алгоритмов расщепления смесей распределений вероятностей, основанных на *EM*-алгоритме (см., например, обзоры [3, 13]). Для доказательства сходимости предлагаемого алгоритма воспользуемся традиционной схемой доказательства, применяемой в случае смесей распределений вероятностей [3].

Пусть $\boldsymbol{\theta}^{(k)}$, $\mathbf{p}_0^{(k)}$ и $l_N^{(k)}$ – значения вектора параметров, вектора апостериорных вероятностей и логарифмической функции правдоподобия по выборке регрессионных наблюдений $X = \{\mathbf{x}_1, ..., \mathbf{x}_T\} \in \Re^{NT}$, $Z = \{\mathbf{z}_1, ..., \mathbf{z}_T\} \in \mathbf{Z}^T$ на *k*-й итерации описанного выше алгоритма расщепления смеси (4).

Теорема 2. *Последовательность значений $l_N^{(k)}$, соответствующих итерациям k = 1, 2, ... алгоритма расщепления смеси (4), не убывает, т. е. справедливо*

$$l_N^{(k)} - l_N^{(k-1)} \ge 0, \; k = 1, 2, \ldots, \tag{18}$$

причем равенство в (18) имеет место тогда и только тогда, когда $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$ и $\mathbf{p}_0^{(k)} = \mathbf{p}_0^{(k-1)}$.

Доказательство. Применяя представление (13) для значений логарифмической функции правдоподобия на итерациях алгоритма, получаем

$$l_N^{(k)} - l_N^{(k-1)} = T \sum_{\alpha \in S} \pi_\alpha^{(k)} \ln\left(\frac{\pi_\alpha^{(k)}}{\pi_\alpha^{(k-1)}} \right) + \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha,t}^{(k)} \ln\left(\frac{\varphi_N(\mathbf{x}_t | B_\alpha^{(k)} \mathbf{z}_t, \Sigma^{(k)})}{\varphi_N(\mathbf{x}_t | B_\alpha^{(k-1)} \mathbf{z}_t, \Sigma^{(k-1)})} \right) + \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha,t}^{(k)} \ln\left(\frac{p_{\alpha,t}^{(k-1)}}{p_{\alpha,t}^{(k)}} \right). \tag{19}$$

Так как значения параметров $\left\{B_\alpha^{(k)}\right\}$, $\Sigma^{(k)}$ на *k*-й итерации согласно (15) максимизируют выражение

$$\sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha,t}^{(k)} \ln\left(\varphi_N(\mathbf{x}_t | B_\alpha \mathbf{z}_t, \Sigma) \right)$$

по возможным значениям параметров $\{B_\alpha\}, \Sigma$, то последнее слагаемое в (19) неотрицательно. Следовательно,

$$l_N^{(k)} - l_N^{(k-1)} \ge T \sum_{\alpha \in S} \pi_\alpha^{(k)} \ln\left(\frac{\pi_\alpha^{(k)}}{\pi_\alpha^{(k-1)}} \right) + \sum_{\alpha \in S} \sum_{t=1}^{T} p_{\alpha,t}^{(k)} \ln\left(\frac{p_{\alpha,t}^{(k-1)}}{p_{\alpha,t}^{(k)}} \right). \tag{20}$$

Последовательности $\{a_l\}$, $\{b_l\}$ $(l = 1, ..., L)$, такие, что $\sum_{l=1}^{L} a_l = \sum_{l=1}^{L} b_l = 1$, удовлетворяют неравенству [13].

$$\sum_{l=1}^{L} a_l \ln \frac{a_l}{b_l} \geq \frac{1}{4} \sum_{l=1}^{L} \left(a_l - b_l\right)^2 \geq 0. \qquad (21)$$

Поэтому слагаемые в правой части (20) неотрицательны, а значит $l_N^{(k)} - l_N^{(k-1)} \geq 0$, причем $l_N^{(k)} - l_N^{(k-1)} = 0$ только в том случае, если $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$ и $\mathbf{p}_0^{(k)} = \mathbf{p}_0^{(k-1)}$. ■

Покажем, что описанный алгоритм обеспечивает нахождение МП-оценок неизвестных параметров (под МП-оценкой понимается любое решение системы уравнений правдоподобия, т. е. любой локальный максимум логарифмической функции правдоподобия).

Следствие 1. *В условиях теоремы 2 справедливы утверждения:*

I) последовательность $\left\{l_N^{(k)}\right\}$ *при* $k \to \infty$ *сходится к одному из локальных максимумов функции правдоподобия* $l_N(\boldsymbol{\theta}; X, Z)$ *вида (12);*

2) соответствующая последовательность $\{\boldsymbol{\theta}^{(k)}\}$ *при* $k \to \infty$ *сходится к точке, доставляющей этот локальный максимум.*

Доказательство. По определению смеси (4) $|\Sigma| \neq 0$ и, следовательно, функция правдоподобия (12) ограничена. Поскольку последовательность $\left\{l_N^{(k)}\right\}$ $\left(k = 0,1,2,...\right)$ не убывает, то справедливо первое утверждение. В силу (20) и неравенства (21) имеем соотношение

$$l_N^{(k)} - l_N^{(k-1)} \geq \frac{1}{4} \sum_{\alpha \in S} \left(\pi_\alpha^{(k)} - \pi_\alpha^{(k-1)}\right)^2 + \frac{1}{4} \sum_{\alpha \in S} \sum_{t=1}^{T} \left(p_{\alpha,t}^{(k-1)} - p_{\alpha,t}^{(k)}\right)^2.$$

Так как левая часть в последнем выражении стремится к нулю при $k \to \infty$, то $\pi_\alpha^{(k)} - \pi_\alpha^{(k-1)} \to 0$, $p_{\alpha,t}^{(k)} - p_{\alpha,t}^{(k-1)} \to 0$ при $k \to 0$ $\left(t = 1,...,T,\ \alpha \in S\right)$, что влечет справедливость второго утверждения. ■

Повторяя вычисления при различных значениях параметров, можно найти все локальные (а также глобальный) максимумы функции правдоподобия.

4. Исследование точности классификации регрессионных наблюдений

Пусть $\{\hat{\pi}_\alpha\}$, $\{\hat{B}_\alpha\}$, $\hat{\Sigma}$, $\left\{\hat{p}_{\alpha t}(\mathbf{x}_t, \mathbf{z}_t)\right\}$ – найденные в результате работы описанного выше алгоритма МП-оценки соответствующих параметров по неклассифицированной выборке регрессионных наблюдений $X = \{\mathbf{x}_1,...,\mathbf{x}_T\} \in \Re^{NT}$, $Z = \{\mathbf{z}_1,...,\mathbf{z}_T\} \in \mathbf{Z}^T$. Для классификации новых наблюдений $(\mathbf{x}_\tau, \mathbf{z}_\tau)$ $(\tau = T+1,...,T+n,\ n \geq 1)$, описываемых моделью (2), (3), можно использовать два следующих эквивалентных в смысле вероятности ошибки решающих правила.

1. Правило вида (17) классификации по максимуму апостериорной вероятности:

$$d_\tau = \operatorname{argmax}\left\{\hat{p}_{\alpha,t}(\mathbf{x}_t, \mathbf{z}_T)\right\} \text{ по } \alpha \in S.$$

2. ПБРП, получающееся из оптимального в смысле минимума вероятности ошибки байесовского решающего правила подстановкой в него состоятельных оценок неизвестных параметров [10]:

$$d_\tau = d\left(\mathbf{x}_\tau, \mathbf{z}_\tau\right) = \begin{cases} 0, & \lambda_\tau\left(\mathbf{x}_\tau, \mathbf{z}_\tau\right) < 0; \\ 1, & \lambda_\tau\left(\mathbf{x}_\tau, \mathbf{z}_\tau\right) \geq 0, \end{cases} \qquad (22)$$

с линейной дискриминантной функцией

$$\lambda_\tau(\mathbf{x}_\tau, \mathbf{z}_\tau) = \mathbf{z}'\left(\hat{B}_1 - \hat{B}_0\right)' \hat{\Sigma}^{-1}\mathbf{x}_\tau - \frac{1}{2}\mathbf{z}'\left(\hat{B}_1 + \hat{B}_0\right)' \hat{\Sigma}^{-1}\left(\hat{B}_1 - \hat{B}_0\right) - \ln\frac{\hat{\pi}_0}{\hat{\pi}_1}.$$

Как показано в [10, 11], асимптотические свойства оценок $\{\hat{\pi}_\alpha\}$, $\{\hat{B}_\alpha\}$, $\hat{\Sigma}$ и, следовательно, величина ошибки классификации ПБРП, использующего эти оценки, при увеличении объема обучающей выборки T существенно зависят от значений факторов (плана регрессионных экспериментов) $Z = \{\mathbf{z}_1, ..., \mathbf{z}_T\}$. Там же с помощью метода асимптотического разложения риска ПБРП (22) найдены количественные оценки приращения риска (вероятности ошибки) классификации, обусловленные отсутствием классификации обучающей выборки. Показано, что увеличение ошибки классификации тем больше, чем сильнее пересечение классов (меньше межклассовое расстояние $\Delta_\tau = \mathbf{z}_\tau\left(B_1 - B_0\right)' \Sigma^{-1}\left(B_1 - B_0\right)\mathbf{z}_\tau$) и меньше объем выборки T.

В таблице приведены результаты численных экспериментов, иллюстрирующие точность классификации с помощью решающего правила (22), использующего оценки параметров по неклассифицированной выборке для следующих тестовых примеров:

Тестовый пример 1. T = 50, n = 50, N = 1, M = 2; истинные значения параметров: $\pi_0 = 0,5$, $\sigma_{11} = 0,09$, $B_0 = (1,\ 1)$, $B_1 = (-1,\ 1)$; начальные значения параметров: $\pi_0^{(0)} = 0,3$, $\sigma_{11}^{(0)} = 1$, $B_0^{(0)} = (0,5;\ 0,5)$, $B_1^{(0)} = (0,3;\ 0,5)$.

Тестовый пример 2. T = 50, n = 50, N = 1, M = 3; истинные значения параметров: $\pi_0 = 0,5$, $\sigma_{11} = 0,09$, $B_0 = (1,\ 1,\ 1)$, $B_1 = (-1,\ 1,\ 1)$; начальные значения параметров: $\pi_0^{(0)} = 0,3$, $\sigma_{11}^{(0)} = 1$, $B_0^{(0)} = (0,5;\ 0,5,\ -0,8)$, $B_1^{(0)} = (0,3;\ 0,5;\ 0,8)$.

Тестовый пример 3. T = 50, n = 50, N = 2, M = 2; истинные значения параметров:

$$\pi_0 = 0,5,\quad B_0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ B_0 = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix},\ \Sigma = \begin{pmatrix} 0,09 & 0 \\ 0 & 0,09 \end{pmatrix};$$

начальные значения параметров:

$$\pi_0^{(0)} = 0,3,\quad B_0^{(0)} = \begin{pmatrix} 0,4 & 0,5 \\ 0,8 & 0,4 \end{pmatrix},\ B_1^{(0)} = \begin{pmatrix} 0,5 & 0,7 \\ -0,3 & 0,5 \end{pmatrix},\ \Sigma = \begin{pmatrix} 0,25 & 0 \\ 0 & 0,64 \end{pmatrix}.$$

Значения компонент вектора $\mathbf{z}_t \in \mathbf{Z} \subset \Re^M\ (M = 1, 2, 3)$ моделировались по равномерному закону на интервале [–1, 1].

Результаты анализа точности классификации

Оценки вероятности ошибки	Тестовые примеры		
	1	2	3
Точечная оценка для обучающей выборки	0,000	0,083	0,000
Точечная оценка для контрольной выборки	0,125	0,178	0,142
Интервальная 95%-я оценка вероятности ошибки	[0,059–0,218]	[0,102–0,289]	[0,077–0,253]

Заключение

Целесообразность применения предлагаемого алгоритма обусловлена следующими проблемами анализа данных:

1. Во многих сферах человеческой деятельности возникают задачи обработки больших массивов многомерных данных. При этом часто анализируемые в процессе наблюдения за объектами признаки являются взаимно зависимыми, а сами объекты принадлежат к различным классам (различаются размером, типом, условиями функционирования и т. п.). В этих случаях данные в

пространстве признаков имеют сложную неоднородную (кластерную) структуру. Для обнаружения таких особенностей данных на этапе их предварительного статистического анализа часто применяются различные алгоритмы из класса *data mining* (добычи данных) [14]. Описанный выше алгоритм может использоваться в указанных случаях как один из подобных алгоритмов с целью проверки сразу двух предположений: о наличии кластерной структуры данных и статистической зависимости применяемых признаков.

2. При использовании методов дискриминантного анализа многомерных данных в слабоформализованных областях может возникать проблема формировании классифицированной обучающей выборки наблюдений в силу нестрогого описания классов объектов. В подобных случаях на этапе формирования обучающей выборки наблюдений часто применяются экспертные оценки при отнесении наблюдений к различным классам на основе автономного анализа признаков. При этом не учитывается существующая между признаками статистическая зависимость. Предлагаемый алгоритм позволяет получить предварительное разбиение неоднородных данных в пространстве статистически зависимых признаков.

С обеими проблемами приходится сталкиваться, в частности, при разработке алгоритмов кредитного скоринга, предназначенных для классификации заемщиков коммерческих банков по степени их платежеспособности [7].

Применение предложенного алгоритма расщепления смесей многомерных регрессионных наблюдений приводит к установлению статистических зависимостей, адекватность которых может быть исследована с помощью стандартного набора тестов, применяемых для проверки адекватности регрессионных моделей, включая анализ статистической значимости коэффициентов регрессии, анализ остатков и т. д. [1, 2]. Примеры использования указанных тестов в задачах дискриминантного анализа многомерных регрессионных наблюдений описаны в [15].

Список литературы

1. Айвазян, С.А. Прикладная статистика. Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М. : Финансы и статистика, 1985. – 487 с.

2. Харин, Ю.С. Эконометрическое моделирование / Ю.С. Харин, В.И. Малюгин, А.Ю. Харин. – Минск : БГУ, 2003. – 313 с.

3. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян [и др.]. – М. : Финансы и статистика, 1989. – 607 с.

4. McLachlan, G.J. Discriminant Analysis and Statistical Pattern Recognition / G.J. McLachlan. – N.-Y. : Wiley, 1992. – 562 p.

5. Харин, Ю.С. Робастность в статистическом распознавании образов / Ю.С. Харин. – Минск : Университетское, 1992. – 232 с.

6. Жук, Е.Е. Устойчивость в кластер-анализе многомерных наблюдений / Е.Е. Жук, Ю.С. Харин. – Минск : БГУ, 1998. – 240 с.

7. Гринь, Н.В. Исследование точности методов классификации многомерных данных в задачах кредитного скоринга / Н.В. Гринь, В.И. Малюгин // Вестник ГрГУ. Сер. 2. – 2008. – № 1. – С. 77–85.

8. Малюгин, В.И. Оценка устойчивости коммерческих банков на основе эконометрических моделей с дискретными зависимыми переменными / В.И. Малюгин, Е.В. Пытляк // Банковский вестник. – 2007. – № 4 (369). – С. 30–36.

9. Egorov, A.A. Analysis of the banking crises by the panel logit model with the application to Belarusian banking system / A.A. Egorov, V.I. Malugin // Proc. of the 8th Intern. Conf. «Computer Data Analysis and Modeling». – Vol. 2. – Minsk : BSU, 2007. – P. 68-71.

10. Малюгин, В.И. Об оптимальности классификации случайных наблюдений, различающихся уравнениями регрессии / В.И. Малюгин, Ю.С. Харин // Автоматика и телемеханика. – 1986. – № 7. – С. 35–46.

11. Малюгин, В.И. Кластер-анализ регрессионных наблюдений / В.И. Малюгин // Статистическая обработка экспериментальных данных : межвуз. сб. науч. тр. ; Новосибирский эл.-техн. ин-т. – Новосибирск, 1986. – С. 43–52.

12. Redner, L. Mixture densities, maximum likelihood and the EM algorithm / L. Redner, J. Walker // SIAM Review. – 1984. – Vol. 26, № 2. – P. 23–31.

13. Кульбак, С. Теория информации и статистика / С. Кульбак. – М.: Наука, 1967. – 408 с.

14. Berry, M.J. A. Data mining techniques: for marketing, sales, and customer relationship management / M. J.A. Berry, G. Linoff. – N.-Y. : Wiley, 2004. – 643 p.

15. Малюгин, В.И. Дискриминантный анализ многомерных автокоррелированных регрессионных наблюдений в условиях параметрической неоднородности моделей / В.И. Малюгин // Информатика. – 2008. – № 3. – С. 17–28.

Поступила 03.05.08

НИИ прикладных проблем математики и информатики
Белорусского государственного университета,
Минск, пр. Независимости, 4
e-mail: Malugin@bsu.by

V.I. Malugin

STATISTICAL ANALYSIS OF THE MIXTURES OF DISTRIBUTIONS OF REGRESSION OBSERVATIONS

The «plug-in» decision rule for classification of a sample from the mixture of the multivariate distributions of regression observations is suggested. The decision rule is based on the EM-algorithm designed both for classification of observations and for estimation of mixture parameters on the basis of the maximum likelihood estimation method. The convergence of the suggested algorithm is proved and the classification error probability is examined on simulated data.