www.esa-conference.ru

Статистическая групповая классификация в случае многомерного распределения Пуассона

Каменева Светлана Владимировна, старший преподаватель,
Пермский государственный научно-исследовательский университет, г. Пермь

Аннотация. *В данной статье представлены методы групповой и поточечной классификации, позволяющие исследовать вероятностные модели многомерных дискретных распределений. Построены оптимальные решающие правила для многомерного распределения Пуассона. В качестве частных случаев приведены решающие правила поточечной классификации объектов из данных совокупностей. Рассмотрены статистические решающие правила групповой классификации, построенные на основе оценок максимального правдоподобия и несмещенных оценок.*

Ключевые слова: *дискретные распределения, многомерное распределение Пуассона, статистическая групповая классификация, оценки максимального правдоподобия, несмещенные оценки.*

На практике широкое применение имеет распределение Пуассона. Изучению его посвящено большое количество работ [1], [2], [3].

Распределение Пуассона является идеальной моделью для описания процесса радиоактивного распада. Радиоактивное вещество испускает альфа - частицы. Случайная величина - число частиц, достигающих заданной части пространства (счетчика) в течение времени τ - имеет распределение Пуассона [4].

Дискретная, случайная величина X, принимающая целые неотрицательные значения $x = 0,1,2,...$ в соответствии с некоторым значением τ - временем наблюдения потока событий, имеет распределение Пуассона, когда вероятность $P_\lambda(X = x \mid \tau)$ определяется формулой

$$P_\lambda(X = x \mid \tau) = \frac{(\tau\lambda)^x}{x!} e^{-\tau\lambda}, \quad (1)$$

где λ - параметр распределения Пуассона.

На практике обычно рассматривают распределение Пуассона с фиксированным временем τ ($\tau = 1$), которое задается соотношением

$$P_\lambda(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}. \quad (2)$$

Математическое ожидание и дисперсия распределения Пуассона (2) совпадают

$M_\lambda(X) = D_\lambda(X) = \lambda$.

Обобщением на многомерный случай распределения (2) является многомерное распределение Пуассона. Для него вероятность того, что случайный вектор $X = (X^{(1)},..., X^{(k)})'$ примет фиксированное значение $x = (x^{(1)},..., x^{(k)})'$ задается выражением

$$P_\lambda(X^{(1)} = x^{(1)},..., X^{(k)} = x^{(k)}) = e^{-\sum_{i=1}^{k}\lambda^{(i)}} \prod_{i=1}^{k} \frac{(\lambda^{(i)})^{x^{(i)}}}{x^{(i)}!}. \quad (3)$$

Здесь $\lambda = (\lambda^{(1)}, \lambda^{(2)},..., \lambda^{(k)})'$ - параметр многомерного распределения Пуассона; $\lambda^{(i)}$ - действительные положительные числа; $x^{(i)}$ - целые неотрицательные числа, $i = \overline{1,k}$.

Математические ожидания и дисперсии для распределения (3) находятся аналогично одномерному распределению по следующим формулам:

$M_{\lambda^{(i)}}(X) = D_{\lambda^{(i)}}(X) = \lambda^{(i)}$, $i = \overline{1,k}$.

Асимптотическое поведение многомерного распределения Пуассона (3) было изучено в работах [60], [64], [65]. Известно, что при $\lambda^{(i)} \to \infty$, $x^{(i)} = \lambda^{(i)} + O\left((\lambda^{(i)})^{1/2}\right)$, $i = \overline{1,k}$ и любом натуральном ν для вероятностей (3) имеет место следующее разложение

$$P_\lambda\left(X^{(1)} = x^{(1)},..., X^{(k)} = x^{(k)}\right) = \frac{1}{(2\pi)^{k/2}\left(\prod_{i=1}^{k}\lambda^{(i)}\right)^{1/2}} \cdot e^{-1/2\sum_{i=1}^{k} y_i^2} \times$$

$$\times \exp\left\{\sum_{s=1}^{\nu-1}(-1)^{s-1}\sum_{j=0}^{[(s+1)/2]}\binom{s-j+2}{j}\frac{B_j}{(s-j+1)(s-j+2)}\sum_{i=1}^{k}\frac{y_i^{s-2j+2}}{\left(\lambda^{(i)}\right)^{s/2}} + \quad (4)\right.$$

$$\left. + O\left(\max_{i=\overline{1,k}}\left(\lambda^{(i)}\right)^{-\nu/2}\right)\right\},$$

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

где $y_i = \dfrac{x^{(i)} - \lambda^{(i)}}{\sqrt{\lambda^{(i)}}}$, $i = \overline{1,k}$. Главный член в разложении (4) является k - мерным, нормальным распределением с вектором математических ожиданий $\mu = (\mu_1, \mu_2, ..., \mu_k)$ и матрицей ковариации Σ = $\|\sigma_{i,j}\|$, $i, j = 1, k$.

$$\mu_i = \lambda^{(i)},\ \sigma_{i,j} = \begin{cases} \lambda^{(i)}, i = j, \\ 0, i \neq j. \end{cases}$$

В данной статье исследован статистический подход к анализу дискретных данных, имеющих многомерное распределение Пуассона (3) с помощью методов статистической групповой классификации. Задачи групповой и поточечной классификации решались ранее в основном для совокупностей, имеющих непрерывные распределения. Реальные задачи классификации, как правило, сводятся к вероятностной модели, основанной на предположении о принадлежности рассматриваемой совокупности данных к известным и наиболее часто встречающимся параметрическим семействам распределений, к таким как многомерное нормальное распределение, многомерное распределение Стьюдента, распределение Уишарта, экспоненциальное распределение.

В статье разработаны методы групповой и поточечной классификации, позволяющие исследовать вероятностные модели дискретных распределений. Построены оптимальные решающие правила групповой классификации для многомерного распределения Пуассона. В качестве частных случаев получены решающие правила поточечной классификации.

Построим оптимальные решающие правила групповой классификации для многомерного распределения Пуассона. Так как случай $M > 2$ классов сводится к последовательному решению задач классификации двух классов, то далее будут приведены правила групповой классификации полностью описанных двух классов π_1 и π_2, подробно они изложены в работе [5].

При отсутствии какой-либо априорной информации о вероятностях заданных распределений $P(\pi_{00} \mid \pi_s)$ $(s = 1,2)$ оптимальное решающее правило групповой классификации, основанное на отношении функций правдоподобия, будет иметь вид

$$\pi_{00} \subset \pi_1, \text{ если } \frac{L(\pi_{00} \mid \pi_1)}{L(\pi_{00} \mid \pi_2)} \geq V, \quad (5)$$

где $V = \dfrac{q_2 C(1/2)}{q_1 C(2/1)}$ - заданный порог.

Если же априори известно, что условные распределения вероятностей $P(\pi_{00} \mid \lambda_s)$ зависят от некоторого параметра λ_s (s=1,2), то оптимальное решающее правило групповой классификации строится как отношение условных распределений вероятностей достаточной статистики t вида

$$\pi_{00} \subset \pi_1, \text{ если } \frac{P(t \mid \lambda_1)}{P(t \mid \lambda_2)} \geq V. \quad (6)$$

Для многомерного распределения Пуассона (3) условная функция правдоподобия задается соотношением

$$L(\pi_{00} \mid \pi_s) = e^{-n_0 \sum_{i=1}^{k} \lambda_s^{(i)}} \cdot \prod_{i=1}^{k} \frac{\left(\lambda_s^{(i)}\right)^{\sum_{j=1}^{n_0} x_{0j}^{(i)}}}{\prod_{j=1}^{n_0} x_{0j}^{(i)}!}, \quad (7)$$

а условные распределения вероятностей достаточной статистики $U_0 = (U_0^{(1)}, U_0^{(2)}, ..., U_0^{(k)})'$ параметра $\lambda_s = (\lambda_s^{(1)}, \lambda_s^{(2)}, ..., \lambda_s^{(k)})'$, где $U_0^{(i)} = \sum_{j=1}^{n_0} x_{0j}^{(i)}$, $i = \overline{1,k}$ могут быть найдены по формуле

$$P(U_0 \mid \lambda_s) = e^{-n_0 \sum_{i=1}^{k} \lambda_s^{(i)}} \cdot \prod_{i=1}^{k} \frac{\left(\lambda_s^{(i)}\right)^{U_0^{(i)}}}{U_0^{(i)}!}. \quad (8)$$

Пусть объекты в совокупностях π_s (s=1,2) имеют многомерное распределение Пуассона с вероятностями (3). При отнесении выборки π_{00} к одной из двух совокупностей π_1, π_2 оптимальное решающее правило групповой классификации (5), построенное на основе отношения функций правдоподобия, запишется следующим образом

$$\pi_{00} \subset \pi_1, \text{ если } e^{n_0 \sum_{i=1}^{k} \left(\lambda_2^{(i)} - \lambda_1^{(i)}\right)} \cdot \prod_{i=1}^{k} \left(\frac{\lambda_1^{(i)}}{\lambda_2^{(i)}} \right)^{U_0^{(i)}} \geq V \ . (9)$$

Для многомерного распределения Пуассона использование оптимального решающего правила (8), построенного на отношении условных распределений вероятностей достаточных статистик приводит к тому же правилу (9).

Если ранее распределения данных внутри совокупностей π_S ($s = \overline{1, M}$) предполагались известными, то далее будем считать, что распределения данных известны лишь частично, т.е. содержат неизвестные параметры. В этом случае совокупности задаются только конечными подмножествами своих представителей $\pi_{s0} = (x_{s1}, x_{s2}, ..., x_{sN_s})$, называемыми обучающими выборками объемов N_s из совокупностей π_s ($s = \overline{1, M}$).

Пусть объекты в совокупностях π_s распределены в соответствии с вероятностями $P(x \mid \lambda_s)$, которые зависят от неизвестного параметра λ_s, а достаточная статистика $t = U_0$. Как и ранее, для простоты рассматривается случай, когда имеются только две совокупности π_1 и π_2, заданные своими обучающими выборками π_{10} и π_{20}, соответственно, и выборка π_{00}, которую нужно отнести к одной из них.

В классической задаче распознавания образов по одному объекту (поточечной классификации) [5],[6] решающие правила строились только по обучающим выборкам π_{s0} (s=1,2). Это означало, что структурно они включают только отношения независимых величин. Позднее, в работах [7], [8] был введен принципиально новый подход к решению задач статистической групповой классификации. Классифицируемая выборка π_{00} стала использоваться вместе с обучающими выборками π_{s0} для построения самих статистических решающих функций. При таком подходе возникает другая структура решающих правил, заключающаяся в том, что эти правила основаны на отношениях зависимых случайных величин.

Найдем оценки максимального правдоподобия $\breve{L}(\pi_{00} \mid \lambda_s)$, $\breve{P}(t \mid \lambda_s)$ и построим статистические решающие правила на их основе для многомерного распределения Пуассона (3), где неизвестный параметр λ_s, достаточная статистика $t = U_0$.

Известно [9], [10], что оценкой максимального правдоподобия для неизвестного параметра $\lambda^{(i)}$ многомерного распределения Пуассона (3) является: $\lambda^{(i)} = \dfrac{x^{(i)}}{N}$, где N - объем выборки, $i = \overline{1, k}$. Тогда по объединенной выборке $\pi_{s0} \cup \pi_{00}$ эта оценка запишется

$$\breve{\lambda}_s^{(i)} = \frac{\sum_{j=1}^{N_s} x_{sj}{}^{(i)} + \sum_{j=1}^{n_0} x_{0j}{}^{(i)}}{N_s + n_0} = \frac{U^{(i)}{}_{N_s+n_0}}{N_s + n_0}, \ (i = \overline{1, k}).$$

Подставив $\lambda_s{}^{(i)}$ (s=1,2; $i = \overline{1, k}$) в выражения (7), (8) получим оценки максимального правдоподобия

$$\breve{L}(\pi_{00} \mid \lambda_s) = e^{-\frac{n_0}{N_s+n_0} \sum_{i=1}^{k} U^{(i)}{}_{N_s+n_0}} \prod_{i=1}^{k} \frac{1}{\prod_{j=1}^{n_0} x_{0j}{}^{(i)}!} \cdot \left(\frac{U^{(i)}{}_{N_s+n_0}}{N_s + n_0} \right)^{U_0{}^{(i)}} . (10)$$

$$\breve{P}(U_0 \mid \lambda_s) = e^{-\frac{n_0}{N_s+n_0} \sum_{i=1}^{k} U^{(i)}{}_{N_s+n_0}} \prod_{i=1}^{k} \frac{1}{U_0{}^{(i)}!} \cdot \left(\frac{U^{(i)}{}_{N_s+n_0}}{N_s + n_0} \right)^{U_0{}^{(i)}} . (11)$$

Пусть объекты в совокупностях π_s имеют многомерное распределение Пуассона (3) с неизвестным параметром λ_s (s=1,2). При отнесении выборки π_{00} к одной из совокупностей π_1, π_2 статистическое решающее правило групповой классификации, основанное на отношении оценок максимального правдоподобия, построенных по объединенной выборке $\pi_{s0} \cup \pi_{00}$ (s=1,2) будет иметь вид

$\pi_{00} \subset \pi_1$, если

$$\frac{e^{-\frac{n_0}{N_1+n_0} \sum_{i=1}^{k} U^{(i)}{}_{N_1+n_0}}}{e^{-\frac{n_0}{N_2+n_0} \sum_{i=1}^{k} U^{(i)}{}_{N_2+n_0}}} \prod_{i=1}^{k} \left(\frac{(N_2 + n_0) U^{(i)}{}_{N_1+n_0}}{(N_1 + n_0) U^{(i)}{}_{N_2+n_0}} \right)^{U_0{}^{(i)}} \geq V. \ (12)$$

Статистическое решающее правило (12), основанное на отношении оценок максимального правдоподобия, построенных по обучающей выборке π_{s0} (s=1,2) примет вид

$$\pi_{00} \subset \pi_1, \text{ если } \frac{e^{-\frac{n_0}{N_1}\sum_{i=1}^{k} U^{(i)}{}_{N_1}}}{e^{-\frac{n_0}{N_2}\sum_{i=1}^{k} U^{(i)}{}_{N_2}}} \prod_{i=1}^{k} \left(\frac{N_2 U^{(i)}{}_{N_1}}{N_1 U^{(i)}{}_{N_2}} \right)^{U_0{}^{(i)}} \geq V. \quad (13)$$

При $n_0 = 1$ из правил (12), (13) могут быть получены статистические решающие правила поточечной классификации многомерного распределения Пуассона (3).

Рассмотрим способы нахождения несмещенных оценок $P(\pi_{00} \mid \lambda_s)$, $P(U_0 \mid \lambda_s)$ и построения статистических решающих правил на их основе в случае многомерного распределения Пуассона (3) при неизвестном значении параметра λ_s (s=1,2).

Найдем несмещенную оценку условной функции правдоподобия $L(\pi_{00} \mid \lambda_s)$ по объединенной выборке $\pi_{s0} \cup \pi_{00}$ (s=1,2)

$$P(\pi_{00} \mid \lambda_s) = \frac{P(\pi_{00}, \pi_{s0} + \pi_{00} \mid \lambda_s)}{P(\pi_{s0} + \pi_{00} \mid \lambda_s)} = \frac{P(\pi_{00} \mid \lambda_s) \cdot P(\pi_{s0} \mid \lambda_s)}{P(\pi_{s0} + \pi_{00} \mid \lambda_s)} =$$

$$= \frac{N_s^{\sum_{i=1}^{k} U^{(i)}{}_{N_s}}}{(N_s + n_0)^{\sum_{i=1}^{k} U^{(i)}{}_{N_s+n_0}} \prod_{j=1}^{n_0} \prod_{i=1}^{k} x_{0j}{}^{(i)}!} \cdot \prod_{i=1}^{k} \frac{U^{(i)}{}_{N_s+n_0}!}{U^{(i)}{}_{N_s}!}, \quad (14)$$

где $P(\pi_{00} \mid \lambda_s)$, $P(\pi_{s0} \mid \lambda_s)$, $P(\pi_{s0} + \pi_{00} \mid \lambda_s)$ находятся путем вычисления функций правдоподобия по выборкам π_{00}, π_{s0}, $\pi_{s0} \cup \pi_{00}$ соответственно.

Аналогично находится несмещенная оценка для вероятности условного распределения достаточной статистики $P(U_0 \mid \lambda_s)$ на основе объединенной выборки $\pi_{s0} \cup \pi_{00}$ (s=1,2), которая имеет вид

$$P(U_0 \mid \lambda_s) = \frac{P(U_0, U_{N_s+n_0} \mid \lambda_s)}{P(U_{N_s+n_0} \mid \lambda_s)} = \frac{P(U_0 \mid \lambda_s) \cdot P(U_{N_s} \mid \lambda_s)}{P(U_{N_s+n_0} \mid \lambda_s)} =$$

$$= \frac{n_0^{\sum_{i=1}^{k} U_0{}^{(i)}} N_s^{\sum_{i=1}^{k} U^{(i)}{}_{N_s}}}{(N_s + n_0)^{\sum_{i=1}^{k} U^{(i)}{}_{N_s+n_0}} \prod_{i=1}^{k} U_0{}^{(i)}!} \cdot \prod_{i=1}^{k} \frac{U^{(i)}{}_{N_s+n_0}!}{U^{(i)}{}_{N_s}!}. \quad (15)$$

где $P(U_0 \mid \lambda_s)$, $P(U_{N_s} \mid \lambda_s)$, $P(U_{N_s+n_0} \mid \lambda_s)$ находятся по формуле (8).

На основе полученных несмещенных оценок строятся статистические решающие правила групповой и поточечной классификации многомерного распределения Пуассона при неизвестном значении параметра λ_s (s=1,2).

Пусть объекты в совокупностях π_s имеют многомерное распределение Пуассона (3) с неизвестным параметром λ_s (s=1,2). При отнесении выборки π_{00} к какой-либо одной из совокупностей π_1, π_2 статистическое решающее правило групповой классификации, основанное на отношении несмещенных оценок, построенных по объединенной выборке $\pi_{s0} \cup \pi_{00}$ (s=1,2) будет иметь вид

$$\pi_{00} \subset \pi_1, \text{ если } \frac{N_1^{\sum_{i=1}^{k} U^{(i)}{}_{N_1}}}{N_2^{\sum_{i=1}^{k} U^{(i)}{}_{N_2}}} \cdot \frac{(N_2 + n_0)^{\sum_{i=1}^{k} U^{(i)}{}_{N_2+n_0}}}{(N_1 + n_0)^{\sum_{i=1}^{k} U^{(i)}{}_{N_2+n_0}}} \cdot \prod_{i=1}^{k} \frac{U^{(i)}{}_{N_2}! U^{(i)}{}_{N_1+n_0}!}{U^{(i)}{}_{N_1}! U^{(i)}{}_{N_2+n_0}!} \geq V. \quad (16)$$

При $n_0 = 1$ из правила (16) может быть получено статистическое решающее правило поточечной классификации многомерного распределения Пуассона (3).

Литература:

1. Беляев Ю.К., Лумельский Я.П. Многомерные пуассоновские блуждания// Статистические методы: Межвуз. сб. научн. тр./ Перм. ун-т. Пермь. 1978. С.12 - 17.
2. Ившин В.В., Лумельский Я.П. Статистические задачи оценивания в модели "нагрузка-прочность"// Изд-во Перм. ун-т. Пермь. 1995. 128с.
3. Лумельский Я.П. Статистические оценки результатов контроля качества. М.: Изд-во стандартов. 1979. 200с.

4. Лумельский Я.П., Шеховцова М.Г. Пуассоновские блуждания и оценка надежности// Статист. методы оценивания и проверки гипотез: Межвуз. сб. науч. тр./ Перм. ун-т. Пермь. 1978. С.88 - 100.

5. Каменева С.В. О некоторых задачах проверки гипотез и групповой классификации. Случай дискретных распределений. LAP Lambert. 2016.104 с.

6. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. М.: Статистика. 1974. 240 с.

7. Абусев Р.А. Групповая классификация. Решающие правила и их характеристики. Изд-во Пермского университета. Пермь. 1992. 220 с.

8. Абусев Р.А., Лумельский Я.П. Статистическая групповая классификация// Изд-во Перм. ун-та. Пермь. 1987. 92с.

9. Абусев Р.А. Многомерные пуассоновские блуждания и оценивание распределений достаточных статистик// Тезисы докл. на VII Вильнюсской конф. по теор. вероятн. и матем. статистике. Вильнюс. 1998. С.17.

10. Бабушкина Е.В., Каменева С.В., Абусев Р.А. Оценивание линейных стохастических неравенств в задачах надежности// Материалы межвузовского научно-технического семинара “Проблемы и перспективы рационального управления эксплуатацией вооружения”// Изд-во ПВКИУ. Пермь. 2002. Вып.9. С.22-29.