ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2017, том 27, № 3, с. 33–42

ПРИБОРОСТРОЕНИЕ ДЛЯ БИОЛОГИИ И МЕДИЦИНЫ

УДК 543.51; 681.2–5

© В. В. Манойлов, Ю. А. Титов, А. Г. Кузьмин, И. В. Заруцкий

АЛГОРИТМЫ ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ КЛАССИФИКАЦИИ МАСС-СПЕКТРОВ ВЫДЫХАЕМЫХ ГАЗОВ

В работе рассматриваются алгоритмы дискриминантного анализа для классификации масс-спектров выдыхаемых газов. Показывается, что для вычисления коэффициентов дискриминантных функций могут быть использованы три алгоритма: 1) алгоритм, основанный на QR-разложении; 2) алгоритм, основанный на вычислении обобщенной корреляционной функции, и 3) алгоритм, основанный на решении переопределенной системы линейных уравнений методом наименьших квадратов. Приводятся таблицы вероятностей принадлежности обрабатываемого масс-спектра к одной из двух групп: здоровых или больных людей. В работе показывается, что в качестве переменных для проведения классификации могут быть использованы как непосредственно амплитуды масс-спектральных пиков выдыхаемого газа, так и переменные, получающиеся в результате сокращения размерности обрабатываемых данных методом главных компонент. При этом показывается, что для оценки принадлежности к одной из указанных групп достаточно использовать первые две главные компоненты. Приводятся примеры апробации предлагаемых методов.

Кл. сл.: масс-спектрометр для анализа выдыхаемых газов, линейный дискриминантный анализ, классификация масс-спектров

ВВЕДЕНИЕ

Настоящая работа является развитием результатов, полученных авторами в работе [1]. Данная работа имеет следующие цели:

- описание трех различных алгоритмов вычисления коэффициентов дискриминантных функций;
- определение минимального количества главных компонент, необходимых для принятия решения о принадлежности объекта к одному из двух классов;
- вычисление вероятностей принятия правильного и ошибочного решений о принадлежности обрабатываемого масс-спектра к одному из двух классов.

Дискриминантный анализ является разделом многомерного статистического анализа [2–5], который включает в себя методы классификации многомерных наблюдений по принципу максимального сходства анализируемых наблюдений с наблюдениями, отнесенными к определенным классам по результатам обучения. В данной работе методы дискриминантного анализа рассмотрены на примере классификации людей по группам здоровья на основе анализа масс-спектров выдыхаемых газов. Такая классификация может быть полезна в том числе при проведении массовых скрининговых профилактических осмотров населения. В качестве примера использования данного метода в аналитическом приборостроении рассматривается классификация масс-спектров, полученных на квадрупольном масс-спектрометре МС7-200, при анализе наличия патологий по спектрам выдыхаемых газов [6]. Исследования возможностей выявления патологий по масс-спектрам выдыхаемых газов были описаны в работах [7–9].

В работе [1] показано, что дискриминантными признаками для классификации могут служить значения амплитуд на определенных массах линейчатого масс-спектра и значения переменных, полученных после преобразования обрабатываемых масс-спектров методом главных компонент.

1. АЛГОРИТМЫ ВЫЧИСЛЕНИЯ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ

1.1. Алгоритм на основе QR-разложений

QR-разложение представляет исходную матрицу **A** в виде $\mathbf{A} = \mathbf{Q}\cdot\mathbf{R}$, где **Q** — ортогональная матрица, т. е. состоящая из ортогональных друг другу векторов единичной длины, а **R** — верхняя треугольная матрица (нули под главной диагональю). Матрица **Q** называется ортогональной, если она удовлетворяет условию:

$$\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{Q} = \mathbf{I}$$

где $\mathbf{I}$ — единичная матрица, $\mathbf{Q}^{\mathrm{T}}$ — транспонированная матрица $\mathbf{Q}$. Свойство сохранения нормы произвольного вектора $\mathbf{x}$ при ортогональных преобразованиях $|\mathbf{Q} \cdot \mathbf{x}| = |\mathbf{x}|$ дает алгоритм поиска приближенных решений систем линейных алгебраических уравнений (СЛАУ) с прямоугольной матрицей $\mathbf{A}$: $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$. Суть алгоритма заключается в том, что задача минимизации с матрицей $|\mathbf{A} \cdot \mathbf{x} - \mathbf{b}|$ заменяется задачей $|\mathbf{Q}^{\mathrm{T}} \cdot (\mathbf{A} \cdot \mathbf{x} - \mathbf{b})|$, в которой $\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{A}$ уже будет "хорошей" благодаря специальному построению матрицы $\mathbf{Q}$, причем как переопределенных, так и недоопределенных, в том числе и плохо обусловленных. Для невырожденной СЛАУ можно сразу записать $\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{Q} \cdot \mathbf{R} \cdot \mathbf{x} = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{b}$, отсюда следует благодаря ортогональности матрицы $\mathbf{Q}$ алгоритм решения системы $\mathbf{R} \cdot \mathbf{x} = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{b}$. Так как матрица $\mathbf{R}$ треугольная, то решение получается по формулам прямого хода [10].

Для нашей задачи дискриминантного анализа масс-спектров выдыхаемых газов $\mathbf{A}$ — матрица исходных данных или значений интенсивностей пиков масс-спектра на определенных массах, или значений переменных, полученных после преобразования исходной матрицы $\mathbf{A}$ методом главных компонент; $\mathbf{b}$ — вектор свободных членов СЛАУ, элементы которого принимают два значения, например: -0.5 — здоров, 0.5 — болен; $\mathbf{x}$ — вектор коэффициентов дискриминантной функции. Для j-объекта (испытуемого) расчетное значение дискриминантной функции будет записываться в следующем виде:

$$F_j = a_{j1} \cdot x_1 + \ldots + a_{ji} x_i, \qquad (1)$$

a_{ji} — элементы матрицы исходных данных $\mathbf{A}$ наблюдений. В нашем случае a_{ji} — интенсивность i-пика в j-масс-спектре.

1.2. Алгоритм на основе вычисления обобщенных ковариационных матриц

Процедура нахождения коэффициентов дискриминантной функции с использованием ковариационных матриц была кратко описана в работе [1]. Опишем этот алгоритм более подробно.

Для расчета коэффициентов дискриминантных функций нужен статистический критерий, оценивающий различия между группами. Очевидно, что классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Один из методов поиска наилучшей дискриминации данных заключается в нахождении таких дискриминантных функций, которые были бы основаны на максимуме отношения межгрупповой вариации к внутригрупповой. Многомерное нормальное распределение случайной величины a_{ji} характеризуется следующими статистическими компонентами:

- $\mathbf{A}_j$ — вектор из m средних значений переменной j по всем классам (общие средние значения используемых признаков);
- $\mathbf{T}$ — матрица размера $m \times m$ сумм квадратов и попарных произведений, которая показывает степень различий между признаками. Элементы матрицы $\mathbf{T}$ задаются соотношением

$$t_{jl} = \sum_{k=1}^{p} \sum_{i=1}^{n} (a_{ijk} - A_j)(a_{ilk} - A_l), \qquad (2)$$

где выражения в скобках — отклонения значений переменных от общего среднего, j и l — номера переменных (номера масс), $j, l = 1, \ldots, m$.

Будем считать, что в обучающей выборке количество здоровых и больных будет одинаковым и равно n.

- $\mathbf{A}_{jk}$ — матрица $m \times p$ из средних значений переменной j для измерений k-го класса (групповые средние);
- $\mathbf{W}$ — матрица, которая используется для определения степени разброса внутри классов и отличается от $\mathbf{T}$ тем, что при ее вычислении используются средние для отдельных классов, а не общие средние:

$$w_{jl} = \sum_{k=1}^{p} \sum_{i=1}^{n} (a_{ijk} - A_{jk})(a_{ilk} - A_{lk}). \qquad (3)$$

Если элементы матрицы $\mathbf{W}$ разделить на $(n - p)$, то получится внутригрупповая ковариационная матрица $\mathbf{S}$, которая рассматривалась в работе [1]. Если расположение центров классов различается между собой, то степень вариации наблюдений внутри классов будет меньше общего статистического разброса $w_{jl} < t_{jl}$, причем чем больше расхождение этих величин, тем ощутимее влияния фактора группировки. Введем матрицу разницы этих двух матриц $\mathbf{B}$, которая представляет собой межгрупповую сумму квадратов отклонений и попарных произведений $\mathbf{B} = \mathbf{T} - \mathbf{W}$ (т. е. $b_{jl} = t_{jl} - w_{jl}$). Величины элементов $\mathbf{B}$ по отношению к величинам элементов $\mathbf{W}$ дают меру различия между группами. Коэффициенты $x_{1k}, x_{2k}, \ldots, x_{mk}$ разделяющих функций могут быть найдены по методу дискриминантного анализа

Фишера [11] как элементы матрицы, обратной к $\mathbf{W}$, что соответствует общей вычислительной процедуре множественной линейной регрессии. Такой метод нахождения коэффициентов разделяющей функции был использован в работе [1]. Классификация проводится по двум группам, и в нашем случае в формулах (2) и (3) $p = 2$.

1.3. Алгоритм на основе решения переопределенной системы линейных алгебраических уравнений методом наименьших квадратов

Для определения компонентов вектора $\mathbf{x}$, которые являются коэффициентами дискриминантной функции, нам нужно решить переопределенную систему линейных алгебраических уравнений $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ с прямоугольной матрицей $n \times m$. Перепишем эту систему в полном виде, т. е. произведем матричное умножение. Получаем систему уравнений следующего вида:

$$\begin{bmatrix} a_{11}\, a_{12} \cdots a_{1m} \\ a_{21}\, a_{22} \cdots a_{2m} \\ \cdots \cdots \cdots \cdots \\ a_{n1}\, a_{n2} \cdots a_{nm} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix}. \qquad (4)$$

Приближенное решение такой системы можно выполнить методом наименьших квадратов [12, 13].

В выражении (4) a_{ij} — значение интенсивности i-пика в j-масс- спектре; b_j — значение признака: -0.5 — здоров, 0.5 — болен, в зарегистрированном j экспериментальном спектре.

Определим невязку $r_j = \sum_{i=1}^{m} a_{ij} x_i - b_j$.

Сумма квадратов невязок имеет вид:

$$\varphi = \sum_{j=1}^{n} r_j^2, \quad \text{или} \quad \varphi = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij} x_i - b_j \right)^2. \qquad (5)$$

Найдем обобщенное решение методом наименьших квадратов: приравняем все частные производные по компонентам дискриминантной функции нулю (условия минимума)

$$\frac{\partial \varphi}{x_i} = 2 \sum_{j=1}^{n} a_{ij} \left(\sum_{i=1}^{m} a_{ij} x_i - b_j \right) = 0, \qquad (6)$$

или

$$\frac{\partial}{\partial x_i} \left\{ \sum_{i=1}^{n} [a_{i1} x_1 + a_{i2} x_2 + \cdots + a_{im} x_m - b_i]^2 \right\} = 0.$$

Дифференцирование дает систему из m линейных уравнений с m неизвестными. Уравнение r системы имеет вид

$$\sum_{j=1}^{n} \left[a_{j1} a_{jr} x_1 + a_{j2} a_{rj} x_2 + \cdots + a_{jm} a_{jr} x_m - a_{jr} b_j \right] = 0. \qquad (7)$$

Выделяя вектор свободных членов и обозначая его $\mathbf{B}$, получаем, что r-й элемент этого вектора равен

$$B_r = \sum_{j=1}^{n} a_{jr} b_j. \qquad (8)$$

Каждый элемент матрицы $\mathbf{C}$ при неизвестных x_r вычисляется по формуле

$$C_{ri} = \sum_{j=1}^{n} a_{jr} a_{ji}, \qquad (9)$$

r и i — соответственно номер столбца и строки матрицы $\mathbf{C}$ размером $m \times m$.

Система из m линейных уравнений с элементами матрицы $\mathbf{C}$ и с m неизвестными решается методом Крамера.

В процессе выполнения настоящей работы все три рассмотренные алгоритма вычисления коэффициентов дискриминантных функций были реализованы в системе Matlab. Выбор одного из трех алгоритмов осуществлялся путем сравнения результатов их работы на одних и тех же данных. Проведенное сравнение показало, что исследуемые алгоритмы дают разные коэффициенты дискриминантной функции и соответственно имеют разную надежность классификации. Для дальнейшей работы выбирался такой алгоритм, который давал наибольшее количество правильных ответов по контрольной выборке.

2. ВЫЯВЛЕНИЕ В МАСС-СПЕКТРАХ ЛИНИЙ, ИНТЕНСИВНОСТИ КОТОРЫХ ОКАЗЫВАЮТ РЕШАЮЩЕЕ ЗНАЧЕНИЕ НА ПРИНЯТИЕ РЕШЕНИЯ О ПРИНАДЛЕЖНОСТИ ОБРАБАТЫВАЕМОГО МАСС-СПЕКТРА К ГРУППЕ БОЛЬНЫХ ИЛИ ГРУППЕ ЗДОРОВЫХ ЛЮДЕЙ

С геометрической точки зрения дискриминантные функции определяют гиперповерхности в p-мерном пространстве. Коэффициенты x_i дискриминантной функции, записанные в формуле (1) выбираются таким образом, чтобы центроиды (средние значения) различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы

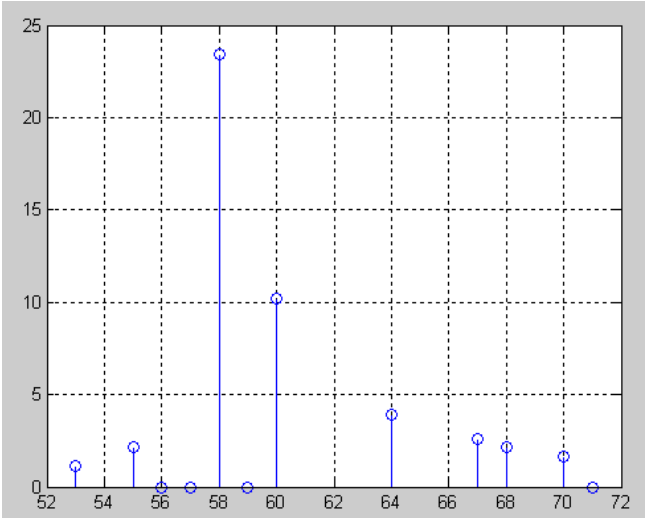

Рис. 1. Масс-спектр здорового человека в котором выделены 12 масс.
По горизонтальной оси значения массовых чисел в а.е.м. По вертикальной оси интенсивности на соответствующих массах. В дискриминантном анализе используются интенсивности масс, которые помечены знаком "o". Нулевая интенсивность означает, что на данной массе пик меньше заданного порога

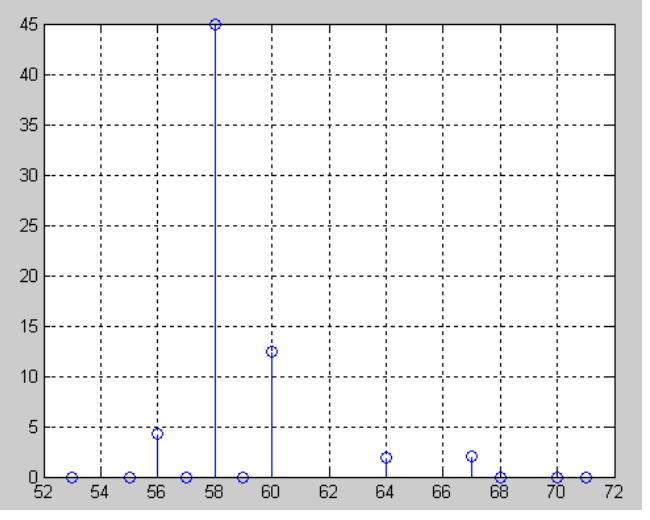

Рис. 2. Масс-спектр человека с возможными патологиями, в котором выделены 12 масс.
По горизонтальной оси значения массовых чисел в а.е.м. По вертикальной оси интенсивности на соответствующих массах

значения второй функции были некоррелированы со значениями первой.

Проводя вычисления дискриминантной функции по формуле (1) для данных, про которые заранее известно, что они принадлежат к группе здоровых или больных людей, мы проводим расстановку коэффициентов x_i таким образом, чтобы первым в получаемой последовательности был наибольший коэффициент, затем следующий по величине и т. д. Каждый из коэффициентов x_i соответствует определенной линии в масс-спектре, и таким образом можно выявить масс-спектрометрические пики, интенсивности которых наибольшим образом влияют на принятие решений о принадлежности масс-спектра к определенной группе.

Для проведения дискриминантного анализа использовались 12 линий, соответствующие массам: 53, 55, 56, 57, 58, 59, 60, 64, 67, 68, 70 и 71. Исходный масс-спектр, содержащий 64 линии [14], преобразуется в вектор, содержащий 12 переменных — интенсивностей пиков на отдельных массах, с помощью которых можно отличить масс-спектры здоровых и больных людей. Интенсивности пиков усредненного масс-спектра здорового человека и человека с возможными патологиями на выбранных массах показаны на рис. 1, 2 и в табл. 1. Эти масс-спектры были взяты в качестве обучающей выборки.

Табл. 1. Интенсивности пиков усредненных масс-спектров здоровых людей и людей с возможными патологиями на 12 выбранных массах. Эти интенсивности являются переменными в дискриминантном анализе. Ниже масс в скобках указаны номера переменных

Номер спектра	Массы в а.е.м. (номера переменных)												Группа: 0 (здоров) 1 (болен)
	53 (1)	55 (2)	56 (3)	57 (4)	58 (5)	59 (6)	60 (7)	64 (8)	67 (9)	68 (10)	70 (11)	71 (12)	
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.1	2.1	0	0	23.4	0	10.2	3.9	2.6	2.1	1.6	0	0

Табл. 1 (*продолжение*)

1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	0	0	0	0	25.9	0	10.3	0	1.3	0	1.1	0	0
3	6.3	2.7	0	0	22.4	0	11.6	4.6	0	0	2.3	0	0
4	5.5	0	3.3	0	22.0	0	11.9	3.4	5.6	0	1.1	1.1	0
5	0	0	5.3	0	24.0	0	10.6	6.2	7.6	0	1.1	1.2	0
6	0	0	4.3	0	45.0	0	12.4	1.9	2.0	0	0	0	1
7	0	0	0	0	43.0	0	8.3	0	0	2.6	0	0	1
8	0	5.9	0	0	81.0	0	8.5	4.1	0	0	0	0	1
9	9.5	0	0	0	12.0	0	10.1	0	1.8	0	0	1.6	1
10	4.7	3.1	1.5	0	18.8	0	10.4	0	5.5	0	0	0	1

Табл. 2. Значения коэффициентов дискриминантной функции, вычисленных по обучающей выборке

Характеристика	Номер переменной											
	1	2	3	4	5	6	7	8	9	10	11	12
Масса	53	55	56	57	58	59	60	64	67	68	70	71
Коэффициент дискриминантной функции	12.0	–8.0	–17.2	–19.8	1.4	0.6	27.9	34.9	–2.7	0.2	17.5	–25.2

Табл. 3. Собственные числа ковариационной матрицы исходных данных

Функция	Номер главной компоненты											
	1	2	3	4	5	6	7	8	9	10	11	12
Собственные числа ковариационной матрицы исходных данных	1826	169	21	10	3	2.4	1.5	1.1	0.8	0.5	0.25	0.16

В табл. 2 приведены значения коэффициентов дискриминантной функции, вычисленных по обучающей выборке. Выполнив упорядочивание коэффициентов дискриминатной функции и сопоставляя эти коэффициенты с массовыми числами, можно выделить те массы, интенсивности пиков на которых оказывают наиболее существенное влияние на результаты дискриминантного анализа. Эти выбранные массы представлены в табл. 2.

3. ПРЕОБРАЗОВАНИЕ ОБРАБАТЫВАЕМЫХ МАСС-СПЕКТРОВ МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ

Метод главных компонент (МГК) позволяет существенно сократить размерность обрабатываемых данных. Количество главных компонент, с помощью которых представляются наблюдаемые данные, можно найти, анализируя вектор собственных значений ковариационной матрицы **Cm**

исходной матрицы данных **A** . Начиная с определенного номера элемента этого вектора, его элементы приближаются к нулю. Если выбрать из этих элементов только превышающие некоторый заданный порог, то таким образом можно определить минимальное количество компонент, достаточное для представления исходных данных по координатам главных компонент. В работе [1] удовлетворительные результаты были получены при использовании шести главных компонент. В данной работе было проведено исследование возможности дискриминантного анализа при использовании только двух главных компонент. В табл. 3 представлены собственные числа ковариационной матрицы **A** . В этой таблице 12 столбцов по числу переменных, т. е. масс в исходных данных.

На рис. 3 показан результат проведения дискриминантного анализа, в котором были использованы только две главные компоненты: самая главная PCA-1 и следующая за ней PCA-2. Для получения результатов, приведенных на рис. 3, был проанализирован массив данных, состоящий из 400 масс-спектров здоровых и 400 масс-спектров больных людей. Обработка данных такого массива показала, что количество масс-спектров больных людей, отнесенных к классу здоровых равно нулю. Количество масс-спектров здоровых людей, отнесенных к классу больных, не превышает 20 из 400. Данный результат показывает принципиальную возможность использования двух главных компонент для проведения дискриминантного анализа масс-спектров выдыхаемых газов.

4. ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ ПРИНЯТИЯ ПРАВИЛЬНОГО И ОШИБОЧНОГО РЕШЕНИЙ О ПРИНАДЛЕЖНОСТИ ОБРАБАТЫВАЕМОГО МАСС-СПЕКТРА К ГРУППЕ ЗДОРОВЫХ ИЛИ БОЛЬНЫХ ЛЮДЕЙ

Используя две главные компоненты, о которых упоминалось выше, было проведено обучение по выборке, состоящей из 400 масс-спектров здоровых

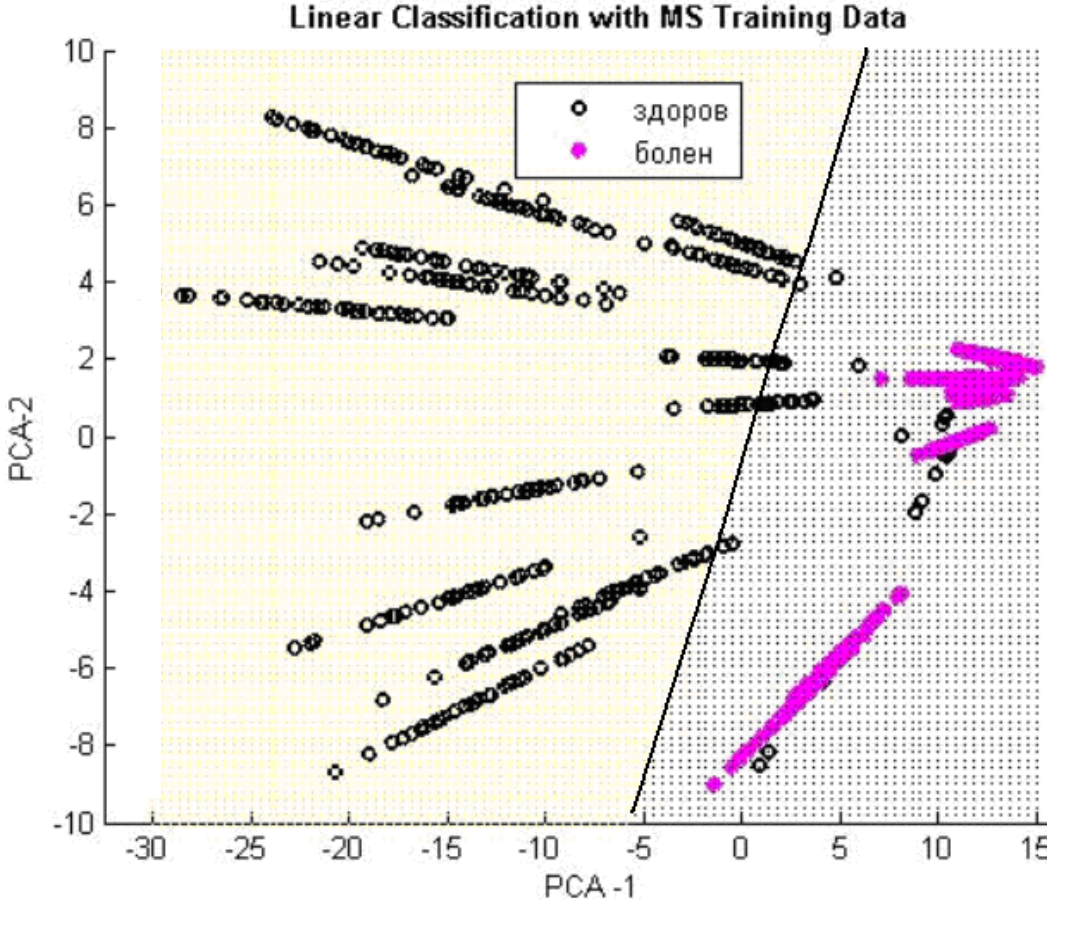

Рис. 3. Проведение дискриминантного анализа для двух переменных, представляющих собой главные компоненты

Табл. 4. Вероятности принятия правильного P_r и ошибочного P_f решений при классификации контрольной выборки здоровых людей

Вероятности	Номер масс-спектра														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P_r	1.0	0.82	0.82	0.80	0.75	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	1.0	0.94
P_f	0.0	0.18	0.18	0.20	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.06

Табл. 5. Вероятности принятия правильного P_r и ошибочного P_f решений при классификации контрольной выборки людей с возможными патологиями

Вероятности	Номер масс-спектра														
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
P_r	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.68	0.54	0.97	0.94	0.96
P_f	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.32	0.46	0.03	0.06	0.04

и 400 масс-спектров больных людей. В результате обучения была найдена дискриминантная функция, которая в случае двух переменных представляет собой линейную функцию. Графически такая функция представляется, например, как наклонная линия на рис. 3. Вероятности принятия правильного и ошибочного решений были вычислены по контрольной выборке реальных масс-спектров, состоящей из 15 масс-спектров здоровых людей (группа 1 — здоровые) и 15 масс-спектров людей с возможными патологиями (группа 2 — больные). Данные масс-спектры в контрольной выборке из 30 масс-спектров были представлены таким образом, что масс-спектры с порядковыми номерами от 1 до 15 принадлежали данным здоровых людей, а масс-спектры с порядковыми номерами от 16 до 30 — людей с возможными патологиями.

В табл. 4 представлены вероятности правильной классификации масс-спектров здоровых людей, а в табл. 5 представлены вероятности правильной классификации масс-спектров людей с возможными патологиями. В этих таблицах P_r — вероятность правильного принятия решения, P_f — вероятность ошибочного решения. Полученные результаты также представлены на рис. 4.

Данные, представленные в табл. 4 и 5, а также на рис. 4 показывают, что для контрольной выборки реальных масс-спектров классификация с помощью дискриминантного анализа прошла успешно: все масс-спектры здоровых и больных людей классифицированы правильно.

ЗАКЛЮЧЕНИЕ

Рассмотренные алгоритмы классификации на основе дискриминантного анализа дают возможность автоматического принятия решений о различии масс-спектров без визуального анализа информации, представленной в графическом виде. Такое автоматическое принятие решений может быть полезным при массовых экспресс-анализах. При возникновении незначительных отклонений вычисленных значений дискриминантных функций от порогов целесообразным является повторение измерений с дополнительным анализом информации по графикам масс-спектров и измерениями на других, отличных от масс-спектрометра приборах. Кроме того, данный метод целесообразно использовать параллельно с другими методами анализа патологий.

Рассмотренные алгоритмы имеют следующие преимущества: простота реализации, возможность автоматического принятия решения о принадлежности проверяемого сигнала к определенному классу.

Указанные преимущества дают возможность рекомендовать описанный подход использования

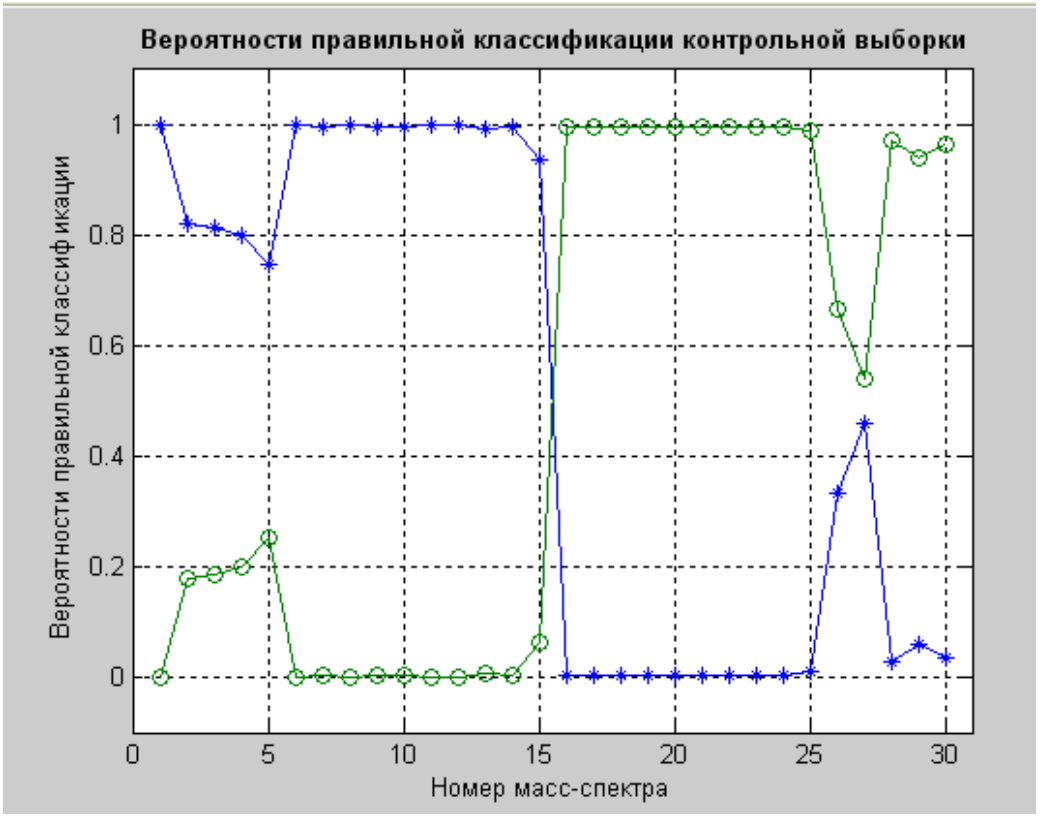

Рис. 4. Вероятности правильной классификации контрольной выборки масс-спектров.
***** — вероятности отнесения к группе 1 (здоровые), **o** — вероятности отнесения к группе 2 (больные)

дискриминантного анализа для обработки информации не только масс-спектрометров, но и других типов аналитических приборов.

СПИСОК ЛИТЕРАТУРЫ

1. *Манойлов В.В., Титов Ю.А., Кузьмин А.Г., Заруцкий И.В.* Методы обработки и классификации масс-спектров выдыхаемых газов с использованием дискримнантного анализа // Научное приборостроение. 2016. Т. 26, № 3. С. 50–57. URL: http://213.170.69.26/mag/2016/abst3.php#abst7.
2. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. Справочное издание / Под ред. проф. С.А. Айвазяна. М.: "Финансы и статистика", 1989. 608 с.
3. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы. Учебник. М.: "Финансы и статистика", 1998. 352 с.
4. Факторный, дискриминантный и кластерный анализ. Пер. с англ. / Под ред. И.С. Енюкова. М.: "Финансы и статистика", 1989. 215 с.
5. Statsoft (электронный учебник по статистике). URL: http://www.statsoft.ru/home/textbook/modules/stdiscan.html.
6. *Кузьмин А.Г.* Квадрупольный масс-спектрометр. Патент на полезн. мод. РФ № 94763, 27.05.2010.
7. *Кузьмин А.Г., Титов Ю.А.* Малогабаритные масс-спектрометры для динамических исследований состава выдыхаемого воздуха // Труды I Международной научно-практической конференции "Высокие технологии, фундаментальные и прикладные исследования в физиологии и медицине", СПб., 23−26 ноября 2010 г. Изд-во СПбГПУ, 2010. Ч. 3. С. 266−270.
8. *Кузьмин А.Г., Ткаченко Е.И., Орешко Л.С., Титов Ю.А.* Перспективы метода масс-спектрометрической аромадиагностики по составу выдыхаемого воздуха // Тезисы докладов X Евразийской научной конференции "ДОНОЗОЛОГИЯ−2014", 18−19 декабря 2014 г. СПб., 2014. С. 229−231.
9. *Кузьмин А.Г., Ткаченко Е.И., Орешко Л.С., Титов Ю.А.* Диагностические возможности масс-спектрометрии выдыхаемого воздуха // Сборник тезисов I Всероссийской конференции с международным участием "Химический анализ и медицина", 09–12 ноября, 2015, Москва. С. 35.
10. *Кирьянов Д.В., Кирьянова Е.Н.* QR- и SVD-разложения: "плохие" СЛАУ. URL: http://www.polybook.ru/comma/2.10.pdf.

11. *Шитиков В.К., Розенберг Г.С., Зинченко Т.Д.* Дискриминантные функции для классификации многомерных объектов. URL: http://www.ievbras.ru/ecostat/Kiril/Library/Book1/Content383/Content383.htm.
12. *Бернар М.* Современная масс-спектрометрия. М.: Изд-во "Иностранная литература", 1963.
13. *Манойлов В.В., Заруцкий И.В.* Оценка амплитуд "наложившихся" масс-спектрометрических пиков при известных положениях на оси масс и известных полуширинах алгебраическим методом // Научное приборостроение. 2007. Т. 17, № 1. С. 98–102. URL: http://213.170.69.26/mag/2007/abst1.php#abst13.
14. *Manoilov V.V., Kuzmin A.G., Titov U.A.* Extraction of information attributes from the mass spectrometric signals air // Journal of Analytical Chemistry. 2016. Vol. 71, no. 14. P. 1301–1308.

Институт аналитического приборостроения РАН, г. Санкт-Петербург (*Манойлов В.В., Титов Ю.А., Кузьмин А.Г., Заруцкий И.В.*)

Университет ИТМО, г. Санкт-Петербург (*Манойлов В.В.*)

Контакты: *Манойлов Владимир Владимирович*, manoilov_vv@mail.ru

Материал поступил в редакцию 21.07.2017

DISCRIMINANT ANALYSIS ALGORITHMS FOR CLASSIFICATION MASS SPECTRA OF EXHALED GASES

V. V. Manoylov[1,2], Yu. A. Titov[1], A. G. Kuzmin[1], I. V. Zarutskiy[1]

[1]*Institute for Analytical Instrumentation of RAS, Saint-Petersburg, Russia*
[2]*ITMO University, Saint-Petersburg, Russia*

In this paper the algorithms of discriminant analysis for the classification of mass spectra of exhaled gases are considered. It is shown that three algorithms can be used to calculate the coefficients of discriminant functions: 1) an algorithm based on the QR decomposition; 2) an algorithm based on the calculation of the generalized correlation function, and 3) an algorithm based on the solution of an overdetermined system of linear equations by the method of least squares. Tables are given for calculating the probabilities of deciding whether a processed mass spectrum belongs to one of two groups: mass spectra belonging to healthy and unhealthy people. It is shown in the paper that the amplitudes of the mass-spectral peaks of the masses of the exhaled gas as well as the variables resulting from the reduction in the dimension of the data processed by the principal component method can be used as variables for carrying out the classification. It is shown that for an estimation of belonging to one of these groups, it is sufficient to use variables corresponding to the first two principal components. Examples of approbation of the proposed methods are given.

Keywords: the mass spectrometer for the analysis of exhaled gases, linear discriminant analysis, classification of mass spectra

REFERENCES

1. Manoylov V.V., Titov Yu.A., Kuz'min A.G., Zaruzkiy I.V. [Methods for data processing and classification for mass spectra of exhaled gases using discriminant analysis]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2016, vol. 26, no. 3, pp. 50–57. (In Russ.). Doi: 10.18358/np-26-3-i5056.
2. Ajvazyan S.A., Buhshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika. Klassifikaciya i snizhenie razmernosti. Spravochnoe izdanie* [Applied statistics. Classification and decrease in dimension. Reference book]. *Ed. S.A. Ajvazyan.* Moscow, Finansy i statistika Publ., 1989. 608 p. (In Russ.).
3. Dubrov A.M., Mhitaryan B.C., Troshin L.I. *Mnogomernye statisticheskie metody. Uchebnik* [Many-dimensional statistical methods. Textbook]. Moscow, Finansy i statistika Publ., 1998. 352 p. (In Russ.).
4. Enyukov I.S., ed. *Faktornyj, diskriminantnyj i klasternyj analiz* [Factor, discriminant and cluster analysis]. Moscow, Finansy i statistika Publ., 1989. 215 p. (In Russ.).
5. Statsoft (electronic textbook statistically). URL: http://www.statsoft.ru/home/textbook/modules/stdiscan.html.

6. Kuzmin A.G. Kvadrupol'nyj mass-spektrometr. Patent RF no. 94763. [Patent for the quadrupole mass spectrometer]. Prioritet 27.05.2010. (In Russ.).
7. Kuzmin A.G., Titov U.A. [Small-size mass spectrometers for dynamic researches of composition of the exhaled air]. *Trudy I Mezhdunarodnoj nauchno-prakticheskoj konferencii "Vysokie tekhnologii, fundamental'nye i prikladnye issledovaniya v fiziologii i medicine". Ch. 3* [Proc. I of the Int. scientific and practical conference "High Technologies, Basic and Applied Researches in Physiology and Medicine", Part 3], Saint Petersburg, 23−26 November, 2010. SPbGPU Publ., 2010, pp. 266−270. (In Russ.).
8. Kuzmin A.G., Tkachenko E.I., Oreshko L.S., Titov U.A. [Prospects of a method of a mass and spectrometer aromadiagnostika for composition of the exhaled air]. *Tezisy dokladov X Evrazijskoj nauchnoj konferencii "DONOZOLOGIYA−2014"* [Theses of reports of the X Euroasian scientific DONOZOLOGIYA−2014 conference]. Saint-Petersburg, 18−19 December, 2014, pp. 229−231. (In Russ.).
9. Kuzmin A.G., Tkachenko E.I., Oreshko L.S., Titov U.A. [Diagnostic opportunities of a mass spectrometry of the exhaled air]. *Sbornik tezisov I Vserossijskoj konferencii s mezhdunarodnym uchastiem "Himicheskij analiz i medicina"* [The collection of theses of the I All-Russian conference with the international participation "A chemical analysis and medicine"], Moscow, 09–12.11.2015. P. 35. (In Russ.).
10. Kir'yanov D.V., Kir'yanova E.N. *QR- i SVD-razlozheniya: "plohie" SLAU* [QR-and SVD decomposition: "poor" SLOUGH]. (In Russ.). URL: http://www.polybook.ru/comma/2.10.pdf.
11. Shitikov V.K., Rozenberg G.S., Zinchenko T.D. *Diskriminantnye funkcii dlya klassifikacii mnogomernyh objektov* [Discriminant functions for classification of many-dimensional objects]. (In Russ.). URL: http://www.ievbras.ru/ecostat/Kiril/Library/Book1/Content383/Content383.htm.
12. Bernar M. *Sovremennaya mass-spektrometriya* [The modern mass spectrometry]. Moscow, IIL Publ., 1963. (In Russ.).
13. Manoylov V.V., Zaruzkiy I.V. [Algebraic estimation of amplitudes of "superimposed" mass spectrum peaks with known half-widths and positions on the mass axis]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2007, vol. 17, no. 1, pp. 98–102. (In Russ.). URL: http://213.170.69.26/en/mag/2007/abst1.php#abst12.
14. Manoilov V.V., Kuzmin A.G., Titov U.A. Extraction of information attributes from the mass spectrometric signals air. *Journal of Analytical Chemistry*, 2016, vol. 71, no. 14, pp. 1301–1308. Doi: 10.1134/S1061934816140094.

Contacts: *Manoylov Vladimir Vladimirovich,*
manoilov_vv@mail.ru

Article received in edition: 21.07.2017