

УДК 51

ИСПОЛЬЗОВАНИЕ ЯДЕРНЫХ ОЦЕНОК ПЛОТНОСТИ В ЗАДАЧАХ ИДЕНТИФИКАЦИИ

Шаманова Любовь Сергеевна

студентка

ФМФиТ Магистры МАТ- 1 КАЛМГУ

Республика Калмыкия, Целинный район, с. Троицкое

Аннотация. Одной из характерных черт современного состояния социально- экономических наук является широкое применение в них методов математической статистики. При этом практика выдвигает все новые и новые задачи, для решения которых требуются и новые методы. Таковыми, в частности, являются задачи классификации.

Байесовское решение задачи классификации. Использование принципа максимума апостериорной вероятности при классификации с обучением. Оценка априорных вероятностей и плотностей вероятностей по обучающей выборке. Параметрический подход с использованием линейного дискриминанта Фишера. Непараметрический подход с использованием оценки плотности распределения методом Розенבלата- Парзена.

Одной из характерных черт современного состояния социально- экономических наук является широкое применение в них методов математической статистики. При этом практика выдвигает все новые и новые задачи, для решения которых требуются и новые методы. Таковыми, в частности, являются задачи классификации.

Любая схема классификации основывается на одном общем фундаментальном принципе, состоящем из двух положений:

1. В один класс объединяются объекты, сходные между собой в некотором смысле.
2. Степень сходства между собой у объектов, принадлежащих к одному классу, должна быть больше, чем степень сходства между собой у объектов, относящихся к разным классам.

Одной из наиболее строгих статистических процедур распознавания является дискриминантный анализ.

Предметом исследования является задача статистической идентификации.

Цель исследования – теоретическое изучение задач дискриминантного анализа.

Задачи исследования:

- Общая характеристика методологического подхода к решению задачи упорядоченной классификации.
- Разработка одной модели дискриминантного анализа, именно классификатора, основанного на принципе максимума апостериорной вероятности с проверкой пяти ядер и четырех параметров сглаживания.

Методы исследования:

- Байесовское решение проблемы идентификации в рамках непараметрических методов;
- Восстановление плотности вероятностей в классе непрерывных функций методами Парзена.

Сущность байесовского метода решения проблемы классификации.

Проблема классификации возникает, когда исследователь делает некоторое число измерений, связанных с каким-то индивидуумом, и на основе этих измерений хочет отнести его к одной из нескольких категорий. Он не может непосредственно определить категорию, к которой относится индивидуум, и вынужден использовать эти измерения. Во многих случаях можно предположить, что имеется конечное число категорий или генеральных совокупностей, из которых мог быть взят рассматриваемый индивидуум, причем каждая из этих категорий характеризуется определенным законом распределения вероятностей для измерений. Таким образом, индивидуум рассматривается как случайное наблюдение над этой генеральной совокупностью. Вопрос ставится так: как по результатам измерений определить, из какой генеральной совокупности взят данный индивидуум. [4]

Байесовское решение задачи классификации

Рассмотрим модельную постановку задачи классификации. Пусть множество исследуемых объектов разбито на K классов или групп H_k . Каждый объект характеризуется набором $\vec{X}$ числовых параметров, которые непосредственно могут быть измерены, т. е. точкой в пространстве признаков. Предполагается, что $\vec{X}$ – случайная величина, а принадлежность объектов к разным классам выражается в том, что для объектов из класса H_i , это случайная величина имеет распределение f_k . Задача состоит в том, чтобы по наблюдению над $\vec{X}$ определить тот класс, к которому принадлежит соответствующий объект. [7]

Использование принципа максимума апостериорной вероятности при классификации с обучением. [1]

Для решения задачи классификации в рамках байесовского подхода с использованием

принципа максимума апостериорной вероятности нам необходимо знать следующее:

1. Априорные вероятности π_k генеральных совокупностей, которые являются относительными частотами объектов составной генеральной совокупности, из которой извлечен объект, подлежащий идентификации.

2. Плотности вероятностей $f_k(\vec{X})$ данного набора измерений $\vec{X}$ над объектом в K альтернативных генеральных совокупностях.

Тогда, если дан объект с измерениями $\vec{X}$, то объект относят к той совокупности, номер которой определяется из условия (1)

$$S_k(\vec{X}) = \max_{1 \leq j \leq K} S_j(\vec{X}), \quad (1)$$

где

$$\{S_k(\vec{X}) = \ln \pi_k + \ln f_k(\vec{X}), k = 1, \dots, K \quad (2)$$

- набор информантов для наблюдения $\vec{X}$. Правило (1) минимизирует математическое ожидание числа ложных идентификаций. Величины $\pi_k f_k(\vec{X})$ пропорциональны апостериорным вероятностям классов, поэтому в данном случае, байесовский принцип максимума апостериорной вероятности сводится к следующему: относить объект с характеристикой $\vec{X}$ к тому классу, апостериорная вероятность которого максимальна.

Оценка априорных вероятностей и плотностей вероятностей по обучающей выборке. [3,6]

Пусть имеющаяся база данных содержит результаты обследования N индивидуумов и пусть эти N объектов разбиты на K групп в соответствии с характером заболевания. Пусть N_k - число объектов k -ой группы. Тогда

$$\sum_{i=1}^K N_k = N, \quad (3)$$

а оценки априорных вероятностей определяются соотношениями

$$\pi_k = \frac{N_k}{N}, k = 1, \dots, K. \quad (4)$$

Параметрический подход с использованием линейного дискриминанта Фишера. [5]

В этом случае распределение случайной величины U предполагается r -мерным нормальным в каждом из совокупностей. Линейный дискриминант Фишера опирается на предположение, что ковариационные матрицы классов равны. В этом случае ковариационная матрица оценивается по обучающим объектам всех классов, что повышает устойчивость оценки, особенно в случае малочисленных классов. Тогда расчет информантов производится по следующему алгоритму. Представлен в формулах (5-11).

Вычисляются групповые средние

$$\bar{x}_k^j = \frac{1}{N_k} \sum_{i=1}^{N_k} x_{ki}^j, k = 1, \dots, K, \quad j = 1, \dots, p. \quad (5)$$

Вычисляются суммы взаимных произведений от средних

$$S_k^{jm} = \sum_{i=1}^{N_k} (x_{ki}^j - \bar{x}_k^j)(x_{ki}^m - \bar{x}_k^m), k = 1, \dots, K, \quad j, m = 1, \dots, p \quad (6)$$

Вычисляется объединенная ковариационная матрица

$$D_{jm} = \frac{1}{N-K} \sum_{k=1}^K S_k^{jm}, \quad j, m = 1, \dots, p \quad (7)$$

Вычисляются общие средние

$$\bar{X}_j = \frac{1}{N} \sum_{k=1}^K N_k \bar{x}_k^j, \quad j = 1, \dots, p. \quad (8)$$

Вычисляется обобщенная D^2 статистика Махаланобиса

$$V = \sum_{i=1}^p \sum_{j=1}^p (D^{-1})_{ij} \sum_{k=1}^K N_k (\bar{x}_k^i - \bar{X}^i)(\bar{x}_k^j - \bar{X}^j). \quad (9)$$

Значение V может быть использовано как хи- квадрат для проверки гипотезы, что средние значения одинаковы во всех K группах для данных p переменных. Вычисляются коэффициенты линейных дискриминантных функций

$$C_0^0 = -\frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p (D^{-1})_{ij} \bar{x}_k^i \bar{x}_k^j, \quad C_k^j = \sum_{i=1}^p (D^{-1})_{ij} \bar{x}_k^i, \quad j = 1, \dots, p \quad (10)$$

Тогда информанты определяются соотношениями

$$S_k(\bar{x}) = C_k^0 + \sum_{j=1}^p C_k^j + \ln \pi_k. \quad (11)$$

Непараметрический подход с использованием оценки плотности распределения методом Розенבלата- Парзена.

Сущность данного метода заключается в следующем. Плотность вероятности $f_k(\vec{X})$ оценивается как функция правдоподобия

$$f_k(\vec{x}) = \prod_{j=1}^p f_k(x^j), \quad (12)$$

где $f_k(\vec{x})$ – одномерная плотность распределения, оценка которой осуществляется с помощью непараметрического метода Парзена

$$f_k(x^j) \approx \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{b_k^j} K\left(\frac{D_{ki}^j}{b_k^j}\right). \quad (13)$$

Здесь где $D_{ki}^j = |x_{ki}^j - x^j|$ – расстояния от j -ой координаты наблюдения $\vec{X}$ от j -ой координаты всех базовых наблюдений k -ого класса, b_k^j – параметры сглаживания, $K(u)$ – функция (борелевское ядро), удовлетворяющее условиям

$$K(u) \geq 0, K(-u) = K(u), \quad \lim_{|u| \rightarrow \infty} K(u) = 0, \quad \int_{-\infty}^{+\infty} K(u) du = 1 \quad (14)$$

При программной реализации описанного алгоритма можно использовать следующие типы борелевских ядер. [2,8]

Стандартное ядро Гаусса

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \quad (15)$$

два типа ядер с компактным носителем: ядро Епанечникова

$$K(u) = \begin{cases} \frac{3}{4} (1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases} \quad (16)$$

и равнобедренное треугольное ядро

$$K(u) = \begin{cases} 1 - |u|, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}, \quad (17)$$

а также следующие два типа ядер: тригонометрическое ядро

$$K(u) = \frac{1}{\pi} \left(\frac{\sin u}{u} \right)^2 \quad (18)$$

Ядро Валле – Пуссена

$$K(u) = \frac{1}{2\pi} \left(\frac{\sin \frac{u}{2}}{\frac{u}{2}} \right)^2. \quad (19)$$

Список литературы:

1. Антонов А. В., Чепурко В. А. Построение непараметрической плотности распределения на основании цензурированной информации // Надёжность. — М.: Издательский дом «Технология», 2005, № 2. - С. 3.
2. Антонов А.В. Системный анализ. — М.: Высшая школа, 2004.—454 с.
3. Деврой Л., Дьёрфи Л. Непараметрическое оценивание плотности. L1-подход. / Пер. с англ.— М.: Мир. 1988.—408с.
4. Пенская М.Я. Об оценках априорной плотности. - В кн.: Статистические методы. Пермь:1982, с. 115-
5. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989;
6. Антонов А. В. Системный анализ. – М.: Высшая школа, 2004. -454с.;
7. Деврой Л., Дьёрфи Л. Непараметрическое оценивание плотности. / Пер. с англ. – М.: Мир 1988.- 408с.;
8. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и её применения. — 1969;
9. Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А. Обучающиеся системы

обработки информации и принятия решений. Непараметрический подход. — Новосибирск: Наука, 1996

10. Пенская М. Я. Об оценках априорной плотности.- В.кн.: Статистические методы. Пермь: 1982.
11. Субботин С. В., Большаков Д. Ю. Применение байесовского классификатора для распознавания классов целей. // «Журнал Радиоэлектроники», 2006, № 4.
12. Хардле В. Прикладная непараметрическая регрессия. — М.: Мир, 1993.

ISSN 2541-9250

НАУКА ЧЕРЕЗ ПРИЗМУ ВРЕМЕНИ

Международный научный журнал

Ежемесячное научное издание

№12 (45) 2020

ISSN 2541-9250



9 772541 925005 >

Библиографическое описание статьи:

Шаманова Л.С. Использование ядерных оценок плотности в задачах идентификации // Наука через призму времени. – 2020. – №12 (45). – с. 12-16



<http://www.naupri.ru/journal/2783>

© Шаманова Л.С., 2020

© Редакция научного журнала «Наука через призму времени», 2020