НЕПАРАМЕТРИЧЕСКАЯ КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ НА ОСНОВЕ ЯДЕРНЫХ ОЦЕНОК ПЛОТНОСТЕЙ С ПРЯМОУГОЛЬНЫМ ЯДРОМ

Паланевич А. С., Жук Е. Е.
Кафедра математического математического моделирования и анализа данных,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: apalanevich@gmail.com, zhukee@mail.ru

В работе рассматривается задача статистической классификации с применением байесовского решающего правила и прямоугольного ядра для оценивания плотностей распределения.

Введение

Статистическая классификация данных является одной из самых известных задач прикладного анализа данных. Она часто возникает в таких областях, как медицина, экономика и т.д. Для решения данной проблемы существует множество методов и алгоритмов, один из которых я бы хотел рассмотреть в своей работе.

Для изучения был выбран метод, основанный на ядерной оценке плотности распределения. В качестве ядра использовалось прямоугольное ядро, так как оно является довольно простым и позволяет особенным образом интерпретировать полученные с его помощью результаты.

I. Описание модели

Пусть регистрируются случайные наблюдения $x = x(w) \in R^N$ над объектами $w \in \Omega$, принадлежащими к L классам $\{\Omega_1, \Omega_2, \dots, \Omega_L\}$:

$$\Omega_i \cap \Omega_j = \emptyset, i \neq j, i, j = 1, 2, ..., L;$$

$$\bigcup_{i=1}^{L} \Omega_i = \Omega;$$

Обозначим истинный номер класса, к которому принадлежит объект w, через $d^0(w)$. Этот номер является дискретной случайной величиной со следующим распределением:

$$P(d^0(w) = i) = \pi_i, i = 1, 2, \dots, L;$$

$$\pi_1 + \pi_2 + ... + \pi_L = 1;$$

Здесь $\{\pi_i, i = \overline{1, L}\}$ – априорные вероятности классов. В рамках каждого из классов наблюдение $x(w)$ описывается условной плотностью распределения:

$$p_i(x) = p(x|d^0(x) = i), i = 1, 2, ..., L, x \in R^N;$$

II. Постановка задачи и построение оценок

Поставим перед собой задачу оценки номера класса для нового наблюдения по имеющейся классифицированной выборке (задача дискриминантного анализа). Для этого будем пользоваться байесовским решающим правилом (БРП) [1]:

$$\widehat{d^0}(x) = \arg\max_{i=1,2,...,L} (\pi_i p_i(x)), x \in R^N,$$

где $\widehat{d^0}(x)$ – оценка номера неизвестного класса для наблюдения x. Мы имеем дело с нерандомизированным решающим правилом [1]. Однако для пользования этим правилом необходимо знать $\{\pi_i, p_i(x), i = 1, 2, \dots, L\}$. Так как их точные значения неизвестны, то укажем способ построения оценок для этих величин.

Пусть $X = \{x_1, x_2, \dots, x_n\} \in R^{nN}$ – классифицированная выборка и n_k – число наблюдений из выборки, которые относятся к классу с номером k. Тогда построим оценку для априорных вероятностей [1]:

$$\widehat{\pi_k} = \frac{n_k}{n}, k = 1, 2, ...L. \quad (1)$$

Условные плотности распределения будем оценивать с помощью ядерных оценок с прямоугольным ядром. Пусть $\Gamma(x)$ – N-мерный параллелепипед с центром в точке x и сторонами $h_i \in R, i = 1, 2, \dots, N$, с «объемом» $V = \prod_{i=1}^{N} h_i$. Вводя функцию-индикатор $I_{\Gamma(x)}(y)$, равную единице, если $y \in \Gamma(x)$, и нулю – в противном случае, оценки плотностей запишем следующим образом [2]:

$$\widehat{p_k}(y) = \frac{1}{n_k V} \sum_{j=1}^{n} I_{\Gamma(y)}(x_j) \delta_{k, d^0(x_j)}; \quad (2)$$

$$x_j \in X, j = 1, 2, ..., n, k = 1, 2, ..., L, y \in R^N.$$

Тогда для оценки номера класса, к которому принадлежит новое наблюдение x^*, получаем следующее подстановочное БРП [1]:

$$\widehat{d}(x^*) = \underset{k=1,2,...,L}{\arg\max} (\widehat{\pi_k}\widehat{p_k}(x^*)), x^* \in R^N. \quad (3)$$

III. Интерпретация результатов

Теперь сформулируем и докажем теорему, придающую построенному БРП содержательный смысл.

Теорема. Пусть по обучающей выборке $X = \{x_1, x_2, \ldots, x_n\} \in R^{nN}$ построены упомянутые выше оценки параметров модели $\{\widehat{\pi_k}, \widehat{p_k}(x), k = 1, 2, ..., L\}$ из (1), (2). Тогда новое наблюдение $x^* \in R^N$, классифицируемое с помощью подстановочного БРП (3), относится к классу Ω_d с номером d, когда количество наблюдений из выборки X, лежащих в $\Gamma(x^*)$ и принадлежащих этому классу, является наибольшим среди остальных классов.

Доказательство. Для простоты рассуждений предположим, что максимум в БРП достигается на одном классе с номером d. Тогда:

$$\widehat{d}(x^*) = d \Leftrightarrow \widehat{\pi_d}\widehat{p_d}(x^*) > \widehat{\pi_k}\widehat{p_k}(x^*), k = \overline{1, L}, k \neq d.$$

Подставляя полученные ранее оценки параметров модели (1), (2) и сокращая множители в дробях, получаем:

$$\sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)\delta_{d,d^0(x_j)} > \sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)\delta_{k,d^0(x_j)}$$

$$\sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)(\delta_{d,d^0(x_j)} - \delta_{k,d^0(x_j)}) > 0.$$

Это неравенство возможно лишь в случае, когда количество наблюдений x, на которых функция $\delta_{d,d^0(x)}$ обращается в единицу (то есть $d = d^0(x)$) больше, чем количество аналогичных наблюдений для функции $\delta_{k,d^0(x)}$. Это как раз и соответствует сформулированной выше интерпретации построенного БРП.

Замечание. В случае, когда максимум подстановочного БРП достигается на нескольких классах (когда в параллелепипеде вокруг нового наблюдения сразу несколько классов имеют наибольшую частоту), можно пропорционально увеличить стороны $\Gamma(x^*)$ до тех пор, пока количество элементов какого-либо класса не станет наибольшим. Аналогично можно поступить, когда в $\Gamma(x)$ вообще не попало ни одно наблюдение.

Замечание. Здесь в качестве параметров классификатора выступают стороны параллелепипеда $h_i, i = 1, 2, ..., N$. Их можно выбрать, например, из тех же соображений, что используются при выборе размеров ячейки для построения гистограмм (правило Стерджеса или правило квадратного корня). В таком случае для каждого измерения длина стороны параллелепипеда будет своей.

Замечание. Вместо параллелепипеда можно рассматривать N-мерный куб, то есть положить $h_i = h, i = 1, 2, ..., N$. В таком случае для выбора длины стороны удобно взять сетку $h^{(1)}, h^{(2)}, ..., h^{(m)}$ и посмотреть, на каком из значений точность классификатора будет наибольшей.

Заключение

Таким образом, мы получили довольно интересную и интуитивно понятную интерпретацию решающего правила для статистической классификации с использованием прямоугольных ядерных оценок. Это делает построенный метод довольно простым для применения на практике.

Также хочется заметить, что полученный результат, в каком-то смысле, роднит исследуемый в этой работе метод с классификацией наблюдений на основе оценок плотностей с применением гистограмм с прямоугольными ячейками. Основное отличие заключается в том, что в гистограммном случае ячейки имеют фиксированные позиции, а в случае ядерных оценок ячейка строится вокруг нового наблюдения.

Список литературы

1. Жук, Е. Е., Харин, Ю. С. Математическая и прикладная статистика // учебное пособие. – Минск, БГУ, 2005. – с. 192–208.
2. Multivariate Kernel Smoothing and its Applications / Jose E. Chacon, Tarn Duong – Taylor & Francis Group, LLC, 2018.