Субективно оценяване на корелационни коефициенти

Наталия Николова, Даниела Тонева, Бисер Стоянов, Кирил Тенекеджиев

Subjective Assessment of Correlation Coefficients: *The paper discusses pattern recognition based on continuous multi-normal features and proposes subjective estimation of the necessary parameters. The main task is to construct the conditional likelihoods, which in turn means to find the covariance matrix. Subjectively elicited quantiles are used to find conditional quantiles, which are employed in the assessment of the covariance matrix elements. Because of the subjective estimates used, the covariance matrix often turns to be fictitious (i.e. with negative eigen values), which is unacceptable in a pattern recognition task. Therefore earlier works proposed ways to transform it into a classical one.*

Keywords: *Statistical pattern recognition, continuous multi-normal features, correlation coefficients, subjective estimates*

ВЪВЕДЕНИЕ

В процеса на диагностика много често се налага решения да се вземат на основата на непълна и понякога изкривена информация за действителното състояние на изследвания обект. Затова се налага да се прибегне до използването на субективна (експертна) информация за причисляването на обекта към предварително определени класове на възможните му състояния. Целта е да се построи класификатор на състоянието S чрез формално описание на това състояние. Това описание приема формата на многомерен вектор на измерването $\vec{x}$, чиито координати при случая на експертна оценка са оценими променливи величини. Дефинират се c на брой класове взаимноизключващи се състояния w_k. Произволен обект S, описан чрез $\vec{x}$ принадлежи само на един от класовете на състоянието с определена вероятност. Разпознаването може да се извърши по четири вида диагностични признаци.

Целта на разпознаването на образи е да се определи вероятността обекта S да принадлежи на всеки от класовете на състояние, чрез изчисляване на апостериорните вероятности по формулата на Бейс [Clemen, 1996]. Това се свежда до задача за намиране на условните правдоподобия да се наблюдава вектор $\vec{x}$ ако обекта S принадлежи на всеки от класовете на състояние. Доклада разглежда тази задача при непрекъснати многомерно-нормално разпределени признаци. Намирането на необходимите параметри използва субективни оценки на квантили на разпределенията на признаците, които се намират по известни техники [Николова, Стоянов, Тенекеджиев, 2010]. На тази база се намират елементите на ковариационната матрица, нужна за построяване на условната плътност на непрекъснатите многомерно-нормални признаци.

ФОРМАЛИЗАЦИЯ НА ЗАДАЧАТА ЗА РАЗПОЗНАВАНЕ НА ОБРАЗИ

Целта на разпознаването на образи е да се определи вероятността обекта S да принадлежи на всеки от класовете w_1, w_2, …, w_c: $P(S \in w_k \mid \vec{x})$ [Fukunaga, 1990]. Тези последни вероятности се наричат апостериорни и включват:

а) базови равнища $P(S \in w_k)$, за k=1,2,…, c, които изразяват вероятността случайният обект S да принадлежи на w_k без да е налично знание за $\vec{x}$;

b) плътностите на условните правдоподобия $f_k(\vec{x} / S \in w_k)$, за k=1,2,…, c, които са пропорционални на вероятността да се изрази обекта S чрез $\vec{x}$ ако принадлежи на w_k. Връзката между априорни вероятности, апостериорни вероятности и условни правдоподобия се дава чрез формулата на Бейс:

(1) $$P(S \in w_k \mid \vec{x}) = \frac{P(S \in w_k) f_k(\vec{x} \mid S \in w_k)}{f(\vec{x})}$$, за k=1,2,…, c,

където $f(\vec{x})$ е безусловното правдоподобие, пропорционално на вероятността да се изрази обекта S чрез $\vec{x}$ без знание за принадлежността на S. Безусловното правдоподобие може да се изчисли по формулата за пълната вероятност:

(2) $$f(\vec{x}) = \sum_{k=1}^{c} P(S \in w_k) f_k(\vec{x} \mid S \in w_k).$$

Известни са четири вида признаци за разпознаване – дискретни, псевдо-дискретни, независими непрекъснати и непрекъснати многомерно-нормално разпределени признаци. Тогава векторът $\vec{x}$ може да се представи като

(3) $$\vec{x} = \left(\vec{x}^d, \vec{x}^p, \vec{x}^c, \vec{x}^i\right)^T.$$

Дискретните, псевдо-дискретните, непрекъснатите многомерно-нормално разпределени и независимите признаци съответно са обединени в следните вектори:

(4) $$\vec{x}^d = \left(x_1^d, x_2^d, ..., x_a^d\right),$$

(5) $$\vec{x}^p = \left(x_1^p, x_2^p, ..., x_t^p\right),$$

(6) $$\vec{x}^c = \left(x_1^c, x_2^c, ..., x_e^c\right),$$

(7) $$\vec{x}^i = \left(x_1^i, x_2^i, ..., x_g^i\right).$$

Видно е, че $\vec{x} = \left(\vec{x}^d, \vec{x}^p, \vec{x}^c, \vec{x}^i\right)^T$ е "*a*+*t*+*e*+*g*"-мерен вектор със смесени признаци. Прието е като стандарт, че дискретните, псевдо-дискретните и независимите непрекъснати признаци са независими помежду си и от непрекъснатите многомерно-нормално разпределени признаци в произволен клас на състояние. Тогава:

(8) $$f_k(\vec{x} \mid w_k) = f_k^d\left(\vec{x}^d \mid w_k\right) f_k^p\left(\vec{x}^p \mid w_k\right) f_k^c\left(\vec{x}^c \mid w_k\right) f_k^i\left(\vec{x}^i \mid w_k\right)$$, k=1,2,…,c.

В (8) $f_k^d\left(\vec{x}^d \mid w_k\right)$ е условната плътност на вероятността дискретните признаци на S да приемат стойности $\vec{x}^d$, ако S принадлежи на клас w_k. Аналогични са значенията и на останалите плътности в (8). Вижда се, че проблемът за откриване на апостериорните вероятности се свежда до оценка на условните правдоподобия.

ИНТЕЛИГЕНТНА ОЦЕНКА ПРИ НЕПРЕКЪСНАТИ МНОГОМЕРНО-НОРМАЛНО РАЗПРЕДЕЛЕНИ ПРИЗНАЦИ

Тъй като $\vec{x}^c$ е многомерно-нормално разпределен, условната плътност може да се изчисли като:

(9) $$f_k^c(\vec{x}^c/w_k) = \frac{1}{(2\pi)^{\frac{e}{2}} |K_k|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(\vec{x}^c - \vec{\mu}_k\right)^T K_k^{-1}\left(\vec{x}^c - \vec{\mu}_k\right)\right]$$, k=1,2,…, c.

Тук, $\vec{\mu}_k$ е вектор на средните стойности и е e-мерен вектор-стълб с i-та координата $\vec{\mu}_k$, която е условната средна стойност на непрекъснатия многомерно-нормално разпределен признак $ncfeature_{i,}$ ако S принадлежи на w_k:

(10) $$\vec{\mu}_k = (\mu_1^{(k)}, \mu_2^{(k)}, ..., \mu_e^{(k)})^T$$, за k=1,2,…, c.

В (9), K_k е ковариационна матрица и е симетрична квадратна матрица с e реда и e колони и с i,j-ти елемент $r_{i,j}^{(k)} \sigma_i^{(k)}$, като $\sigma_i^{(k)}$ е стандартното отклонение на непрекъснатия многомерно-нормално разпределен признак $ncfeature_i$, ако S

принадлежи на w_k, докато $r_{i,j}^{(k)}$ е коефициент на корелация между непрекъснатия многомерно-нормално разпределен признак i и j ако S принадлежи на w_k. В сила са следните условия:

(11) $\sigma_i^{(k)} \geq 0,\ r_{i,j}^{(k)} \geq -1,\ r_{i,j}^{(k)} \leq 1,\ r_{i,j}^{(k)} = r_{j,i}^{(k)},\ r_{i,i}^{(k)} = 1$, за k=1,2,..., c.

Тогава

(12) $$K_k = \begin{pmatrix} \left[\sigma_1^{(k)}\right]^2 & r_{1,2}^{(k)}\sigma_1^{(k)}\sigma_2^{(k)} & \dots & r_{1,e}^{(k)}\sigma_1^{(k)}\sigma_e^{(k)} \\ r_{2,1}^{(k)}\sigma_2^{(k)}\sigma_1^{(k)} & [\sigma_2^{(k)}]^2 & \dots & r_{2,e}^{(k)}\sigma_2^{(k)}\sigma_e^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ r_{e,1}^{(k)}\sigma_e^{(k)}\sigma_1^{(k)} & r_{e,2}^{(k)}\sigma_e^{(k)}\sigma_2^{(k)} & \dots & \left[\sigma_e^{(k)}\right]^2 \end{pmatrix}$$, за k=1,2,..., c.

Така, за достигане до необходимата субективна информация при непрекъснато многомерно нормално разпределение е необходимо да се изчислят параметрите:

(13) $\mu_i^{(k)}$, за i=1,2,..., e; k=1,2,.., c;

(14) $\sigma_i^{(k)}$, за i=1,2,..., e; k=1,2,.., c;

(15) $r_{i,j}^{(k)}$, за i=1,2,..., e-1; j=i+1, i+2, ..., e; k=1,2,.., c;

Векторът на средните стойности съвпада с безусловния квантил $x_{0.5}$ (медианата на разпределението). Квантил x_α се дефинира неявно:

(16) $$\int_{-\infty}^{x_\alpha} f(x,\mu,\sigma)dx = \alpha .$$

Чрез екстрахиране и след осредняване, за $x_{0.5}$ се получава:

(17) $x_{0.5}^{ext} \in \left[x_{0.5}^{down}; x_{0.5}^{up}\right]$,

(18) $$\overline{x}_{0.5} = \frac{x_{0.5}^{down} + x_{0.5}^{up}}{2},$$

(19) $\hat{\mu}_x = \overline{x}_{0.5}$.

По аналогичен начин се намира μ за всяка отделна координата. Дефинира се неявна функция:

(20) $NI(\alpha, \mu_x, \sigma) = x_\alpha$.

При α =0.25 и $\mu_x = \hat{\mu}_x$ се търси $x_{0.25}(\sigma)$ като:

(21) $x_{0.25}(\sigma) = NI(0.25, \overline{x}_{0.5}, \sigma)$,

(22) $x_{0.25}^{ext} \in \left[x_{0.25}^{down}; x_{0.25}^{up}\right]$,

(23) $$\overline{x}_{0.25} = \frac{x_{0.25}^{down} + x_{0.25}^{up}}{2}.$$

Аналогично се подхожда при α =0.75:

(24) $x_{0.75}(\sigma) = NI(0.75, \overline{x}_{0.5}, \sigma)$,

(25) $x_{0.75}^{ext} \in \left[x_{0.75}^{down}; x_{0.75}^{up}\right]$,

(26) $$\overline{x}_{0.75} = \frac{x_{0.75}^{down} + x_{0.75}^{up}}{2}.$$

Въвеждат се следните функции:

(27) $$Z_{0.25} = \frac{x_{0.25} - x_{0.5}}{\sigma},$$

(28) $$Z_{0.75} = \frac{x_{0.75} - x_{0.5}}{\sigma}.$$

Сега могат да се извършат оценки за σ. Първа възможна оценка се дава чрез:

(29) $$\sigma_1 = \frac{x_{0.25} - x_{0.5}}{Z_{0.25}}.$$

Чрез използване на $x_{0.75}$ се намира следната оценка за параметъра σ:

(30) $$\sigma_2 = \frac{x_{0.75} - x_{0.5}}{Z_{0.75}}.$$

Трета оценка за този параметър може да се получи след осредняване на оценките от формули (29) и (30)

(31) $$\sigma_3 = \frac{\sigma_1 + \sigma_2}{2}.$$

Друга оценка на параметъра може да се намери като аргумента, при който функцията (32) има минимална стойност:

(32) $$\chi^2(\sigma) = \left(\frac{x_{0.25}(\sigma) - \overline{x}_{0.25}}{x_{0.25}^{up} - x_{0.25}^{down}} \right)^2 + \left(\frac{x_{0.75}(\sigma) - \overline{x}_{0.75}}{x_{0.75}^{up} - x_{0.75}^{down}} \right)^2.$$

Аналогично се извършват оценки на σ по всички останали димензии.

За построяването на ковариационната матрица е необходимо да се намерят оценки на коефициента на корелация *r*, което се извършва с използването на условни квантили (както и някоя от получените оценки за σ). Използват се следните условни квантили:

(33) $$y_{0.5} \mid x_{0.25} = \overline{y}_{0.5} - r \frac{\hat{\sigma}_y}{\hat{\sigma}_x} (\overline{x}_{0.25} - \overline{x}_{0.5}),$$

(34) $$y_{0.5} \mid x_{0.75} = \overline{y}_{0.5} - r \frac{\hat{\sigma}_y}{\hat{\sigma}_x} (\overline{x}_{0.75} - \overline{x}_{0.5}),$$

(35) $$x_{0.5} \mid y_{0.25} = \overline{x}_{0.5} - r \frac{\hat{\sigma}_x}{\hat{\sigma}_y} (\overline{y}_{0.25} - \overline{y}_{0.5}),$$

(36) $$x_{0.5} \mid y_{0.75} = \overline{x}_{0.5} - r \frac{\hat{\sigma}_x}{\hat{\sigma}_y} (\overline{y}_{0.75} - \overline{y}_{0.5}).$$

Оценката на тези условни квантили се получава по следният начин:

(37) $$(y_{0.5} \mid \overline{x}_\alpha)^{ext} = \left[(y_{0.5} \mid \overline{x}_\alpha)^{down}; (y_{0.5} \mid \overline{x}_\alpha)^{up} \right],$$

(38) $$(y_{0.5} \mid \overline{x}_\alpha) = \frac{(y_{0.5} \mid \overline{x}_\alpha)^{down} + (y_{0.5} \mid \overline{x}_\alpha)^{up}}{2}.$$

Аналогични изчисления се извършват и за квантилите за *x* при фиксирана стойност на *y*.

Тогава оценки за корелационния коефициент *r* могат да се получат като:

(39) $$r_1 = \frac{y_{0.5} - (y_{0.5} \mid x_{0.25})}{y_{0.5} - y_{0.25}},$$

(40) $$r_2 = \frac{(y_{0.5} \mid x_{0.75}) - x_{0.5}}{y_{0.75} - y_{0.5}},$$

(41) $$r_3 = \frac{r_1 + r_2}{2}.$$

Четвърта оценка на параметъра е аргумента, при който функцията (42) се минимизира:

(42) $$\chi_4^2(r) = \left[\frac{y_{0.5} \mid x_{0.25} - y_{0.5} \mid \bar{x}_{0.25}}{\left(y_{0.5} \mid \bar{x}_{0.25}\right)^{up} - \left(y_{0.5} \mid \bar{x}_{0.25}\right)^{down}}\right]^2 + \left[\frac{y_{0.5} \mid \bar{x}_{0.75} - y_{0.5} \mid \bar{x}_{0.75}}{\left(y_{0.5} \mid \bar{x}_{0.75}\right)^{up} - \left(y_{0.5} \mid \bar{x}_{0.75}\right)^{down}}\right]^2.$$

Аналогично на формули (39) и (40) се изчисляват оценките на условните квантили при разменени стойности за *x* и *y*:

(43) $$r_5 = \frac{x_{0.5} - \left(x_{0.5} \mid y_{0.25}\right)}{x_{0.5} - x_{0.25}},$$

(44) $$r_6 = \frac{\left(x_{0.5} \mid y_{0.75}\right) - x_{0.5}}{x_{0.75} - x_{0.5}}.$$

Аналогично на формула (42), корелационния коефициент може да се оцени като аргумента, който минимизира (45):

(45) $$\chi_5^2(r) = \left[\frac{x_{0.5} \mid y_{0.25} - x_{0.5} \mid \bar{y}_{0.25}}{\left(x_{0.5} \mid \bar{y}_{0.25}\right)^{up} - \left(x_{0.5} \mid \bar{y}_{0.25}\right)^{down}}\right]^2 + \left[\frac{x_{0.5} \mid \bar{y}_{0.75} - x_{0.5} \mid \bar{x}_{0.75}}{\left(x_{0.5} \mid \bar{y}_{0.75}\right)^{up} - \left(x_{0.5} \mid \bar{y}_{0.75}\right)^{down}}\right]^2.$$

Още една оценка за коефициента на корелация може да се получи при минимизирането на следната функция:

(46) $$\chi_6^2(r) = \chi_4^2 + \chi_5^2.$$

Така вече са пресметнати всички необходими елементи за създаването на ковариационната матрица.

ЗАКЛЮЧЕНИЕ

Докладът разгледа задача за разпознаване на образи при непрекъснати многомерно-нормално разпределени признаци. Тя се свежда до построяване на условната плътност, която от своя страна изисква намирането на ковариационна матрица. За целта се използват (най-често) субективни оценки на квантили, на база на които се намират и условни квантили на разпределенията на признаците. Представените зависимости използват тези квантили при намиране на корелационните коефициенти между многомерно-нормалните признаци.

Ковариационната матрица е квадратна матрица с неотрицателни собствени стойности. В алгоритмите за субективна оценка на нейните елементи, често се получават отрицателни стойности. Такава матрица се означава като фиктивна в [Tenekedjiev, Karakatsanis, Bekiaris, 2000] и не може да се използва в задачите за статистическо разпознаване на образи. Затова в [Tenekedjiev, Karakatsanis, Bekiaris, 2000; Николова, Стоянов, Тенекеджиев, 2010] се представят алгоритми за преобразуване на фиктивни в класически ковариационни матрици.

ЛИТЕРАТУРА

[1]. Николова, Н.Д., Стоянов, Б., Тенекеджиев, К., Експертна оценка на коефициенти на корелация при многомерно нормално разпределение с условни квантили. Автоматика и информатика'2010, 2010, 3-6.10, София, стр.II-403 – II.406.

[2]. Clemen, R., Making Hard Decisions: an Introduction to Decision Analysis. Second Edition. Duxbury Press, Wadsworth Publishing Company, 1996.

[3]. Fukunaga, K. Introduction to the Statistical Pattern Recognition. Second Edition. Academic Press, 1990.

[4]. Tenekedjiev, K., Karakatsanis, N., Bekiaris, A., Fictitious Covariance Matrices, Proc. Forth International Conference, Adaptive Computing in Design and Manufacture ACDM'2000, 2000, pp. 23-26, Plymouth, UK.

За контакти:

Доц. д-р Наталия Николова, катедра „Икономика и мениджмънт", Технически университет - Варна, тел.: 052-383 670, e-mail: natalianik@gmail.com

Докладът е рецензиран.