Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

НОВЫЕ РЕЗУЛЬТАТЫ В МАТЕМАТИЧЕСКОЙ ТЕХНИКЕ ИЗУЧЕНИЯ ХРОНИЧЕСКИХ БОЛЕЗНЕЙ*

Л. Д. МЕШАЛКИН

В докладе, который можно рассматривать как продолжение ряда предшествующих выступлений (Мешалкин (1970, 1972), дается обзор двух направлений исследований, получивших в последние годы значительное развитие. Первое из них — это асимптотическая теория классификации многомерных объектов в условиях дефицита выборочных данных. Толчком к ее развитию послужило то обстоятельство, что в последние годы центр тяжести в применении формул классификации переместился с диагностики четко определенных состояний, как это было в 1962—1965 годах, в область использования прогнозирования для выделения переменных, наиболее тесно связанных с течением болезни, оценки их прогностической силы и построения на их базе системы контроля за состоянием больного. С математической точки зрения этот переход означал, что стали рассматриваться задачи дискриминации тесно пересекающихся многомерных статистических совокупностей, в которых вероятности ошибки при отнесении индивидуального наблюдения к той или иной совокупности велики, а центр тяжести исследования лежет в оценке силы разделения.

Второе направление связано с классическими задачами определения пределов физиологической „нормы“ и оценки влияния, зависимости между переменными. Оно требовало разработки методов устойчивой к отклонениям от гауссовости параметризации многомерных распределений и регрессионных зависимостей; такого видоизменения процедур ковариационного и факторного анализа, при котором отдельные „дикие“ наблюдения мало влияют на интерпретацию данных. Применяемая здесь техника получила название λ-моментов и λ-регрессии.

1. Асимптотическое исследование вероятностей ошибочной классификации. В этом параграфе под термином „классификация“ мы должны понимать отнесение обьекта по наблюдаемым признакам к одной из двух статистических совокупностей, заданных обучающими выборками и априорной информацией о виде распределений. Мы будем исследовать задачу классификации в условиях дефицита выборочных данных, когда объемы выборок делаются сравнимы с числом оцениваемых параметров. При этом будет использоваться предложенный А. Н. Колмогоровым прием рассматривать не одну изолированную задачу классификации, а последовательность (по $m \to \infty$) классификационных задач, в которых число оцениваемых параметров, размерность выборочного пространства и объемы выборок, используемых для обучения, растут неограниченно с ростом m.

С тем, чтобы лучше почувствовать новую асимптотику, рассмотрим случай двух p-мерных нормальных распределений с общей известной единичной ковариационной матрицей и неизвестными векторами средних μ_1 и μ_2. Для оценки средних будем использовать обучающие выборки обьемов n_1 и n_2. Предположим далее, что

---

* Доклад, прочитанный на 12-ом Европейском совещании статистиков, Варна, 3—7 сентября 1979 г.

СЕРДИКА Българско математическо списание. Том 6, 1980, с. 391—396.

(1) $$p/n_i \to \lambda_i < \infty \quad (i=1,2);$$

(2) $$(\mu_1-\mu_2)^T(\mu_1-\mu_2) \to J < \infty;$$

и классификатор строится путем замены в логарифме отношения плотностей неизвестных параметров их оценками максимального правдоподобия, т. е

$$g(x) = -(x-\widehat{\mu_1})^T(x-\widehat{\mu_1})/2 + (x-\widehat{\mu_2})^T(x-\widehat{\mu_2})/2 \gtrless 0,$$

где x — классифицируемое наблюдение.

Пусть E_i $(i=1,2)$ — символ условного математического ожидания при условии, что x принадлежит i-ой совокупности. Легко проверяется, что

(3) $$\lim_{n\to\infty}(E_1 g(x) - E_2 g(x)) = J, \quad \lim_{n\to\infty} E_i(g(x)-E_i g(x))^2 = J+\lambda_1+\lambda_2 \quad (i=1,2).$$

Поскольку $g(x)$ — функция по x линейная, а x нормально распределено отсюда следует, что при $m \to \infty$ минимаксная вероятность ошибочной классификации

(4) $$\alpha \to \Phi(J/2\sqrt{J+\lambda_1+\lambda_2}),$$

где $\Phi(t) = (2\pi)^{-1/2}\int_{-\infty}^{t} \exp\{-u^2/2\}\,du$.

В 1970 году Деев [2] исследовал случай дискриминации между двумя невырожденными p-мерными нормальными законами с общей, но неизвестной ковариационной матрицей и нашел, что

(5) $$\alpha \to \Phi\left(-\frac{J(1-\lambda_1\lambda_2/(\lambda_1+\lambda_2))^{1/2}}{2\sqrt{J+\lambda_1+\lambda_2}}\right).$$

Мешалкин и Сердобольский [8] рассмотрели общий негауссовский случай, когда переменные и параметры в каждой из классифицируемых совокупностей можно разложить на независимые блоки ограниченного размера. При наложении ограничения на скорость сближения соответствующих параметров в совокупностях и дополнительных требований типа регулярности им удалось получить для α формулу типа (4), только роль J играло асимптотическое расстояние Кульбака между совокупностями, а $\lambda_i = l/n_i$, где l — число оцениваемых по выборочным данным различающих параметров. При этом допускалось существование порядка l неизвестных, но общих для обеих совокупностей параметров. Заметим, что результат Деева отсюда не следует, поскольку у Деева $p(p+1)/2$ совпадающих параметров.

В применении к нормальным распределениям предположение блочности представляется слишком сильным ограничением, накладываемым на структуру вектора. Заруцкий изучил древовидные зависимости типа Коу [12], в которых предполагается, что можно так переименовать переменные, что в новых обозначениях плотность $f(x) = \Pi_{k=1}^{p} f(x_k \mid x_{i_k})$, где $i_k < k$, и их обобщения. Им было показано [3], что в нормальном случае при минимальных ограничениях на ковариационную матрицу неизвестная структура зависимостей восстанавливается с вероятностью, стремящейся к единице, и что, если в случае, изучавшимся Деевым, структура зависимостей древообразна, то при использовании специального правила построения дискриминантной функции α удовлетворяет формуле (4), а не (5). Древообразные зависимости уже нашли применение в кардиологических исследованиях [11].

Указанные теоретические результаты могут служить своего рода обоснованием для использования двойной нормальной бумаги для представления результатов прогноза. См., например, рис. 1 в [10].

Формулы (4), (5) широко используются при отборе признаков, информативных для разделения совокупностей. Правда, как еще в 1965 г. указывалось Эстесом [13], полученная таким путем оценка для вероятности ошибочной классификации является заниженной, но теоретического обьяснения этому факту дано не было.

По-видимому, первая математическая постановка, четко выявляющая механизм этого занижения, дана в [9]. Вопросы, связанные с оптимальным взвешиванием и отбором факторов в условиях работы [8], изучает в настоящее время Сердобольский.

2. λ-моменты. Хорошо известно, что многие процедуры многомерного статистического анализа быстро теряют свои оптимальные свойства, когда нарушаются предположения нормальности [17]. В этом и следующем параграфах описываются модификации стандартных методов, резко повышающие их устойчивость по отношению к отклонениям от базовых предположений. В теоретическом плане рост устойчивости достигается за счет переопределения исходной статистической модели, такой ее репараметризации, при которой новые параметры имеют хорошие оценки в широком классе шумов, а в идеальном случае совпадают с традиционными. Этим наш подход отличается от работ по робастности, в которых решается задача оптимальной, в некотором смысле, оценки параметров идеальной схемы на фоне шума [14].

Пусть $x \in R^p$, $|u|$ — абсолютное значение u, A — любое выпуклое множество в R^p и $\varrho(F, G) = \sup\limits_A \left| \int\limits_A (dF(x) - dG(x)) \right|$.

Определение 1. *Пусть $w(x)$ — весовая функция, тогда $d = d_w(F) = \int xw(x)dF(x)/e$ и $\Lambda = \Lambda_w(F) = \int (x-d)(x-d)^T w(x)dF(x)/e$, где $e = e_w(F) = \int w(x)dF(x)$, будем называть w-взвешенным средним и w-взвешенной ковариационной матрицей.*

Определение 2. *Если два распределения F и G имеют совпадающие w-взвешенные средние и w-взвешенные ковариационные матрицы, то F и G w-подобны.*

Концепция w-подобия дает возможность связать произвольное распределение F с w-подобным ему нормальным законом N и использовать первые и вторые моменты N при описании F. Однако при этом остается одна трудность — неоднозначность выбора $w(x)$.

Определение 3. *Пусть $\psi = \psi(x, a, \Sigma)$ — плотность нормального закона N с вектором средних a и ковариационной матрицей Σ. Будем называть N (λ, C)-связанным (или короче, λ-связанным) с F, если N ψ^λ-подобен F и $\varrho(F, N) < C$.*

Последнее условие введено для того, чтобы гарантировать при малых C единственность λ-связанного с F нормального закона, так как в общем случае может быть несколько ψ^λ-подобных F нормальных законов.

Определение 4. *Пусть N — λ-связанный с F нормальный закон. Будем называть среднее и ковариационную матрицу N соответственно λ-средним и λ-ковариационной матрицей F.*

Пусть $\mathfrak{N}(p)$ — множество всех несингулярных p-мерных нормальных распределений и $\mathfrak{M}(p, \varepsilon) = \{F: \inf\limits_{N \in \mathfrak{N}(p)} \varrho(F, N) \leq \varepsilon\}$.

Теорема 1[6]. *Для любого* $\lambda>0$ *существуют такие* $C=C(p,\lambda)>0$ *и* $\varepsilon=\varepsilon(p,\lambda,C)>0$, *что для любого* $F\in\mathfrak{M}(p,\varepsilon)$

а) *существует одно и только одно* (λ, C)*-связанное с* F *нормальное распределение;*

б) λ*-среднее* $(a(F))$ *и* λ*-ковариационная матрица* $(\Sigma(F))$ — *непрерывные функции* F (*в смысле* ϱ*-расстояния*);

в) *если* $\eta=A\xi+b$, *где* A — *любая несингулярная квадратная матрица и распределение* $\xi\ F\in\mathfrak{M}(p,\varepsilon)$, *то распределение* $\eta\ G\in\mathfrak{M}(p,\varepsilon)$ *и* λ*-среднее и* λ — *ковариационные матрицы* F *и* G, *связаны соотношениями* $a(G)=Aa(F)+b$ *и* $\Sigma(G)=A\Sigma(F)A^T$.

Для нахождения λ-моментов может быть предложен следующий итеративный процесс:

1) выбрать начальное приближение. Пусть это будет $a=a_0$ и $\Sigma=\Sigma_0$;

2) положить $w=\psi^\lambda(x,a,\Sigma)$ и найти $d=d_w(F)$ и $\Lambda=\Lambda_w(F)$;

3) произвести коррекцию на взвешивание, для этого сначала найти Σ_n из уравнения $\Sigma_n^{-1}=\Lambda^{-1}-\lambda\Sigma^{-1}$, а затем a_n из уравнения $\Sigma_n^{-1}a_n=\Lambda^{-1}d-\lambda\Sigma^{-1}a$;

4) проверить насколько близки (a,Σ) и (a_n,Σ_n). Если различие существенно, положить $a=a_n$, $\Sigma=\Sigma_n$ и повторить вычисления, начиная со второго шага.

3. γ-регрессия. Простейшая модель в R: $x=g(t)+\xi_t$, $E\xi_t=0$. Классические предложения:

а) $g(t)=f(t,\theta)$, где f — известная функция, а θ — вектор неизвестных параметров;

б) ξ_t имеет нормальное распределение;

в) распределение ξ_t не зависит от t.

Рассмотрим сначала случай, когда не имеет места б). Для этого введем понятия λ-регрессии [4]. Пусть $\Phi(t,x)$ — двумерная функция распределения t и x, $F(x\mid t)$ — условия ф. р. x при фиксированном t, $H(t)$ — ф. р. t.

Определение 5. *Пусть* $\psi=\psi(x,a(t),\Sigma(t))$ — *плотность* $N(x\ t)$, (λ, C)*-связанного с* $F(x\mid t)$. *Будем называть* $a(t)$ λ*-регрессией* x *на* t *и* $\Sigma(t)$ λ*-дисперсией* x *относительно* $a(t)$.

Определение 6. *Двумерная ф.р.* $\Phi_\lambda(t,x)$ λ*-регрессионно связана с* $\Phi(t,x)$, *если* $d\Phi_\lambda(t,x)=\psi(x,a(t),\Sigma(t))dx\,dH(t)$.

Аналогия с принципом наименьших квадратов:

Минимум $E_\lambda(x-g(t))^2\equiv\int(x-g(t))^2d\Phi_\lambda(t,x)$ достигается при $g(t)=a(t)$.

Итеративная вычислительная процедура WREG (в полиноминальном случае):

1) положить $a=a_0(t)$, $\Sigma=\Sigma_0$;

2) найти θ_k из уравнений $\partial I/\partial\theta_k=0$, где

$$I=\sum_t\sum_x(x-g(t))^2\psi^\lambda(x-a(t),\ 0,\ \Sigma)\ \text{ и }\ g(t)=\sum_0^s\theta_kt^s;$$

3) найти $\Lambda=I(g(t))/e$, где $e=\sum_x\sum_t\psi^\lambda(x-a(t),\ 0,\ \Sigma)$;

4) найти $a_n(t)$ и Σ_n из уравнений $\Sigma_n^{-1}=\Lambda^{-1}-\lambda\Sigma^{-1}$, $a_n(t)=a(t)+\Sigma_n\Lambda^{-1}(g(t)-a(t))$;

5) сравнить (a,Σ) и (a_n,Σ_n). Если различие существенно, положить $a=a_n$, $\Sigma=\Sigma_n$ и перейти к шагу 2).

Доказано, что

1) $\forall F(x, t) \in \mathfrak{M}(\varepsilon)$ существует одно и только одно λ-регрессионно связанное с $\Phi(t, x)$ распределение $\Phi_\lambda(t, x)$, причем λ-регрессия $a(t)$ и λ-дисперсия $\Sigma(t)$ — непрерывные функции относительно $F(x \mid t)$;

2) при замене в процедуре WREG случайных величин их математическими ожиданиями $a(t)$ и Σ — решения процедуры;

3) если $F(x \mid t)$ нормальны, то при замене в WREG сумм математическими ожиданиями и любом выборе начального приближения $a(t)$, Σ будут найдены после первой итерации;

4) $\forall \lambda > 0$, $\exists c(\lambda) > 0$ и $\varepsilon(\lambda, c) > 0$, что для $F(x \mid t) \in \mathfrak{M}(\varepsilon)$ решения WREG— состоятельная оценка уравнения λ-регрессии.

В случае, когда не имеет места а), но справедливы б) и в), часто используют непараметрические локальные оценки регрессии вида $\hat{x}(t) = \Sigma x_i K(t - t_i)/b)b$, где K—парзеновское окно [1].

Нами было предложено и в этом случае использовать локально обычные параболические оценки линии регрессии [5]. Объединение этого подхода с λ-регрессией позволяет построить технику и для случая, когда не имеет места ни а), ни б).

ЛИТЕРАТУРА

1. Н. Н. Апраушева., В. Д. Конаков. Использование непараметрических оценок в регрессионном анализе. *Заводская лаборатория*, 1973, № 5.
2. А. Д. Деев. Представление статистик дискриминантного анализа и асимптотическое разложение при размерности пространства, сравнимой с объемом выборок. *Доклады АН СССР*, **195**, 1970, 759—762.
3. В. И. Зарудкий. О классификации нормальных векторов простой структуры зависимостей в пространстве большой размерности. *Теория вероятностей и ее применения*, **23**, 1978, 473—475.
4. А. И. Курочкина. Оптимальные свойства главных компонент λ-взвешенной ковариационной матрицы. В сб. *Алгоритмическое и программное обеспечение прикладного статистического анализа*. Москва, 1980.
5. Л. Д. Мешалкин. Использование весовой функции при оценке регрессионной зависимости. В сб. *Многомерный статистический анализ в социально-экономических исследованиях*. Москва, 1974, 25—30.
6. Л. Д. Мешалкин. Параметризация многомерных распределений. В сб. *Прикладной многомерный статистический анализ*. Москва, 1978, 11—18.
7. Л. Д. Мешалкин, А. И. Курочкина. Новый подход к параметризации регрессионных зависимостей. *Зап. научн. семинаров ЛОМИ АН СССР*, **87**, 1979, 79—86.
8. Л. Д. Мешалкин, В. И. Сердобольский. Ошибки при классификации многомерных распределений. *Теория вероятностей и ее прим.*, **23**, 1978, 772—781.
9. Л. Д. Мешалкин. Теория статистического исследования хронически протекающих болезней (методы и частные аспекты методологии). Автореферат докторской диссертации, Москва, 1977.
10. В. Г. Попов, В. С. Юрасов, Л. Д. Мешалкин и др. *Кардиология*, 1976, № 4, 14—20.
11. Р. П. Прохоркас, В. Е. Жютнис, С. В. Мисюнене. Применение некоторых классификаторов для прогнозирования отдаленных исходов инфаркта миокарда. *Проблемы ишемической болезни сердца*. Вильнюс, 1976.
12. C. K. Chow, C. N. Lie. Approximating discrete probablity with dependence trees. *IEEE Trans. Information theory IT—14*, 1968, 462-467.
13. S. E. Estes. Measurment selection for discriminants used in pattern classification Ph. D. Dissertation, Stanford.

14. P. J. Huber. Robust Statistics: A review. *Ann. Math. Stat.*, **43**, 1972, 1041-1067.
15. L. D. Meshalkin. Mathematical methods of chronic desease study in 1970. *Adv. App. Probab.*, **3**, 1971, 194-196.
16. L. D. Meshalkin. Some mathematical methods for the study of noncommunical deseases, Sixth Meeting of the International Epidemiological Association, Primosten, Yugoslavia, 1972, 250-257.
17. J. W. Tukey. A survey of sampling from contaminated distributions. *Contribution to Probability and Statistics* (Ed. Olkin et al.). Stanford, 1960, 448-485.

Центральная научно-исследовательская лаборатория
ул. Маршала Тимошенко 21 Москва 121359

Поступила 6. 9. 1975