

ПРИМЕНЕНИЕ КЛАССИФИКАЦИОННЫХ ПРОЦЕДУР ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ УМЕНЬШЕНИЯ ИНФОРМАЦИОННОЙ НЕОПРЕДЕЛЕННОСТИ МНОГОМЕРНЫХ ИЗМЕРИТЕЛЬНЫХ СИГНАЛОВ

Предложена процедура классификации результатов измерений уровней метрологически неопределенного параметра контроля, стохастически связанного с многомерным сигналом измерительной информации. Показано, что оптимальное число уровней (классов) определяется минимумом погрешности измерения.

параметрическая дискриминация, функция правдоподобия, вероятность ошибки, погрешность измерения

Постановка проблемы. Основной проблемой измерительного контроля параметров биологически сложного сельскохозяйственного сырья является отсутствие стандартных образцов, воспроизводящих заданные уровни этих параметров. Это приводит к метрологической неопределенности уровней косвенного измерения, коэффициентами которых являются единичные показатели контроля, стохастически связанные с контролируемым параметром.

Анализ литературы. Чаще всего для снижения априорной неопределенности результатов измерительного эксперимента при градуировке (обучении) системы измерительного контроля применяют регрессионные модели, основанные на минимизации остаточных погрешностей [1], или методы фиксации воспроизводимых при градуировке уровней показателей контроля на основе физического моделирования структуры контролируемого сырья [2].

Цель исследований – раскрытие возможностей дискриминантного анализа многомерных данных для получения дополнительной информации о метрологически неопределенных уровнях параметра контроля сыпучих материалов со сложной биологической структурой.

Модель параметрической дискриминации. Рассмотрим параметр контроля Y , значения которого принадлежат диапазону A_y . Косвенное измерение значений Y осуществляют по набору x_1, \dots, x_q показателей контроля X_1, \dots, X_q . Такие показатели образуют q -многомерный измерительный сигнал, содержащий дополнительную информацию о параметре Y в форме вероятностной модели взаимного влияния $(q + 1)$ -ой случайных величин (Y, X_1, \dots, X_q) [3].

Если диапазон A_y разделить на K поддиапазонов (классов) и если для j -го ($j = 1, K$) класса w_j известна совместная плотность распределения вероятности $f(x_1, \dots, x_q/w_j)$, то вероятность отнесения вектора $\bar{X} = (X_1, \dots, X_q)$ к классу w_j вычисляют как

$$P(\bar{X} \in w_j) = \sum_{j=1}^k P_j \iint_{R_j} \dots \int f(x_1, \dots, x_q / w_j) dx_1 dx_2 \dots dx_q, \quad (1)$$

где R_j – подмножество значений X_1, \dots, X_q , принадлежащих классу w_j с априорной вероятностью P_j . Основная трудность, возникающая при использовании выражения (1) – это выбор R_1, \dots, R_q , максимизирующих вероятность $P(\bar{X} \in w_j)$.

Множество R_1, \dots, R_q взаимнооднозначно определяет число и размер поддиапазонов $\Delta A_1, \dots, \Delta A_k$, причем всегда существует минимальное k , максимизирующее вероятность правильной классификации номера j любого из поддиапазонов $\Delta A_j \in A_y$ [4].

При отсутствии априорной информации о вероятностях P_j и виде условного распределения $f(x_1, \dots, x_q/w_j)$, последнее достаточно хорошо [5] моделируется нормальным распределением

$$f(x_1, \dots, x_q / w_j) = \frac{|C_j|^{1/2}}{(2\pi)^{q/2}} \times \exp \left[-\frac{1}{2} \sum_{r=1}^q \sum_{s=1}^q c_{rsj} (x_r - \bar{X}_{rj})(x_s - \bar{X}_{sj}) \right], \quad (2)$$

где \bar{X}_{rj} – математическое ожидание показателя контроля X_r в j -м классе (w_j); матрица $C = (c_{rsj})$ является обратной к матрице дисперсий и ковариационной матрице показателей X_1, \dots, X_q в j -ом классе.

Функция (2) при замене постоянных C , \bar{X}_{rsj} , \bar{X}_{sj} их оценками может рассматриваться как оценка функции правдоподобия выборочных значений x_1, \dots, x_q и служить целевой функцией

$$g_j(\bar{X}) = f(x_1, \dots, x_q / w_j)$$

выбора одного из набора решений: $\begin{cases} \gamma_1 : Y \in \Delta A_1; \\ \gamma_k : Y \in \Delta A_k. \end{cases}$

Тогда принимают решение γ_j , если для всех $j \neq i$

$$g_i(\bar{X}) > g_j(\bar{X}).$$

В этом случае:

$$g_j(\bar{X}) = \max P(\bar{X} \in w_j).$$

Априорные вероятности P_j считают одинаковыми для всех $j = \overline{1, k}$ и равными $P_j = 1/k$.

Число классов w_j не может выбираться произвольно, особенно, если мал объем N выборочных значений вектора \overline{X} , используемых для градуировки системы измерительного контроля. Для иллюстрации этого факта достаточно рассмотреть случай двухкомпонентного вектора $\overline{X}(q=2)$, для которого показатель, (обозначим его как ξ_j^2), экспоненты в правой части выражения (2) является линейно преобразованной случайной величиной с центральным χ -квадрат распределением и двумя степенями свободы. Фактически, это экспоненциальное распределение с параметром [4]

$$\lambda = \frac{N-2m}{2N\sqrt{1+\rho^2}},$$

где ρ – нормированный коэффициент парной корреляции между показателями X_1 и X_2 :

$$f(\xi_j^2) = \lambda \exp(-\lambda \cdot \xi_j^2). \quad (3)$$

Используя уравнение (3) можно определить вероятности ошибок классификации номера поддиапазона, если заданы:

а) диапазоны A_1 и A_2 возможных значений показателей контроля X_1 и X_2 , обеспечивающих условие $Y \in A_y$;

б) дисперсии σ_1^2 и σ_2^2 этих сигналов.

Тогда, например, вероятность ошибки первого рода, при условии $Y \in A_j$, равна

$$\alpha_j = \int_{Y \notin \Delta A_j} f(\xi_j^2) d\xi_j^2.$$

Интегрирование функции (3) и усреднение вероятности ошибки по всем классам дает выражение:

$$M[\alpha_k] = \exp \left\{ - \frac{(N-2K) \left[\sum_{r=1}^2 \left(\frac{A_r}{\sigma_r} \right)^2 - 2|\rho| \prod_{r=1}^2 \left(\frac{A_r}{\sigma_r} \right)^2 \right]}{8Nm^2 (1-\rho^2)(1+\rho^2)^{1/2}} \right\}. \quad (4)$$

Выражение (4) показывает, что вероятность ошибки, например, первого рода, является функцией:

1. Числа образцов, используемых для обучения (градуировки) системы измерительного контроля, определяющих априорную метрологическую неопределенность средних значений показателей контроля в поддиапазонах ΔA_j параметра Y .

2. Числа k поддиапазонов измерения.

3. Коэффициента парной корреляции ρ между показателями контроля.

Из уравнения (4) видно, что при фиксированных N и ρ математическое ожидание $M[\alpha_k]$ является не-

линейной функцией числа k . Учет вероятности ошибки второго рода и оценки полной вероятности ошибки классификации, как

$$\overline{P} = \frac{1}{2} \left\{ \alpha_k + \sum_{r=1}^{k-1} \left[\int_{Y \in \Delta A_k} f(\xi_r^2) d\xi_r^2 \right] \right\},$$

позволяет выбрать такое значение k , для которого \overline{P} – минимальна.

Дисперсия σ_y^2 результата измерения значения параметра Y может быть определена, как математическое ожидание дискретной случайной величины:

$$\Delta Y = \begin{cases} y_1, & \text{если } Y \in \Delta A_j; \\ y_2, & \text{если } Y \notin \Delta A_j, \end{cases}$$

при известных вероятностях ошибок первого (α_k) и второго ($\beta_k = 2\overline{P} - \alpha_k$) рода

$$\sigma_y^2 = y_1^2 \cdot \alpha_k + y_2^2 \cdot \beta_k.$$

Например, при $\rho \neq 0$, имеем:

$$\sigma_y^2 = \left(\frac{A_y}{2k} \right)^2 \left[1 + \left(\frac{k^2}{4} - 1 \right) \alpha_k \right], \quad (5)$$

где α_k определяется выражением (4).

Практическое использование моделей дискриминации. Для экспериментальной проверки моделей дискриминации на объектах измерительного контроля в качестве последних использовались сыпучие материалы:

а) семена подсолнечника (параметр контроля – относительна влажность);

б) зерно пшеницы (параметр контроля – процентное содержание клейковины).

Показателями контроля являлись соответственно:

а) X_1 – насыпная плотность семян подсолнечника;
 X_2 – тангенс угла диэлектрических потерь;

б) X_1 – относительная влажность пшеницы;
 X_2 – масса 1000 семян;

X_3 – коэффициент теплопроводности дозированной пробы.

В качестве функции $f(x_1, \dots, x_q/w_j)$ использовалось, соответственно, двух- и трехмерное невырожденное нормальное распределение.

В табл. 1 представлены значения среднеквадратических погрешностей оценивания уровней параметров контроля при заданном числе k числа поддиапазонов (классов) измерения.

Таблица 1

Среднеквадратические погрешности измерения, %

Объект контроля	Подсолнечник	Пшеница
Число показателей (q)	2	3
Параметр контроля	влажность	клейковина
k = 2	1,19	2,71
k = 3	0,96	2,07
k = 4	1,12	2,22
k = 5	1,27	2,39
k = 6	1,31	2,45

Количество многократных измерений по пробам фиксированного уровня равнялось $n = 30$.

Вывод. Из табл. 1 видно, что минимум погрешности измерения соответствует $k = 3$, что указывает на оптимизационный характер классификационных моделей измерительного контроля и соответствие функциональной зависимости, задаваемой уравнением (5).

ЛИТЕРАТУРА

1. *Петров И.К.* Технологические измерения и приборы в пищевой промышленности. – М.: Агропромиздат, 1985. – 224 с.

2. *Овчаренко А.И., Комирный А.С.* Исследование инструментальной погрешности влагомера ВЦЛ-11 и спосо-

бы ее уменьшения // Вестник НТУ «ХПИ». – Х.: НТУ «ХПИ», 2003. – № 7, т. 3. – С. 123-126.

3. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 432 с.

4. *Щапов П.Ф., Качанов М.П.* Классификационные задачи при двухпараметровом измерении влажности. // Украинский метрологический журнал. – 2002. – № 4. – С. 12-14.

5. *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке: Методы планирования эксперимента: Пер. с англ. – М.: Мир, 1981. – 588 с.

Поступила 15.03.2006

Рецензент: канд. техн. наук И.П. Захаров, Харьковский национальный университет внутренних дел.