индекс 3624

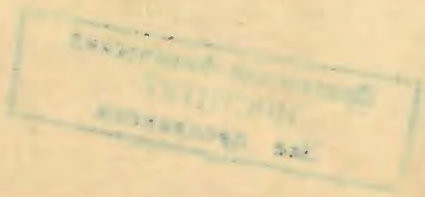
ԵՐԵՎԱՆԻ ՖԻԶԻԿԱՅԻ ԻՆՍՏԻՏՈՒՏ
ЕРЕВАНСКИЙ ФИЗИЧЕСКИЙ ИНСТИТУТ

ЕФИ-683(73)-83

С.Х.АРУТЮНЯН

ВЛИЯНИЕ ИСПОЛЬЗУЕМОЙ ФУНКЦИИ РАССТОЯНИЯ И КОРРЕЛЯЦИИ МЕЖДУ КОМПОНЕНТАМИ МНОГОМЕРНОГО НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ НА ОШИБКУ КЛАССИФИКАЦИИ

ԵՐԵՎԱՆ 1983 ЕРЕВАН



Введение

Наиболее эффективным критерием использования того или иного классификатора является величина вероятности ошибки классификации, основанной на байесовском критерии, так например, байесовская ошибка классификации (БОК). Важными факторами, влияющими на величину этой ошибки, являются величина корреляции между компонентами многомерного вектора наблюдений и вид используемой функции расстояния.

Наличие корреляции создает известные трудности в обработке экспериментальных данных обычными статистическими методами (метод максимального правдоподобия, расчет Монте-Карло). Между тем, наличие корреляции в качестве дополнительной информации улучшает оценки статистической обработки данных методом распознавания образов. Использованию того или иного решающего правила при обработке данных реального эксперимента методами распознавания образов должно предшествовать исследование классификационных возможностей решающего правила в зависимости от условий задачи, т.е. в зависимости от объемов используемых выборок,

расстояний между ними и т.д. Результаты таких исследований показали актуальность задачи выявления механизма влияния корреляции между компонентами вектора наблюдения и используемой функции расстояния на БОК для дальнейших работ по корреляционному анализу многомерных данных методами распознавания образов. Имитационный эксперимент по определению эффективности действия классификатора строится так: генерируются генеральные совокупности n - мерных точек, представляющих альтернативные распределения в признаковом пространстве. Из них случайным образом и в равных объемах компонуются обучающие выборки. Контрольными выборками служат генеральные совокупности. Процедура разпознавания сводится к ранжированию расстояний от контрольной точки до точек матожиданий (квадратичный классификатор) или от контрольной точки до всех точек обучающих выборок (КБС - классификатор) [I] . При ранжировании могут быть использованы различные функции расстояния (ФР). В настоящей работе рассматриваются две функции расстояния [2] :

ФР Евклида $R_E^2 = (\bar{x}-\bar{y})^T(\Sigma')^{-1}(\bar{x}-\bar{y})$, (I)

ФР Махалонобиса $R_M^2 = (\bar{x}-\bar{y})^T\Sigma^{-1}(\bar{x}-\bar{y})$, (2)

где Σ - ковариационная матрица распределения случайных векторов $\bar{x}$, $\bar{y}$; Σ' - диагональная матрица, составленная из диагональных элементов матрицы Σ .

При использовании ФР Махалонобиса обычные евклидовые представления о близости точек неприемлемы. В данной работе делается попытка определения некоторого евклидового аналога ФР Махалонобиса, как геометрической иллюстрации этой функции. Для простоты рассмотрения генерируемые выборки расположены симметрично. Ограничимся двухмерными признаковыми пространствами $\bar{z} = (x, y)$.

Для определения зависимости БОК от величины корреляции $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ рассматриваются два случая:

$$\rho = 0 \quad M_1 = (m_{x1}, m_y); \quad M_2 = (m_{x2}, m_y);$$
$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad (3)$$

и

$$\rho \neq 0 \quad M_1 = (m_{x1}, m_y); \quad M_2 = (m_{x2}, m_y);$$
$$\Sigma_1 = \begin{pmatrix} \sigma^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma^2 \end{pmatrix}; \quad \Sigma_2 = \begin{pmatrix} \sigma^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma^2 \end{pmatrix}. \qquad (4)$$

Легко проверить, что распределения объектов с соотношениями параметров нормальных распределений (3) , (4) симметричны относительно решающей границы, уравнение которой в обоих случаях имеет вид

$$X = (m_{x1} + m_{x2})/2. \qquad (5)$$

2. Функция расстояния Махалонобиса

При обработке многомерных данных методами распознавания образов в силу своей безразмерности широко используется ФР Махалонобиса между точками $\bar{x}$, $\bar{y}$ как некоторая мера их близости, которая представляет собой евклидово расстояние между ними после преобразования декорреляции [2]

$$(\bar{x}-\bar{y})' = \Lambda^{-1/2}\Phi^T(\bar{x}-\bar{y}), \qquad (6)$$

где Λ , Φ ,соответственно, матрицы собственных значений и собственных векторов матрицы Σ : $\Sigma = \Phi\Lambda\Phi^T$.

Тогда $(R_E^2)' = (\bar{x}'-\bar{y}')^T(\bar{x}'-\bar{y}') = (\bar{x}-\bar{y})^T\Phi\Lambda^{-1}\Phi^T(\bar{x}-\bar{y})$ (7)

и $(R_E^2)' = (\bar{x}-\bar{y})^T\Sigma^{-1}(\bar{x}-\bar{y}) = R_M^2$.

Геометрическим местом точек, равноудаленных в смысле Махалонобиса от фиксированной точки, является эллипс, параметры

которого определяются элементами матрицы Σ . В частном случае нормальных распределений эти эллипсы совпадают с ковариационными эллипсами постоянных значений функции плотности распределения вероятности

$$(2\pi)^{-1}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\bar{x}-\bar{M})^{T}\Sigma^{-1}(\bar{x}-\bar{M})\right\}=const, \qquad (8)$$

где $|\Sigma|$ - определитель матрицы Σ .

Выражение (2) для двухмерных пространств имеет вид

$$R_M^2=\frac{(x-m_x)^2}{\sigma^2}+\frac{(y-m_y)^2}{\sigma^2}+\Delta(\bar{z},\bar{M},\bar{\Sigma}), \qquad (9)$$

где $\Delta(\bar{z};\bar{M},\Sigma)$ - корреляционная добавка

$$\Delta(\bar{z},\bar{M},\bar{\Sigma})=\frac{\sigma_{xy}^2}{|\Sigma|}\left\{\frac{(x-m_x)^2}{\sigma^2}+\frac{(y-m_y)^2}{\sigma^2}-\frac{2(x-m_x)(y-m_y)}{\sigma_{xy}}\right\}. \qquad (10)$$

При $\sigma_{xy}=0$ $\Delta(\bar{z},\bar{M},\Sigma)=0$.

Рассмотрим эллипс $R_M^2(\bar{z},\bar{M},\Sigma)=C_{\neq o}$, так что

$$C_{\neq o}=C_o+\frac{\sigma_{xy}^2}{|\Sigma|}\left\{C_o-\frac{2(x-m_x)(y-m_y)}{\sigma_{xy}}\right\}, \qquad (11)$$

где C_o - окружность, вписанная в квадрат с эллипсом $C_{\neq o}$ (рис.1). Сторона квадрата равна $2\sigma\sqrt{C_o}$. При $|\rho|\to 1$ эллипс вытягивается в диагональ квадрата [3] . Окружность C_o пересекает эллипс $C_{\neq o}$ в четырех точках A,B,C,D координаты которых определяются из решения системы уравнений:

$$\begin{cases} 2\dfrac{(x-m_x)(y-m_y)}{\sigma_{xy}}=C_o & (12a)\\[2ex] \dfrac{(x-m_x)^2}{\sigma^2}+\dfrac{(y-m_y)^2}{\sigma^2}+\Delta(\bar{z},\bar{M},\Sigma)=C_{\neq o} & (12б)\end{cases}$$

В точках A,B,C,D определяется евклидовый аналог расстояния Махалонобиса, гипербола (12а) проходит через эти точки и отсекает от эллипса дуги с разными соотношениями между ФР Евклида и Махалонобиса, точки эллипса на дугах A,B,C,D удалены от центра $O=(m_x,m_y)$ больше по ФР Евклида, чем ФР Махалонобиса. И наоборот, точкам эллипса дуг A,B,C,D соответствуют большие значения ФР Махалонобиса, чем Евклида.

3. Зависимость БОК от функции распределения и величины коэффициента корреляции (ρ)

Исследуем особенности ФР Махалонобиса в применении к задаче распознавания. Определим БОК как оценку качества используемого критерия отбора, т.е. проведем процедуру распознавания по классифицированным объектам. Обратимся к рис.2. Здесь точки $O_1=(m_{x1},m_y)$, $O_2=(m_{x2},m_y)$ - центры ковариационных эллипсов $C_{\neq o1}$, $C_{\neq o2}$ первого и второго классов, соответственно (случай $\rho\neq 0$, (4)) с равными значениями C_o . Решающая граница (5) отсекает от эллипсов дуги $A_o B_o$ и дуги AB, на которых процедура распознавания дает ошибочный результат. Рассмотрим два классификатора - параметрический квадратичный классификатор и непараметрический КБС классификатор [2] : В случае нормальных распределений байесовский критерий параметрической классификации сводится к определению расстояния от контрольной точки $\bar{x}$ до точек матожиданий O_1 , O_2 . Этот критерий по ФР Евклида ошибочно отнесет точки дуги $c\bar{x}D$ ко второму классу, а по ФР Махалонобиса - к первому, т.е. классификация будет правильной.

Здесь $\begin{cases} C_{o1}\geqslant C_{o2} \\ C_{\neq o1}\leqslant C_{\neq o2}\end{cases}$

Точки дуг AC обеими ФР классифицируются неправильно

$$\begin{cases} C_{01} \geq C_{02} \\ C_{\neq 01} \geq C_{\neq 02} \end{cases}$$

Таким образом длина дуги $c\bar{x}D$ может служить оценкой эффективности использования ФР Махалонобиса в сравнении с ФР Евклида. В предельном случае $|\rho| \to 1$ дуга $c\bar{x}D$ представляет уже всю дугу ошибочных решений, и эффективность использования ФР Махалонобиса максимальна. Данному анализу влияния ФР на БОК соответствует результат имитационного эксперимента.

В нижеследующей таблице сведены результаты классификации контрольных совокупностей по 500 элементов в каждой. M - число элементов в обучающих выборках.

Таблица

M = 10	M = 20	M = 30	M = 40	M = 50	M = 100	M =500
0,33	0,33	0,31	0,35	0,33	0,35	0,31
					(ФР Евклида)	
0,2	0,19	0,19	0,18	0,18	0,18	0,17
					(ФР Махалонобиса)	

Коэффициент корреляции $\rho_{1,2} = \pm 0.8$; расстояние Бхатачария [2] между классами $R_B^2 = 1,62$. Данному значению расстояния Бхатачария соответствуют следующие оценки верхней и нижней границ Чернова для ошибки классификации, вычисленных по формулам [1]

$\varepsilon_B = \frac{1}{2} \times \exp(-R_E^2) = 0,27$; $\varepsilon_H = \frac{1}{2} - \frac{1}{2} \times (1 - 4 \times \varepsilon_B^2)^{1/2} = 0,08$. При использовании ФР Евклида возможен точный расчет БОК по формуле

БОК $= \frac{1}{2} - \frac{1}{2} \times erf\left(\left(\frac{1}{2}R_E^2 + T\right)/\sqrt{2}R_E\right)$, где T - логарифм отношения априорных вероятностей [4], БОК = 0,31.

При аналогичном анализе непараметрических классификаторов, определяющим становится особенность непараметрической оценки функции плотности - локальное оценивание, сводящееся к определению преобладания числа представителей конкурирующих классов, т.е. решающим фактором является плотность распределения точек (число объектов в единичном объеме) классов в окрестности контрольной точки. Областям постоянных значений функции плотности, например, ковариационным n - мерным эллипсоидам, соответствует равномерное распределение объектов. С её уменьшением плотность распределения точек уменьшается и наоборот. На рис.2 контрольная точка $\bar{x}$ занимает промежуточное положение на дуге между точками A, D. С приближением к точке B, лежащей на большой полуоси эллипса $C_{\neq 02}$, плотность распределения точек второго класса максимальна, в точке C плотности выравниваются (здесь $C_{\neq 01} = C_{\neq 02}$). На интервале CD плотность распределения точек первого класса (которая постоянна) всюду больше плотности распределения точек второго класса. С увеличением коэффициента корреляции ($\rho \to 1$) дуга CB охватывает все большую часть дуги ACD ошибочных решений. Поэтому увеличение корреляции приводит к уменьшению БОК. На рис.3 сравниваются результаты оценки БОК КБС классификатором для двух случаев распределений ($\rho = 0$; $\rho_{1,2} = \pm 0,8$). По оси абсцисс отложены значения K - числа ближайших соседей.

Однако необходимо уточнить разницу оценок плотности распределения точек в признаковом пространстве в связи с использованием разных ФР. Пусть точки $\bar{z}$ и $\bar{y}$ представляют II и I - классы в K = 2 окружении контрольной точки $\bar{x}$ (рис.2). Оценим разность

$$C_{\neq o2}(\bar{x},\bar{z}) - C_{\neq o1}(\bar{x},\bar{y}) = C_{o2}(\bar{x},\bar{z}) - C_{o1}(\bar{x},\bar{y}) + \Delta_2(\bar{x},\bar{z},\bar{\Sigma}_2) - \Delta_1(\bar{x},\bar{y},\Sigma_1) \quad (13)$$

Равным плотностям распределения точек (в понимании ФР Евклида) соответствует

$$C_{o2}(\bar{x},\bar{z}) = C_{o1}(\bar{x},\bar{y}). \quad (14)$$

Знак разности в левой части равенства (13) определяется знаком разности корреляционных добавок, которая при условии (14) порядка нуля.

С увеличением положительной разности плотностей распределения точек первого и второго классов

$$C_{o2}(\bar{x},\bar{z}) > C_{o1}(\bar{x},\bar{y})$$

знак разности $C_{\neq o2}(\bar{x},\bar{z}) - C_{\neq o1}(\bar{x},\bar{y})$ определяется из выражения

$$\left\{C_{o2}(\bar{x},\bar{z}) - C_{o1}(\bar{x},\bar{y})\right\} + \frac{\sigma^2_{xy}}{|\Sigma|}\left\{\left[C_{o2}(\bar{x},\bar{z}) - C_{o1}(\bar{x},\bar{y})\right] - 2\left[\frac{(x_1-z_1)(x_2-z_2)}{\sigma_{12(2)}} - \frac{(x_1-y_1)(x_2-y_2)}{\sigma_{12(1)}}\right]\right\} \gtrless 0. \quad (15)$$

Определение знака этой суммы сводится к сравнению величин, стоящих в квадратных скобках. С уменьшением плотности распределения точек второго класса $C_{o1}(\bar{x},\bar{y}) \simeq const$ увеличивается значение $C_{o2}(\bar{x},\bar{z})$. Однако примерно во столько же раз растет значение $2(x_1-z_1)(x_2-z_2)/\sigma_{12(2)}$. Поскольку $|\sigma_{12}| \leq \sigma^2$, разность

$$C_{o2}(\bar{x},\bar{z}) - 2(x_1-z_1)(x_2-z_2)/\sigma_{12(2)} \quad (16)$$

может принимать отрицательные значения и, таким образом, обусловить некоторую разницу в оценках ошибки классификации в пользу использования ФР Махалонобиса. С увеличением числа K этот эффект проявляется слабо даже для значительных величин ρ (рис.4).

Вместе с тем, из рассмотрения (16) следует, что при $|\rho| \to 1 (|\sigma_{12}| \to \sigma^2)$ эта разность порядка нуля для любых расположений точек $\bar{z}$, $\bar{y}$, $\bar{x}$ и оценка БОК КБС классификатором не чувствительна к виду используемой ФР.

4. Заключение

Использование ФР Махалонобиса в задачах определения ошибок классификации имеет существенный недостаток – обработка данных на ЭВМ требует больших затрат машинного времени. В ряде случаев эти затраты оправданы меньшими ошибками классификации (например, для параметрических квадратичных классификаторов). При использовании непараметрического КБС классификатора использование ФР Евклида приведет к примерно вдвое меньшим затратам машинного времени, чем использование ФР Махалонобиса; значение БОК при этом не изменится. Анализ БОК в зависимости от величины корреляции (рис.3) подтвердил дополнительную разделительную ценность корреляции между компонентами многомерного нормального распределения.

В заключение автор выражает благодарность Чилингаряну А.А. за предоставление некоторых подпрограмм.

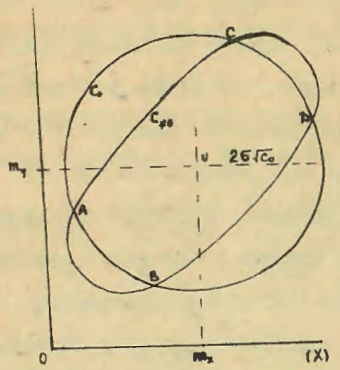


Рис.I

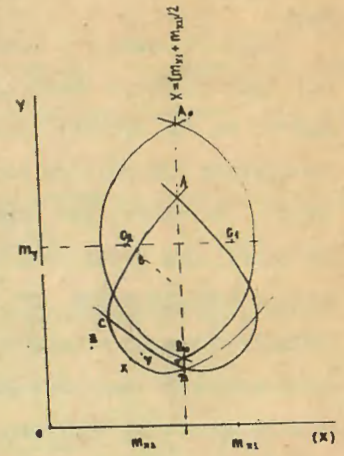


Рис.2

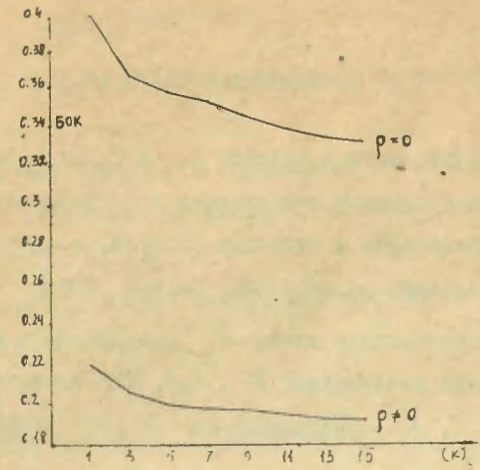


Рис.3

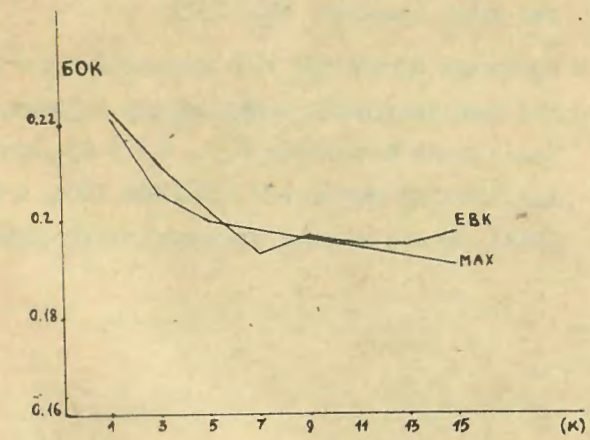


Рис.4

ПОДПИСИ К РИСУНКАМ

Рис.I Ковариационные эллипсы C_0 и $C_{\neq 0}$ нормальных распределений с равными дисперсиями и с коэффициентом корреляции, равным нулю и отличным от нуля, соответственно, вписаны в квадрат со стороной, равной $2\delta\sqrt{c_0}$

Рис.2 Распознавание точки $\bar{x}$ квадратичным классификатором (точки матожиданий 0_I , 0_2), КБС классификатором (точки $\bar{y}$, $\bar{z}$, принадлежащие I и 2 классам соответственно)

Рис.3 Сравнение значений БОК КБС классификатором для нормальных распределений с коэффициентом корреляции, равным нулю ($\rho=0$) и отличным от нуля ($\rho_{1,2}=\pm 0{,}8$). По оси абсцисс отложены "число ближайших соседей". Расстояние между классами $R_Б^2 = 0{,}62$

Рис.4 Сравнение оценок БОК КБС классификатором двух нормальных распределений с коэффициентом корреляции $\rho_{1,2}=\pm$ 0.8; (расстояние Бхатачария $R_Б^2$ = 0,62) при использовании двух функций расстояния - Евклида (ЕВК) и Махалонобиса (МАХ). По оси абсцисс отложены" число ближайших соседей"

СПИСОК ЛИТЕРАТУРЫ

I. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука, I979.

2. Рао С.Р. Кластер-анализ в применении к изучению перемешивания рас в популяциях людей. В сб. Классификация и кластер. М.: Мир, I980, с.I55.

3. Брант З. Статистические методы анализа наблюдений. М.:Мир, I975, с.67.

Рукопись поступила I9 июля I983 г.

Редактор Л.П.Мукаян

Тех.редактор А.С.Абрамян

Заказ 360 ВФ-04572 Тираж 299

Препринт ЕФИ Формат издания 60x84/16

Подписано к печати 21/XI-83 1,0 уч.-изд.л.Ц. 15 к.

Издано Отделом научно-технической информации

Ереванского физического института. Ереван 36, Маркаряна 2