Нелинейные преобразования признакового пространства и их аналитические представления

Саидов Дониёр Юсупович

Ассистент

Механико–математический факультет НУУз имени Мирзо Улугбека, Ташкент, Узбекистан

E-mail: doniyor_2286@mail.ru

Рассматривается двухклассовая задача распознавания в стандартной постановке. Объекты выборки $E_0 = \{S_1, \ldots, S_m\}$ принадлежат одному из классов K_1 или K_2 ($E_0 = K_1 \cup K_2$) и описываются с помощью набора признаков $X(n) = (x_1, \ldots, x_n)$.

На E_0 определена операция попарного объединения признаков x_i и x_j путем нелинейного отображения их значений в описании объектов на числовую ось. Процесс объединения реализуется с помощью иерархической агломеративной процедуры на основе специального правила. Результатом объединения является латентный признак.

Пусть $\{x_i^p\}_{i \in I}$ — множество признаков, полученное на p–ом ($0 \le p < n$) шаге иерархической агломеративной процедуры, I – множество номеров исходных признаков и $I = \{1, \ldots, n\}$ при $p = 0$. Решение об объединении признаков принимается посредством критерия

$$R(x_i^p) = \left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} u_i^d (u_i^d - 1)}{\sum\limits_{i=1}^{2} |K_i|(|K_i| - 1)} \right) \left(\frac{\sum\limits_{d=1}^{2} \sum\limits_{i=1}^{2} u_i^d (|K_{3-i}| - u_{3-i}^d)}{2|K_1||K_2|} \right) \to \max\limits_{c_0 < c_1 < c_2}, \tag{1}$$

используемого для разбиения упорядоченного набора из m значений $r_{i_1}, r_{i_2}, \ldots, r_{i_m}$ признака x_i^p в описании объектов E_0 на интервалы $[c_0, c_1]$ и $(c_1, c_2]$, $c_0 = r_{i_1}, c_2 = r_{i_m}$. Переменная $u_1^1(u_1^2)$ представляет количество значений x_i^p у объектов из K_1 в интервале $[c_0, c_1]((c_1, c_2])$. Соответственно интерпретируется переменная $u_2^1(u_2^2)$ для объектов из класса K_2.

Экстремум (1) соответствует значению $R\,(x_i^p) = 1$ и достигается при условии, что в каждом из интервалов $[c_0, c_1]$ или $(c_1, c_2]$ содержатся все представители (значения признака) только одного класса K_1 или K_2.

Значение нового латентного признака на p–ом шаге ($p \geq 1$) иерархической агломеративной процедуры при объединении признаков $x_i^{p-1}, x_j^{p-1}, (i < j)$ вычисляется по формуле

$$x_i^p = (1 - \alpha_{ij}) \left(t_i w_i^{p-1} \frac{(x_i^{p-1} - c_{i1}^{p-1})}{(c_{i2}^{p-1} - c_{i0}^{p-1})} + t_j w_j^{p-1} \frac{(x_j^{p-1} - c_{j1}^{p-1})}{(c_{j2}^{p-1} - c_{j0}^{p-1})} \right) +$$

$$+ \alpha_{ij} t_{ij} w_{ij}^{p-1} \frac{(x_i^{p-1} x_j^{p-1} - c_{ij1}^{p-1})}{(c_{ij2}^{p-1} - c_{ij0}^{p-1})}, \qquad (2)$$

где $0 \leq \alpha_{ij} \leq 1, t_i, t_j, t_{ij} \in \{-1, 1\}, w_i^{p-1} = R(x_i^{p-1}), w_j^{p-1} = R(x_j^{p-1})$, $w_{ij}^{p-1} = R(x_i^{p-1} x_j^{p-1}), c_{i0}^{p-1}, c_{i1}^{p-1}, c_{i2}^{p-1}, c_{j0}^{p-1}, c_{j1}^{p-1}, c_{j2}^{p-1}, c_{ij0}^{p-1}, c_{ij1}^{p-1}, c_{ij2}^{p-1}$ — значения границ интервалов по (1).

Обозначим через F – множество парных комбинаций номеров признаков из I. Для объединения признаков $x_i^{p-1}, x_j^{p-1}, i, j \in I$ при $i < j$ по (2) используется правило из двух условий:

$$R(x_i^p) > \max_{\{u,v\} \in F \setminus \{i,j\}} R(x_u^p), u < v; \qquad (3)$$

$$R(x_i^p) > \max \left(R(x_i^{p-1}), R(x_j^{p-1}) \right). \qquad (4)$$

При выполнении условий правила $I = I \setminus \{j\}$. Если условие (3) истинно а (4) ложно, то производится вывод значений латентного признака x_i^{p-1} и $I = I \setminus \{i\}$.

Вычислительный эксперимент проводился на медицинских данных с показателями гипертонической болезни. Выборка из 147 объектов была разделена на два класса: K_1 (здоровые) содержал показатели 111 объектов, K_2 (больные) – 36 объектов. Каждый объект описывался 29-ю признаками. Последовательность вычисления значений первого (в порядке формирования) латентного признака для описания объектов была такой:

$x_4^1 = -0.00695(\texttt{АДС} - 140) + 0.39493(\texttt{ДИАСТОЛА} - 0.42) - 0.00413(\texttt{АДС} * \texttt{ДИАСТОЛА} - 68.2)$; $x_4^2 = 0.80135(x_4^1 - 0.0094) + 3.12871(\texttt{QRS} - 0.08)$;

$x_4^3 = 0.58119 x_4^2 + 0.55186(\texttt{СИСПОК} - 0.485) + 0.89202(x_4^2 * \texttt{СИСПОК})$.

Значение $R\left(x_4^3\right) = 1$ указывает на то, что нелинейные отображение по (2) на числовую ось формируют латентный признак, значения которого позволяют корректно (без ошибок) разделять объекты выборки E_0 на классы с использованием 4-х (из 29) исходных признаков.