Апраушева Н. Н., Сорокин С. В.

ОБ ОШИБКЕ КЛАССИФИКАЦИИ ЭЛЛИПСОИДАЛЬНЫХ КЛАССОВ

Учреждение Российской академии наук Вычислительный центр
им. А. А. Дородницына РАН, Москва, Россия

Смеси нормальных распределений широко используются в распознавании образов (классификации) как в режиме с обучением, так и в режиме без обучения [1, 2]. При классификации многомерных выборок из смеси нормальных распределений по алгоритму Дэя-Шлезингера [3-5] было обнаружено, что ошибка классификации $P_{ош}$ зависит не только от объёма выборок n, но и от расположения классов эллипсоидальной формы. При фиксированных значениях n и расстояния Махаланобиса ρ ошибка $P_{ош}$ минимальна, если вектор информативного признака коллинеарен большим осям эллипсоидов равных вероятностей (рис. 1), и она максимальна, если вектор информативного признака коллинеарен их малым осям (рис. 2).

Исследовались двумерные выборки из смеси двух нормальных распределений с плотностью вероятности

$$f(X) = \frac{|\Lambda|^{-\frac{1}{2}}}{2\pi}\sum_{i=1}^{2}\pi_i exp(2^{-1}(X-\mu_i)\Lambda^{-1}(X-\mu_i)'),$$

μ_i — вектор математического ожидания i–й компоненты, π_i — её априорная вероятность, Λ — диагональная ковариационная матрица каждой компоненты, $\pi_i \in (0,1)$, $\pi_1 + \pi_2 = 1$.

Уравнение границы классов имеет вид [6]

$$X\Lambda^{-1}(\mu_1-\mu_2)' - \frac{1}{2}(\mu_1+\mu_2)\Lambda^{-1}(\mu_1-\mu_2)' + ln(\pi_1\pi_2^{-1}) = 0. \qquad (1)$$

Границы классов — это прямая $a\ b$ (рис. 1) и прямая $c\ d$ (рис. 2).

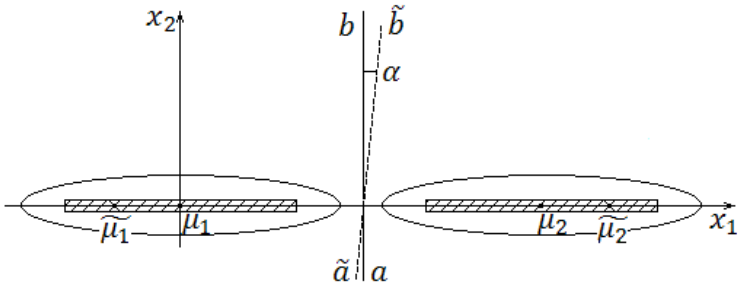

Рис. 1.

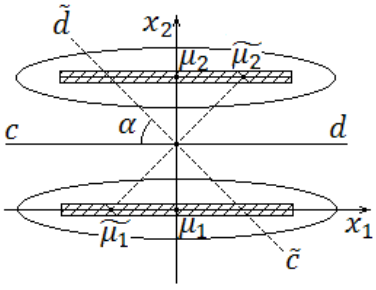

Рис. 2.

Для координат выборочных векторов $\tilde{\mu}_1$, $\tilde{\mu}_2$, вычисляемых по алгоритму Дэя-Шлезингера, в силу неравенства Бьенэме-Чебышева [7] имеем

$$P\{|\tilde{\mu}_{i1} - \mu_{i1}| \le \sigma_{11}/\sqrt{n\gamma}\} > 1-\gamma, \qquad (2)$$

$$P\{|\tilde{\mu}_{i2} - \mu_{i2}| \le \sigma_{22}/\sqrt{n\gamma}\} > 1-\gamma, \quad i = 1,2, \quad 0 < \gamma < 0.1. \qquad (3)$$

σ_{11}^2, σ_{22}^2 — дисперсии признаков x_1, x_2 соответственно. Из неравенств (2), (3) следует, что если $\sigma_{11} \gg \sigma_{22}$, то для абсолютных погрешностей координат векторов μ_1, μ_2 при фиксированных значениях n, γ имеем

$$|\widetilde{\mu_{i1}} - \mu_{i1}| \gg |\widetilde{\mu_{i2}} - \mu_{i2}|, \quad i = 1, 2.$$

На рис 1, 2 даны двумерные доверительные интервалы векторов μ_1, μ_2, (штриховкой), отмечены точками концы векторов μ_1, μ_2, и выборочных векторов $\widetilde{\mu_1}$, $\widetilde{\mu_2}$, границы классов — прямые $a\,b$, $c\,d$ и соответствующие им выборочные прямые $\tilde{a}\,\tilde{b}$, $\tilde{c}\,\tilde{d}$.

В ситуации, представленной на рис. 1, выборочный вектор $\widetilde{\mu_2} - \widetilde{\mu_1}$ почти коллинеарен вектору $\mu_2 - \mu_1$ и выборочная граница классов — прямая $\tilde{a}\,\tilde{b}$ слегка отклоняется от прямой $a\,b$, ошибка классификации, определяемая расстоянием Махаланобиса, равна 6%. В ситуации, представленной на рис. 2, выборочный вектор $\widetilde{\mu_2} - \widetilde{\mu_1}$ значительно отклоняется от вектора $\mu_2 - \mu_1$ (на угол α) и выборочная граница классов — прямая $\tilde{c}\,\tilde{d}$ существенно отклоняется от прямой $c\,d$, что приводит к большой ошибке классификации ($P_{ош}$ ≈50%). Это явление объясняется тем, что вектор $\Lambda^{-1}(\mu_2 - \mu_1)$ является нормалью к прямой (1), разделяющей классы, а его выборочный вектор $\widetilde{\Lambda^{-1}}(\widetilde{\mu_2} - \widetilde{\mu_1})$ в ситуации рис. 1 имеет малое смещение, а в ситуации рис. 2 имеет большое смещение.

Для уменьшения ошибки классификации во 2-м случае следует увеличить объём выборки n или исключить из описания объектов неинформативный признак x_1, или все его значения разделить на такое число $s > 1$, что его дисперсия стала бы меньше дисперсии информативного признака x_2 ($\sigma_{11}^{'} < \sigma_{22}$) [5].

Список литературы

1. Апраушева Н.Н., Горлач И. А. и др. Об опыте автоматического статистического распознавания облачности // Ж. вычисл. математики и мат. физики, 1998, т. 38, №10, с. 1788-1792.
2. Carreira-Perpiñán M. A. Mode-finding for mixture of Gaussian distributions // IEEE Trans. on Patlern Analys. and Mach. Intell., 2000, v. 22, n. 11, p. 1318-1323.
3. Day N. E., Estimating the Components of a Mixture of Normal Distributions. Biometrika, v. 56, n. 3, 1968.
4. Шлезингер М. М. Взаимосвязь обучения и самообучения в распознавании образов. Кибернетика, № 2, Киев, 1969.
5. Апраушева Н. Н. Преобразование признаков при статистическом решении одной задачи автоматической классификации. Изв. АН СССР, Сер. Техническая кибернетика,№ 2, 1985.
6. Андерсон Т. Введение в многомерный статистический анализ. М.: Наука, 1963.
7. Ермаков С. М. Метод Монте-Карло и смежные вопросы. М.: Наука, 1971.