

Математические методы исследования

Mathematical methods of research

DOI: 10.26896/1028-6861-2018-84-1-I-71-84

УДК (UDC) 519.234+519.237.8:620.165.29

НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ ПОСТРОЕНИЯ КЛАССИФИКАЦИОННОЙ СТРУКТУРЫ ХАРАКТЕРИСТИК ИСПЫТАНИЙ НА ГЕРМЕТИЧНОСТЬ

© Сергей Артурович Бушин

Всероссийский научно-исследовательский институт автоматики им. Н. Л. Духова, Москва, Россия;
e-mail: vnii4@vniia.ru, 2009-01-07@mail.ru*Статья поступила 13 сентября 2017 г.*

Рассмотрены проблемы правильного и эффективного выбора методов и средств при проведении контроля герметичности проникающими веществами (газами) на базе определенных критериев. На основе методов математической статистики в рамках проведенных автором теоретических исследований осуществлен классификационный анализ основных характеристик методов и способов испытаний на герметичность, целью которого являлось обозначение приоритетов последних на основе действующего ГОСТ Р 51780–2001. Другими словами, предпринята попытка поиска и выявления эффективных путей упорядочивания классификационной структуры из основных характеристик методов и средств испытаний на герметичность, необходимой для исследователя, при их выборе в процессе подготовки к данным испытаниям. Отчасти это связано с отсутствием в указанном стандарте четких рекомендаций относительно степени значимости данных характеристик для решения задачи контроля герметичности применительно к реальным условиям и возможностям ее реализации. Методика исследований базируется на предложенном непараметрическом алгоритме, который включает проведение классификации данных характеристик, основанной на парной кластеризации. Для эффективного решения задачи непараметрического анализа в работе использован метод экспертных оценок. Предложены решения в области многомерного анализа, относящиеся к задаче обратного сведения множества характеристик к небольшому ряду из групп, формируемых по обобщающим признакам, с использованием дискриминантного анализа на основе непараметрических критериев проверки статистически значимых различий. В результате проведенных исследований характеристик, относящихся главным образом к средствам испытаний на герметичность (как наиболее многочисленным в отличие от методов), определена сравнительно высокая апостериорная вероятность отнесения данных классификационных признаков к группам, распределенным по шести кластерам. Оценена возможность внесения в действующий вышеупомянутый стандарт некоторых уточнений в целях его совершенствования.

Ключевые слова: классификационные признаки; кластеризация; непараметрический анализ; ранговый коэффициент корреляции; дендрограмма; Евклидова метрика; дискриминантный анализ.

NON-PARAMETRIC ALGORITHM FOR CONSTRUCTING CLASSIFICATION STRUCTURE OF THE LEAKAGE TEST CHARACTERISTICS

© Sergey A. Bushin

N. L. Dukhov All-Russia Research Institute of Automatics, Moscow, Russia, e-mail: vnii4@vniia.ru, 2009-01-07@mail.ru

Submitted September 13, 2017.

The problems of correct and effective choice of the methods and means used for leakage monitoring by penetrating substances (gases) on the basis of certain criteria are considered. Methods of mathematical statistics are used for classification analysis of the main characteristics of the leakage test methods and approaches in the framework of general theoretical research to prioritize them in compliance with current GOST 51780–2001. In other words, an attempt has been made to search for and identify the effec-

tive ways of ordering the classification structure from the main characteristics of the methods and means of leak testing necessary for the researcher at the preparatory stage of testing. This is partly due to the lack of clear recommendations in this standard regarding the significance of these characteristics for leakage control in real conditions and possibility of its implementation. The procedure rests on the proposed nonparametric algorithm which includes classification of the characteristics based on pairwise clustering. The expert evaluation method is used for effective addressing the problem of nonparametric analysis. Solutions of multivariate analysis are given for the problem of inverse reduction of characteristics to a small number of groups formed on the basis of generalizing features using discriminative analysis on the basis of nonparametric criteria to check statistically significant difference. Comparatively high posterior probability of attributing the classification features to the groups distributed among six clusters is determined as a result of studying characteristics pertaining mainly to the leakage test facilities (as the most numerous ones compared to test methods). We also consider and assess the possibility of introducing some clarifications to refine the current standard.

Keywords: classification features; clusterization; nonparametric analysis; rank correlation coefficient; dendrogram; Euclidean metric; discriminatory analysis.

Вопросы выбора методов и средств контроля герметичности с использованием определенных критериев и способов при формировании оптимального алгоритма для решения прикладной задачи, например, для промышленного применения и/или лабораторно-исследовательских целей, всегда актуальны и нередко сопряжены со значительными временными и материальными затратами. Для координации действий целесообразно обращаться к стандартам и руководящим материалам, упорядочивающим перечни характеристик методов и средств, для решения подобных задач. Одним из таких руководящих документов является ГОСТ Р 51780–2001 (далее — стандарт), в котором подробно представлены методы и средства испытаний на герметичность, а также обозначены порядок и критерии их выбора [1]. Однако в стандарте выявлена некоторая неопределенность: отсутствуют четкие указания (рекомендации), связанные с порядком применения основных характеристик в соответствии со значимостью решаемых задач контроля герметичности в реальных условиях.

В работе на основе кластерного и дискриминантного анализа [2] определены эффективные пути создания упорядоченной по значимости

классификационной структуры из характеристик методов и средств испытаний на герметичность. Необходимость в такой структуре обусловлена, главным образом, неоднозначностью существующих подходов исследователей в решении этих вопросов из-за различных уровней их знаний и квалификации.

Согласно стандарту первым номером в процедуре выбора значится составление перечня обязательных и дополнительных требований, сформированного с учетом приоритетов, номером два — наличие характеристик методов и средств измерения.

Схема выбора критериев, согласно стандарту, построена следующим образом. Методы и средства должны выбираться исходя из их возможности выполнять обязательные и дополнительные требования. Если нет жесткой пороговой привязки нормы на выполнение требований к показателям испытаний на герметичность, критерием выбора становится соотношение цена — качество (стоимость — эффективность). В случае их равенства разработчики предлагают использовать такой критерий, как повышение культуры производства. Когда по какой-либо причине провести однозначный объективный выбор не удастся, необходимо совместно с заказчиком испытаний принять согласованное решение.

В стандарте содержатся восемь характеристик методов (табл. 1) и двадцать четыре характеристики, относящиеся к средствам испытаний на герметичность (табл. 2), являющиеся, по существу, классификационными признаками.

Какой из характеристик отдать предпочтение — порогу чувствительности, диапазону регистрации потоков вещества, стоимости средства испытаний или вероятности обнаружения течи и т. д., — на первый взгляд, понять сложно. Поэтому цель данного исследования — детальный анализ приоритетов характеристик, позволяющий более четко разобраться в иерархической упорядоченности (последовательности) ука-

Таблица 1. Характеристики методов испытаний на герметичность

1	Возможность испытания на суммарную и (или) локальную герметичность
2	Порог чувствительности
3	Диапазон регистрации потоков вещества
4	Применяемые вещества (пробное, балластное, индикаторное, вещество-носитель)
5	Избирательность к применяемым веществам
6	Вредное воздействие на людей, объект испытаний, окружающую среду
7	Границы применимости метода в конкретных условиях испытаний
8	Точность локализации течей

Таблица 2. Характеристики средств испытаний на герметичность

1	Соответствие выбранному методу
2	Порог чувствительности
3	Диапазон регистрации потоков вещества
4	Значение испытательного давления
5	Производительность
6	Стоимость средства испытаний
7	Квалификация и численность обслуживающего персонала
8	Затраты на эксплуатацию
9	Вероятность обнаружения течи
10	Время подготовки к работе
11	Уровень вредных и опасных воздействий на людей, объект испытаний, окружающую среду
12	Вероятность безотказной работы
13	Вид выдаваемой информации
14	Возможность фиксации результатов испытаний
15	Качество обработки полученной информации (простое измерение, прогноз и т. п.)
16	Возможность объединения средств испытаний в информационную систему
17	Границы применимости средства испытаний в конкретных условиях испытаний
18	Место и условия применения средства испытаний (лаборатория, заводские условия, ремонтные работы и т. п.)
19	Вид средства испытаний (стационарное, мобильное, переносное и т. п.)
20	Возможность модернизации
21	Вид конструктивного исполнения, габаритные размеры, форма
22	Условия эксплуатации и хранения
23	Устойчивость к внешним воздействиям
24	Требования к источникам энергии для функционирования

Таблица 3. Результаты экспертных оценок рангов по характеристикам методов испытаний ($n = 8$)

Эксперты ($m = 5$)	1	2	3	4	5	6	7	8
1(АИ)	7	8	6	5	4	3	1	2
2(ЛИА)	8	2	7	1	6	5	4	3
3(СА)	7	8	5,5	2	5,5	4	1	3
4(СС)	3	8	2	6	4	7	1	5
5(СС)	4	8	3	2	5	7	1	6
Сумма рангов, полученных каждым признаком S	29	34	23,5	16	24,5	26	8	19
Отклонение от средней суммы рангов SS	6,5	11,5	1,0	-6,5	2,0	3,5	-14,5	-3,5
Квадрат отклонения	42,25	132,25	1	42,25	4	12,25	210,25	12,25

занных признаков с присвоением им так называемой степени значимости.

В качестве усовершенствования стандарта предложен непараметрический алгоритм [3], который включает проведение классификации [4] данных характеристик — морфологических вариативных признаков (статистическая терминология), основанной на парной кластеризации. Для непараметрического анализа использован метод экспертных оценок [5]; с этой целью проведен соответствующий опрос мнений у пяти независимых экспертов (исследователей). В отсутствие беспристрастности, интуиции и независимости суждений получить компетентное мнение

практически невозможно, так же как и невозможно подобрать абсолютно одинаковых экспертов. Поэтому при анализе результатов экспертизы для оценки их общности суждений оценивали ранговый коэффициент корреляции [6], по величине которого судили о том, согласованы ли мнения экспертов из-за случайных совпадений или это статистически значимое совпадение.

В табл. 3 и 4 приведены данные, которые характеризуют ранги (рейтинги), выставленные независимыми экспертами. Средние суммы рангов S_1 и S_2 соответственно составляют:

$$S_1 = \frac{29 + 34 + 23,5 + 16 + 24,5 + 26 + 8 + 19}{8} = 22,5;$$

$$S_2 = \frac{98 + 111,5 + 92,5 + 66 + 106 + 100 + 70 + 74}{24} + \\ + \frac{81 + 79,5 + 90,5 + 74,5 + 52,5 + 64,5 + 50 + 61}{24} + \\ + \frac{22 + 49,5 + 39,5 + 27 + 45 + 17 + 14 + 14}{24} = 1499,5.$$

Отклонения от средних сумм: $SS_1 = 456,5$; $SS_2 = 20561,74$.

Подставив значения SS_1 и SS_2 в формулу оценки степени согласованности W_i , получим:

$$W_1 = \frac{12SS_1}{m^2(n_1^3 - n_1)} = \frac{12 \cdot 456,5}{25(512 - 8)} = 0,4348;$$

$$W_2 = \frac{12SS_2}{m^2(n_2^3 - n_2)} = \frac{12 \cdot 20561,74}{25(13824 - 24)} = 0,7152.$$

Видно, что эксперты, выставившие оценки, согласованы друг с другом, при этом в первом случае степень согласованности средняя ($W = 0,3 - 0,7$), а во втором случае — сильная ($W >$

$> 0,7$) [7]. На рис. 1 представлены диаграммы размаха для независимых переменных методов и средств испытаний на герметичность. На этих диаграммах центральная точка или линия указывает на положение центральной области (т. е. на среднее или медиану), прямоугольник — на характер изменчивости вокруг центрального положения (квартили, стандартные ошибки или стандартные отклонения), отрезки вокруг прямоугольников — на диапазон значений переменной (медиана, 25 – 75 %, минимум — максимум).

Интересно было выяснить, какой из уровней качественных признаков, обозначенных в стандарте, имеющих отношение к характеристикам методов и средств, наиболее значим для эксперта (исследователя). С этой целью определяли, насколько среднегрупповой профиль коррелирует с эталонным профилем.

В табл. 5, 6 представлены средние оценки рангов, полученные для каждого из качеств среднегруппового профиля (по восьми и 24 характеристикам методов и средств испытаний соответственно), из значений экспертных оценок и вы-

Таблица 4. Результаты экспертных оценок рангов по характеристикам средств испытаний ($n = 24$)

Метод испытаний	Эксперты ($m = 5$)					Сумма рангов, полученных каждым признаком S	Отклонение от средней суммы рангов SS	Квадрат отклонения
	1(АИ)	2(ЛИА)	3(СА)	4(СС)	5(СС)			
1	21,5	22	22,5	11	21	98	35,5	1261,7
2	21,5	22	24	20	24	111,5	49,0	2403,0
3	21,5	22	20	9	20	92,5	30,0	901,3
4	12	22	16	12	4	66	3,5	12,4
5	21,5	22	21	19	22,5	106	43,5	1894,1
6	17	17,5	19	24	22,5	100	37,5	1407,8
7	13,5	8,5	15	17	16	70	7,5	56,6
8	13,5	8,5	12	22	18	74	11,5	132,7
9	21	17,5	22,5	15	5	81	18,5	343,0
10	21,5	15	18	10	15	79,5	17,0	289,7
11	17	17,5	14	23	19	90,5	28,0	785,2
12	17	17,5	17	14	9	74,5	12,0	144,5
13	11	13	7,5	8	13	52,5	-10,0	99,6
14	9	13	7,5	21	14	64,5	2,0	4,1
15	9	13	13	3	12	50	-12,5	155,7
16	15	5	6	18	17	61	-1,5	2,2
17	1	6	5	2	8	22	-40,5	1638,6
18	9	7	10,5	16	7	49,5	-13,0	168,5
19	7	10	10,5	1	11	39,5	-23,0	528,0
20	2	4	4	7	10	27	-35,5	1258,8
21	6	11	9	13	6	45	-17,5	305,5
22	5	3	3	5	1	17	-45,5	2068,4
23	4	2	2	4	2	14	-48,5	2350,2
24	3	1	1	6	3	14	-48,5	2350,2

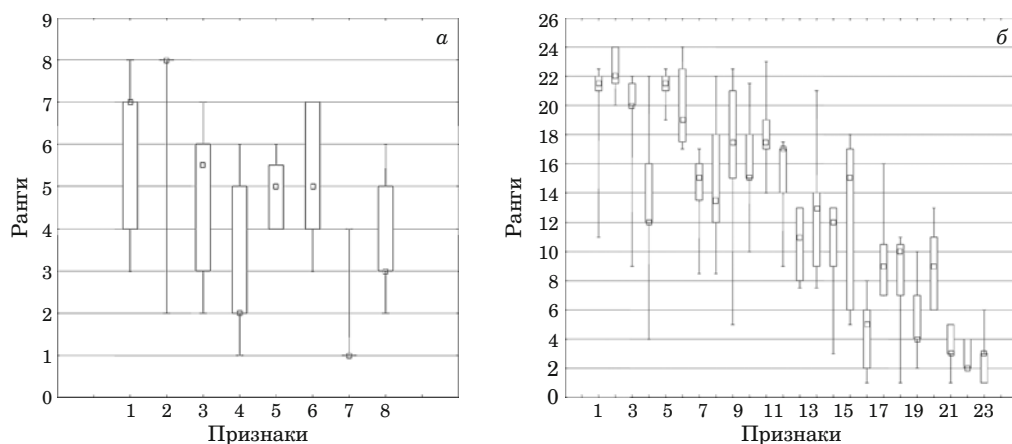


Рис. 1. Диаграммы размаха для независимых переменных методов (а) и средств испытаний (б) на герметичность: □ — медиана, □ — 25 – 75 %; ⊥ мин. – макс

борки рангов «идеального ряда», построенного на основе предложенного разработчиками стандарта: качества перечислены в последовательности, отражающей эталонный профиль. Кроме того, в табл. 5, 6 приведены расчетные значения разности рангов (d) и их квадратов (d^2) для определения рангового коэффициента r_s корреляции Спирмена между среднегрупповым и эталонным профилями.

Отметим, что ранжирование в данном случае проведено таким образом, что меньшему по значимости признаку присвоено низшее значение ранга.

Как принято в математической статистике, при проведении непараметрического анализа сформулированы две гипотезы: H_0 — корреляция между среднегрупповым профилем, построенным по оценкам экспертов, и эталонным профилем не отличается от нуля; H_1 — данная корреляция статистически значимо отличается от нуля.

Для расчета эмпирического коэффициента r_s , в отсутствие одинаковых рангов в группах использована формула

$$r_s = 1 - \frac{6}{N(N^2 - 1)} \sum d^2.$$

В рассматриваемом случае

$$r_{сэм} = 1 - \frac{6 \cdot 23,46}{8 \cdot (8^2 - 1)} = 1 - \frac{140,76}{504} = 0,7207 —$$

для показателей из табл. 5,

$$r_{сэм} = 1 - \frac{6 \cdot 207,82}{24(24^2 - 1)} = 1 - \frac{1246,92}{13\,800} = 0,9096 —$$

для показателей из табл. 6 (см. рис. 2, а, б соответственно).

По табл. XVI Приложения 1 из работы [8] определены критические значения r_s выборочного

Таблица 5. Средние групповые оценки рангов экспертов и «эталонного ряда» (стандарт) по восьми характеристикам методов испытаний на герметичность

Характеристики методов испытаний на герметичность ($N = 8$)	Усредненные групповые оценки рангов экспертов	«Эталонный ряд» (стандарт)	Разности рангов d	Сумма квадратов разностей рангов d^2
1. Возможность испытания на суммарную и (или) локальную герметичность	5,8	8	-2,2	4,84
2. Порог чувствительности	6,8	7	-0,2	0,04
3. Диапазон регистрации потоков вещества	4,7	6	-1,3	1,69
4. Применяемые вещества (пробное, балластное, индикаторное, вещество-носитель)	3,2	5	-1,8	3,24
5. Избирательность к применяемым веществам	4,9	4	0,9	0,81
6. Вредное воздействие на людей, объект испытаний, окружающую среду	5,2	3	2,2	4,84
7. Границы применимости метода в конкретных условиях испытаний	1,6	2	-0,4	0,16
8. Точность локализации течей	3,8	1	2,8	7,84
Сумма	36	36	0	23,46

коэффициента корреляции рангов (рис. 2). При $N = 8$ —

$$r_{\text{скр}} = \begin{cases} 0,72 (p \leq 0,05), \\ 0,88 (p \leq 0,01), \end{cases} \quad r_{\text{сэм}} > r_{\text{скр}} (p \leq 0,05);$$

при $N = 24$ —

$$r_{\text{скр}} = \begin{cases} 0,41 (p \leq 0,05), \\ 0,52 (p \leq 0,01), \end{cases} \quad r_{\text{сэм}} > r_{\text{скр}} (p \leq 0,01).$$

Для случая характеристик методов из табл. 1 гипотеза H_0 отвергается, а H_1 принимается — коэффициент корреляции практически совпал с границей зоны значимости на уровне $p \leq 0,05$. Для данных табл. 2 также гипотеза H_0 отвергается

и принимается гипотеза H_1 об отличии коэффициента корреляции от нуля со значимостью $p \leq 0,01$, которая является положительной. Интерпретация полученного результата такова. Из табл. 5 и 6 видно, что по сравнению со стандартом мнения экспертов имеют более низкий ранг по шкалам возможности испытания на суммарную и локальную герметичность и применяемых веществ из характеристик методов, а также соответствия выбранному методу и испытательного давления из числа качественных признаков, относящихся к средствам испытаний. При этом более высокие ранги мнения экспертов имеются по шкалам точности локализации и уровня вредных и опасных воздействий на людей и объект испытаний для характеристик методов и уровня

Таблица 6. Средние групповые оценки экспертов и «эталонного ряда» (стандарт) по 24 характеристикам средств испытаний на герметичность

Характеристики средств испытаний на герметичность ($N = 24$)	Усредненные групповые оценки рангов экспертов	«Эталонный ряд» (стандарт)	Разности рангов d	Сумма квадратов разностей рангов d^2
1. Соответствие выбранному методу	19,6	24	−4,4	19,36
2. Порог чувствительности	22,3	23	−0,7	0,49
3. Диапазон регистрации потоков вещества	18,5	22	−3,5	12,25
4. Значение испытательного давления	13,2	21	−7,8	60,84
5. Производительность	21,2	20	1,2	1,44
6. Стоимость средства испытаний	20	19	1	1
7. Квалификация и численность обслуживающего персонала	14	18	−4	16
8. Затраты на эксплуатацию	14,8	17	−2,2	4,84
9. Вероятность необнаружения течи	16,3	16	0,3	0,09
10. Время подготовки к работе	15,9	15	0,9	0,81
11. Уровень вредных и опасных воздействий на людей, объект испытаний, окружающую среду	18,1	14	4,1	16,81
12. Вероятность безотказной работы	14,9	13	1,9	3,61
13. Вид выдаваемой информации	10,5	12	−1,5	2,25
14. Возможность фиксации результатов испытаний	12,9	11	1,9	3,61
15. Качество обработки полученной информации (простое измерение, прогноз и т. п.)	10	10	0	0
16. Возможность объединения средств испытаний в информационную систему	12,2	9	3,2	10,24
17. Границы применимости средства испытаний в конкретных условиях испытаний	4,4	8	−3,6	12,96
18. Место и условия применения средства испытаний (лаборатория, заводские условия, ремонтные работы и т. п.)	9,9	7	2,9	8,41
19. Вид средства испытаний (стационарное, мобильное, переносное и т. п.)	7,9	6	1,9	3,61
20. Возможность модернизации	5,4	5	0,4	0,16
21. Вид конструктивного исполнения, габаритные размеры, формы	9	4	5	25
22. Условия эксплуатации и хранения	3,4	3	0,4	0,16
23. Устойчивость к внешним воздействиям	2,8	2	0,8	0,64
24. Требования к источникам энергии для функционирования	2,8	1	1,8	3,24
Сумма	300	300	0	207,82

вредных и опасных воздействий на людей и возможности объединения средств испытаний в информационную систему для признаков, относящихся к средствам испытаний. Этими расхождениями в основном и объясняется некоторое снижение полученных $r_{сэм}$ [8].

Однако возникает вопрос, можно ли переносить результаты тестовой выборки на генеральную совокупность, т. е. является ли она представительной (репрезентативной)? Любой исследователь в своей работе сначала устанавливает подгруппу внутри генеральной совокупности, подробно изучает ее, а затем, если это позволяют результаты статистического анализа, распространяет свои выводы на всю генеральную совокупность. В этой связи наиболее важным является вопрос о приемлемом объеме выборки.

Одни методы требуют большого числа испытуемых в выборке, другие могут применяться при относительно малом их количестве. Например, некоторые непараметрические критерии различий используются при сравнении групп численностью в 5 – 7 человек, а факторный анализ наиболее адекватен, если объем выборки составит около 100 человек [9]. Оптимальная численность группы участников находится эмпирическим путем: наиболее продуктивными признаны группы в 10 – 15 человек [10]. Однако если речь заходит о решении вопроса о приемлемости объема выборки, то, строго говоря, исследователю требуется, «увязать» его с вопросом о необходимом уровне точности результата (Δ_p), гарантируемого с некоторой заранее заданной доверительной вероятностью (P). Оценку требуемого объема выборки n для оценки исследуемого признака в генеральной совокупности можно получить расчетным путем, задавшись значениями допустимой ошибки и доверительной вероятности, на основе зависимости [11]

$$n = \frac{Npq t^2}{N\Delta_p^2 + pqt^2}, \quad (1)$$

где N — объем генеральной совокупности; p — доля исследуемого признака в генеральной совокупности; $q = 1 - p$; t — коэффициент соответствия доверительной вероятности P ; Δ_p — допустимая ошибка. Исходя, например, из нормированного нормального распределения при доверительной вероятности $P = 0,95$, получим, что коэффициенту t соответствует значение 1,96. Однако существует некоторая ограниченность: для получения результата по зависимости (1) необходима предварительная информация о доле признака испытуемых, а это как раз то, что исследователю и требуется определить. При этом

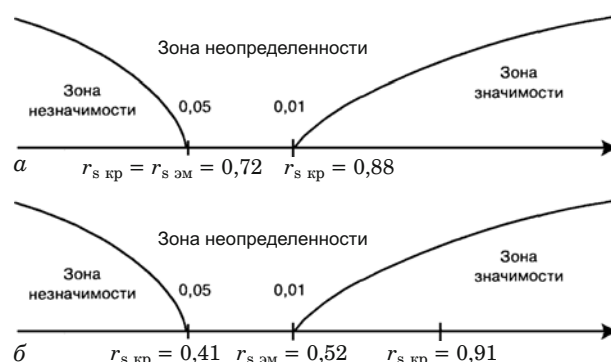


Рис. 2. Разложение эмпирического коэффициента корреляции рангов $r_{сэм}$ по зонам (участкам) на «оси значимости» для характеристик методов (а) и средств (б) испытаний на герметичность

для $p = 0,5$ произведение pq является максимальным, а поэтому n также будет иметь максимальное значение. Подставив в формулу (1) значение $p = 0,5$, получим выражение, которым можно пользоваться при любых долях признака в генеральной совокупности испытуемых, а объем выборки при этом будет получаться с некоторым запасом (при $t = 1,96$ и $P = 0,95$), а именно: при $N = 100$ значение $n = 79$; при $N = 20 - n = 19$, при $N = 5 - n \approx 5$ (4,94) [11].

Несмотря на сравнительно невысокий представительный объем выборки испытуемых полученные результаты имеют достаточно высокую сходимость мнений в установленной последовательности при решении задачи по выбору методов и средств контроля герметичности для испытателя.

Совершенно очевидно, что по мере возрастания количества характеристик качественных признаков все более теряется восприятие результатов. При этом закономерность начинает искусственно маскироваться из-за большого числа недостаточно значимых связей. Поэтому более рациональным для испытателя было бы исключить признаки, являющиеся малозначимыми. Такая задача обратного сведения множества характеристик к небольшому ряду из групп, формируемых по обобщающим признакам, присуща многомерному анализу. Достаточно выразительно отражает черты многомерного анализа в классификации объектов кластерный анализ [12].

Для решения данной задачи использование кластерного анализа наиболее эффективно. В общем случае он предназначен для объединения некоторых объектов в кластеры таким образом, чтобы в один кластер попадали максимально схожие, а объекты различных классов максимально отличались друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, харак-

теризующих объекты. Все кластерные алгоритмы нуждаются в оценках расстояний между кластерами или объектами.

Результаты исследований и их обсуждение

С помощью статистического программного комплекса Statistica 10.0 [13] проведен кластерный анализ для характеристик методов и средств (рис. 3).

Испробовав несколько вариантов числа группировок с использованием агломеративной стратегии, выбрали Евклидову метрику с методом Варда (Уорда), что минимизировало внутрикластерный разброс объектов и позволило получить дендрограмму (дерево объединения кластеров) с глубоко гипертрофированно разделенными кластерами [14].

При построении кластеризации важна проблема выбора числа кластеров. Предполагается, что кластеризация должна выявить естествен-

ные локальные сгущения объектов. Для определения количества классов K применили специальный инструмент из программного пакета Statistica для объединения признаков (объектов) в классы. Для этого использовали точку «излома» и номер шага m , на котором он произошел. В частности, при методе Варда (Уорда) на шкале «Расстояние объединения» признаков средств на рис. 4, б (22 единицы) с номером «Шага» $m = 18$, объединение произошло на существенно большем расстоянии, чем на шаге $(m - 1)$. Количество классов вычисляли по разнице значений n и m , где n — количество признаков в выборке ($n = 24$), откуда получено число классов, равное шести. Количество классов, оцениваемое цифрой 4, при анализе рис. 4, а можно охарактеризовать тем, что на шаге шкалы «Расстояние объединения», равном 4, скачок расстояния произошел более чем на три единицы, в то время как на предыдущих шагах скачки не превышали единицы. Однако четко выраженной направленности

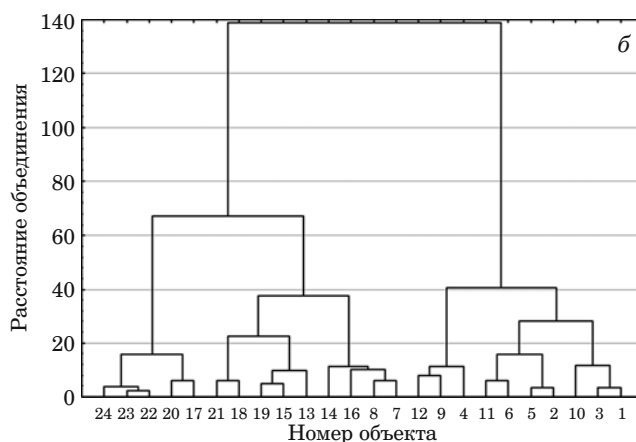
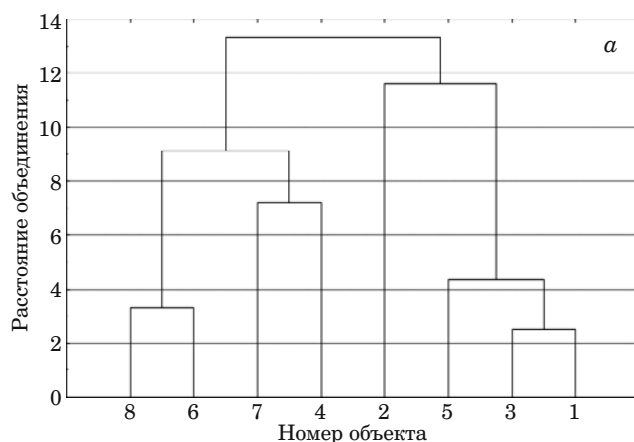


Рис. 3. Дендрограммы классификации характеристик методов испытаний на герметичность на основе Евклидовой метрики и метода Варда (Уорда) с использованием статистического комплекса Statistica 10.0: а и б — для восьми и 24 переменных

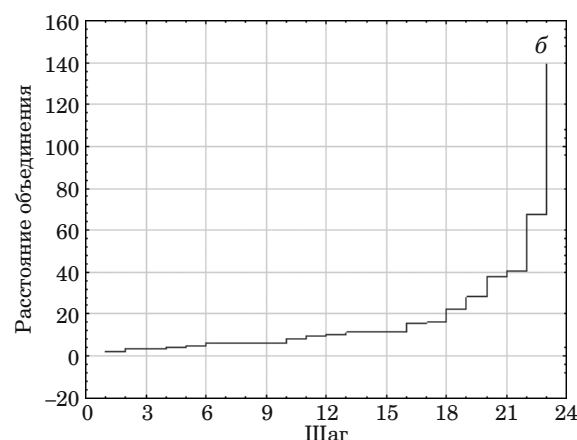
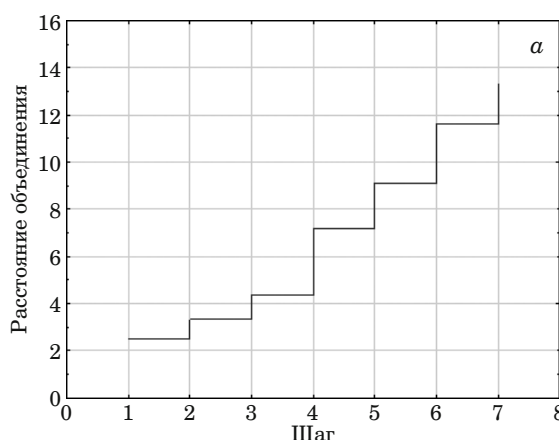


Рис. 4. График объединения признаков-объектов в классы с использованием Евклидовой метрики и метода Варда (Уорда): а и б — методы и средства испытаний на герметичность

на объединение нет, а поэтому выделение в классы характеристик методов может быть оспорено.

Анализ совокупностей из характеристик методов, на усмотрение испытателя, может быть оставлен без внимания с точки зрения необходимости формирования отдельных классов-кластеров. Однако аналогичные допущения не могут быть распространены на группу из характеристик средств испытаний ввиду их сравнительной многочисленности. Из рис. 3 и 4 можно заключить, что по схожести иерархий (признаков групп) классификационной структуры порядка характеристик средств испытаний последние целесообразно разделить на шесть классов (кластеров).

В табл. 7 представлен состав полученных классов, выделенных методом максимизирования расстояния между кластерами.

Из анализа графика (рис. 5), построенного с использованием метода *K*-средних, можно дать следующую интерпретацию полученным классам, начиная с первого.

Сформированный первый кластер (класс) является абсолютным лидером, в его состав вошли четыре признака из характеристик средств контроля: соответствие выбранному методу; порог чувствительности; диапазон регистрации потоков вещества; производительность контроля. По существу все они вошли в первую пятерку признаков, обозначенных в действующем стандарте.

Содержание второго кластера со средними показателями рангов составили следующие четыре признака: вид выдаваемой информации; качество обработки полученной информации (простое измерение, прогноз и т. п.); вид средства испытаний (стационарное, мобильное, перенос-

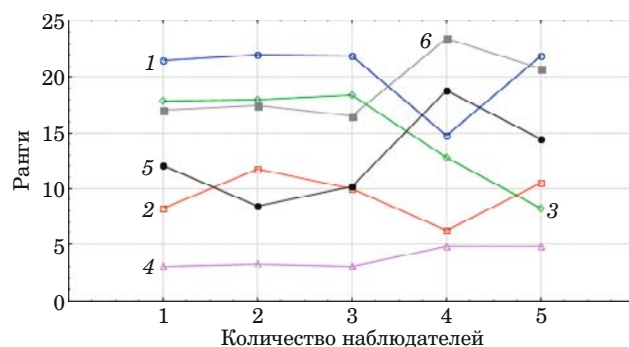


Рис. 5. График средних значений признаков в классах, полученных методом Варда (Уорда): 1 – 6 — кластеры

ное и т. п.); вид конструктивного исполнения, габаритные размеры и форма.

Третий класс, содержащий такие признаки, как значение испытательного давления, вероятность необнаружения течи, время подготовки к работе, вероятность безотказной работы, содержит достаточно высокие ранговые значения, имеющие тенденцию спада из-за мнения 4-го и 5-го экспертов.

Четвертый класс характеризуется самыми низкими уровнями среди всех классов по значимости мнений экспертов и состоит из пяти объектов-признаков: границы применимости средства испытаний в конкретных условиях испытаний; возможность модернизации; условия эксплуатации и хранения; устойчивость к внешним воздействиям; требования к источникам энергии для функционирования.

Характерной особенностью объектов, вошедших в пятый класс, является их удовлетвори-

Таблица 7. Результаты классификационного анализа характеристик средств испытаний на герметичность, сгруппированных в кластеры с применением программного комплекса STATISTICA 10.0

Номер класса	Количество объектов в классе	Состав класса (признаки)
1	4	1 — Соответствие выбранному методу; 2 — Порог чувствительности; 3 — Диапазон регистрации потоков вещества; 5 — Производительность.
2	4	13 — Вид выдаваемой информации; 15 — Качество обработки полученной информации (простое измерение, прогноз и т. п.); 19 — Вид средства испытаний (стационарное, мобильное, переносное и т. п.); 21 — Вид конструктивного исполнения, габаритные размеры, формы.
3	4	4 — Значение испытательного давления; 9 — Вероятность необнаружения течи; 10 — Время подготовки к работе; 12 — Вероятность безотказной работы.
4	5	17 — Границы применимости средства испытаний в конкретных условиях испытаний; 20 — Возможность модернизации; 22 — Условия эксплуатации и хранения; 23 — Устойчивость к внешним воздействиям; 24 — Требования к источникам энергии для функционирования.
5	5	7 — Квалификация и численность обслуживающего персонала; 8 — Затраты на эксплуатацию; 14 — Возможность фиксации результатов испытаний; 16 — Возможность объединения средств испытаний в информационную систему; 18 — Место и условия применения средства испытаний (лаборатория, заводские условия, ремонтные работы и т. п.).
6	2	6 — Стоимость средства испытаний; 11 — Уровень вредных и опасных воздействий на людей, объект испытаний, окружающую среду.

тельная востребованность. В их число вошли: квалификация и численность обслуживающего персонала; затраты на эксплуатацию; возможность фиксации результатов испытаний; возможность объединения средств испытаний в информационную систему; место и условия применения средства испытаний (лаборатория, заводские условия, ремонтные работы и т. п.).

Наиболее малочисленный, но имеющий сравнительно высокие показатели (выше среднего) — шестой кластер, включающий: средства испытаний; уровень вредных и опасных воздействий на людей, объект испытаний, окружающую среду.

Аналогичный кластерный анализ с использованием Евклидовой метрики был проведен с ис-

Таблица 8. Кластерный анализ; файл: Евклид + Дивизивная; таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(2)	9,605									
(3)	3,35	12,37								
(4)	20,65	24,94	19,37							
(5)	8,276	3,5	10,36	22,59						
(6)	15,78	10,43	17,26	23,44	9,721					
(7)	19,48	20,4	19,16	18,93	18,56	14,2				
(8)	22,42	21,12	22,34	22,37	19,3	12,94	6,164			
(9)	17,22	20,32	17,08	12,39	18,68	20,6	17,96	21,67		
(10)	10,5	16,43	9,11	16,38	14,12	16,99	13,01	17,3	12,32	
(11)	16,9	14,12	17,28	20,11	11,95	6,185	11,84	10,01	18,9	15,38
(12)	15,09	18,83	14,12	8,675	16,36	17,18	12,62	16,35	7,935	8,916
	14,11									
(13)	22,14	27,01	19,96	15,88	24,2	23,47	13,48	16,64	21,43	15,28
	19,38	14,12								
(14)	25,04	25,03	24,28	18,63	22,62	17,28	10,81	8,818	22,59	19,91
	12,51	15,89	13,53							
(15)	21,76	28,15	19,68	15,65	25,86	26,22	16,29	21,15	21,18	15,63
	23,46	15,21	7,762	19,16						
(16)	26,16	26,91	25,16	24,64	24,57	19,91	9,874	8,216	24,93	18,93
	15,95	19,11	14,4	10,97	20,2					
(17)	35,23	40,33	33,22	24,8	37,95	35,81	23,51	26,69	31,9	28
	32,18	26,04	14,84	22,74	14,07	23,15				
(18)	27,61	29,74	26,56	17,12	27,51	23,49	11,17	13,48	20,18	19,51
	19,51	15,12	12,57	10,91	15,66	12,82	17,12			
(19)	26,44	32,67	24,3	19,24	30,36	30,1	18,83	23,36	25,08	19,74
	27,06	19,3	9,327	20,98	4,924	21,05	9,811	16,22		
(20)	34,53	38,45	32,76	25,1	36,09	32,89	20,26	22,51	31,23	27,06
	29,12	25,04	13,72	18,85	15,33	18,57	5,916	13,83	12,01	
(21)	27,81	30,99	26,37	14,56	28,54	25,93	14,95	17,48	21,34	20,69
	21,75	15,44	10,26	12,34	12,85	16,76	13,96	6,801	13,16	12,08
(22)	38,19	42,86	36,17	25,26	40,31	37,91	24,83	27,61	31,04	29,38
	33,57	26,39	17,7	23,71	18,57	23,22	9,327	15,69	15,14	9,798
	13,93									
(23)	39,31	44,04	37,26	27	41,49	39,02	25,84	28,52	32,87	30,61
	34,75	28	18,47	24,7	19,29	23,83	8,602	17,1	15,47	9,22
	15,33	2,236								
(24)	40,04	44,25	38,1	28,02	41,73	38,74	25,56	27,74	33,83	31,44
	34,44	28,73	18,93	23,86	20,45	22,93	9,274	16,71	16,74	8,307
	15,33	4,123	2,828							

пользованием программного пакета STADIA 8.0 со стратегией Варда (Уорда) [14]. Данная стратегия позволяет группировать объекты (переменные) в заданное число кластеров. Полученные результаты имеют определенную общность с кластерами, выделенными с помощью программного комплекса Statistica 10. В этой связи для получения вышеуказанного числа кластеров можно применить дивизивную стратегию динамических сгущений. Результаты используемых стратегий Варда (Уорда) и дивизивной стратегии на основе Евклидовой метрики получили достаточно четкое группирование в кластеры, но при этом несколько отличную классификацию: 1) порядковые номера у кластеров не совпадают, но у кластеров 2, 3, 6 наблюдается полное совпадение с кластерами 1, 3, 4 соответственно; 2) в кластере 4 имеется всего три одинаковых признака из четырех определенных в сравниваемой структуре, а в кластере 5 количество признаков увеличилось ровно на два, ранее не принадлежащих кластеру пять; «пополнение» произошло как раз за счет тех признаков, которые образовывали шестой кластер в комплексе Statistica 10.0; 3) первый кластер пакета STADIA и шестой — из Statistica по-прежнему имеют всего два признака, но совершенно разных.

Обе классификации выделяют четко шесть кластеров. В табл. 8 представлены показатели расстояний.

Ниже представлены кластеры со значением среднего внутрикластерного расстояния, равного 5,638:

- 1 = (18*, 21)
- 2 = (1*, 2, 3, 5)
- 3 = (4, 9, 10, 12*)

- 4 = (13, 15*, 19)
- 5 = (6, 7, 8*, 11, 14, 16)
- 6 = (17, 20, 22, 23*, 24),

где значения со звездочкой — центры групп.

В табл. 9 представлены результаты объединения в классы (кластеры) на основе программного пакета STADIA 8.0 с уровнями значимости (в порядке ранговой упорядоченности).

На рис. 6 представлена одна из проекций кластеров на плоскость $X - Y$ (2D-формат). Как видно, разделение данных на шесть кластеров вполне состоятельно. Отсюда можно заключить, что по схожести иерархий (признаков групп) классификационной структуры порядка и критериев выбора методов и средств (в данном случае средств) характеристики средств целесообразно разделять на шесть классов. Причем порядок расположения кластеров в табл. 7 и 9 не может интерпретироваться как ранжирование (по возрастанию/по убыванию), наиболее важно содержимое построенных «групп», т. е. то, к какой совокупности классов принадлежит тот или иной качественный признак.

Следует отметить, что кластерный анализ не содержит вычислительного механизма проверки гипотезы об адекватности получаемых классификаций. Результаты можно обосновать с использованием метода дискриминантного анализа [14]. Метод дискриминантного анализа предполагает использование функции $f(x) = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n$ и нахождение на ее основе значений искомых параметров a_i, g_i, x_i и далее самой функции f . При использовании дискриминантного анализа необходимо учитывать его связь с алгоритмами кластерного анализа как наиболее близкими по получаемым результатам.

Таблица 9. Результаты классификационного анализа для характеристик средств испытаний на герметичность, сгруппированных в кластеры с применением программного пакета STADIA 8.0

Номер класса	Количество объектов в классе	Состав класса (признаки)
1	2	18 — Место и условия применения средства испытаний (лаборатория, заводские условия, ремонтные работы и т. п.); 21 — Вид конструктивного исполнения, габаритные размеры, формы.
2	4	1 — Соответствие выбранному методу; 2 — Порог чувствительности; 3 — Диапазон регистрации потоков вещества; 5 — Производительность.
3	4	4 — Значение испытательного давления; 9 — Вероятность необнаружения течи; 10 — Время подготовки к работе; 12 — Вероятность безотказной работы.
4	3	13 — Вид выдаваемой информации; 15 — Качество обработки полученной информации (простое измерение, прогноз и т. п.); 19 — Вид средства испытаний (стационарное, мобильное, переносное и т. п.).
5	6	6 — Стоимость средства испытаний; 7 — Квалификация и численность обслуживающего персонала; 8 — Затраты на эксплуатацию; 11 — Уровень вредных и опасных воздействий на людей, объект испытаний, окружающую среду; 14 — Возможность фиксации результатов испытаний; 16 — Возможность объединения средств испытаний в информационную систему.
6	5	17 — Границы применимости средства испытаний в конкретных условиях испытаний; 20 — Возможность модернизации; 22 — Условия эксплуатации и хранения; 23 — Устойчивость к внешним воздействиям; 24 — Требования к источникам энергии для функционирования.

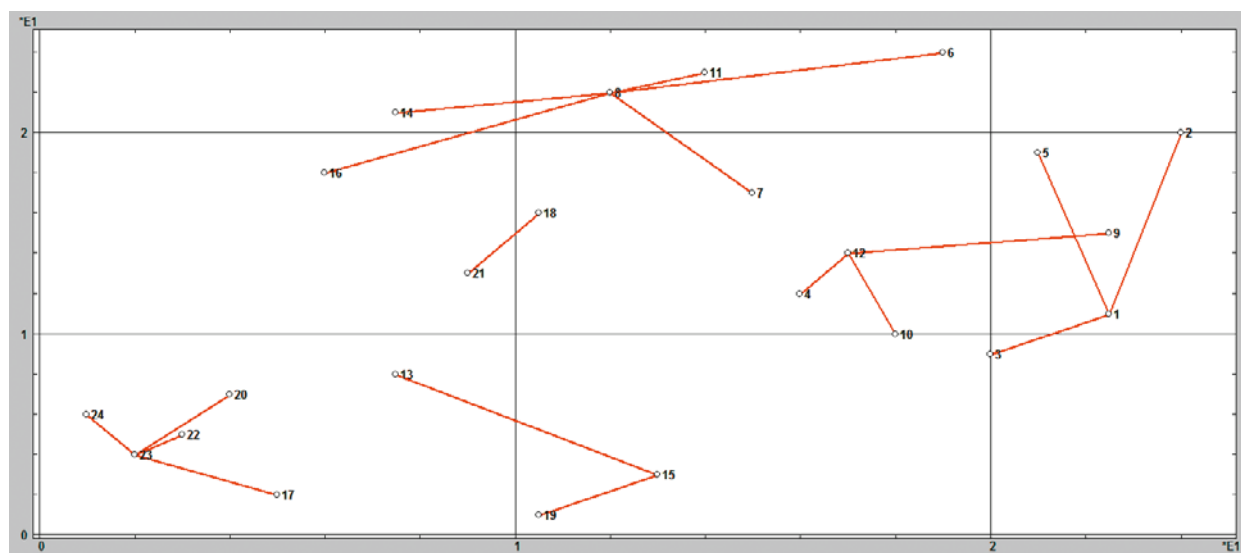


Рис. 6. Проекция шести кластеров на плоскость $X - Y$ (в качестве примера использована плоскость $X3 - Y1$)

Основное его назначение — идентификация и определение однородности групп наблюдаемых объектов и эмпирических данных на основе экспериментальных (эталонных) оценок или выборок [15].

Как показал дискриминантный анализ, предлагаемая классификация — упорядоченность признаков (см. табл. 7) оказалась эффективной ($K = 6$), поскольку определенный уровень значимости равен нулю ($P < 0,05$) для гипотезы о нулевом межкластерном расстоянии D^2 . Причем у всех объектов-признаков из характеристик средств определена сравнительно высокая апостериорная вероятность отнесения к вышеуказанным классам: данный показатель близок к единице. Анализ проводили с использованием программного пакета STADIA 8.0. Результаты представлены ниже.

Дискриминантный анализ. Файл: расстояние Махаланобиса = 244,2, значимость = 0.

Гипотеза 1: <Межкластерное расстояние отличается от нуля>

Класс <— Коэффициенты дискриминантной функции: $a[0], a[1], \dots$ —>

1	-216,8	8,541	10,3	1,475	0,5721	-0,7958
2	-47,21	3,727	4,983	0,6684	0,2046	-0,2704
3	-158,8	7,798	8,878	1,348	0,51	-1,556
4	-4,844	1,054	1,35	0,1963	0,2914	0,04602
5	-54,46	3,751	4,094	0,7196	1,215	-0,04689
6	-144,6	6,405	7,804	1,086	1,368	-0,2997
Объект	Класс	D^2	Значим	Апостеор. вероят.		

1	1	1,208	0,9441	0,9999
2	1	3,359	0,6449	1
3	1	3,547	0,6162	0,9995
4	3	6,478	0,2624	1

5	1	1,525	0,9102	0,9999
6	6	1,23	0,942	1
7	5	3,634	0,6032	0,9991
8	5	2,005	0,8484	1
9	3	7,187	0,2071	0,9999
10	3	5,504	0,3575	0,9981
11	6	1,23	0,942	1
12	3	1,004	0,9623	0,9999
13	2	5,565	0,3509	0,9992
14	5	6,16	0,291	0,9979
15	2	2,753	0,738	1
16	5	5,864	0,3197	1
17	4	4,066	0,54	1
18	5	6,386	0,2704	0,9934
19	2	4,424	0,4902	1
20	4	3,624	0,6047	1
21	2	6,269	0,2809	0,8742
22	4	3,535	0,6181	1
23	4	1,336	0,9311	1
24	4	2,106	0,8343	1

Аналогичный анализ с заданным количеством кластеров, проведенный посредством того же статистического пакета STADIA 8.0, наиболее эффективно определялся цифрой шесть практически с той же самой группировкой признаков по средствам контроля, но на базе группируемых переменных, получаемых при выполнении программы. Приведем результаты анализа.

Дискриминантный анализ. Файл: расстояние Махаланобиса = 79,76, значимость = $7,632E-6$.

Гипотеза 1: <Межкластерное расстояние отличается от нуля>

Класс <— Коэффициенты дискриминантной функции: $a[0], a[1], \dots$ —>

1	-13,8	0,728	0,4998	0,3097	1,016	-0,01648
2	-66,96	3,136	2,367	-0,132	-0,06268	0,8335
3	-44,82	2,969	1,958	0,2093	0,1297	-0,5066
4	-16,83	1,377	1,376	-0,1554	-0,407	0,6652
5	-35,31	1,496	0,8478	-0,2452	1,277	0,8976
6	-2,069	0,2785	0,2332	-0,06242	0,2768	0,2949
Объект	Класс	D^2	Значим	Апостеор. вероят.		
1	2	1,671	0,8925	0,9998		
2	2	4,094	0,536	1		
3	2	4,487	0,4816	0,9998		
4	3	6,49	0,2614	1		
5	2	2,137	0,83	0,9999		
6	5	6,355	0,2732	0,9998		
7	5	5,338	0,376	0,928		
8	5	2,242	0,8147	0,9994		
9	3	6,346	0,274	1		
10	3	7,941	0,1595	0,5942		
11	5	6,024	0,3039	0,9995		
12	3	0,7324	0,9811	0,9998		
13	4	4,707	0,4526	0,9996		
14	5	5,26	0,3849	0,9828		
15	4	1,091	0,9549	1		
16	5	6,636	0,2492	0,9998		
17	6	4,239	0,5155	0,9964		
18	1	1,081	0,9558	0,999		
19	4	2,105	0,8344	0,9988		
20	6	4,095	0,5358	0,9979		
21	1	1,081	0,9558	0,9928		
22	6	3,187	0,6712	0,9849		
23	6	1,406	0,9236	0,9982		
24	6	1,254	0,9396	0,9981		

Для сравнения результатов можно воспользоваться расстоянием Махаланобиса. Чтобы использовать расстояние Махаланобиса в задаче определения принадлежности заданной точки одному из n -го числа классов, нужно найти матрицы ковариации всех классов на основе известных выборок из каждого класса. После этого необходимо рассчитать данное расстояние от заданной точки до каждого класса и выбрать класс, для которого это расстояние минимально. Использование вероятностной интерпретации равносильно (эквивалентно) выбору класса с помощью метода максимального правдоподобия [16]. При сравнительном анализе двух значений расстояний — 244,2 (Statistica) и 79,76 (STADIA) — становится очевидно, что оптимальная упорядоченность должна соответствовать версии результатов, полученных посредством автоматического

выбора группирующих кодов, сгенерированных программным пакетом STADIA 8.0. Тем не менее выбор всегда остается за испытателем, которому с учетом различного рода обстоятельств и с «привязкой» к конкретным условиям применения придется решать, какую из предложенных последовательностей ему выбрать (см. табл. 7 или табл. 9).

Таким образом, задача классификации и необходимость учета при ее решении ряда признаков и порядка выбора этих признаков тесно связаны с качественным анализом [12]. На основе метода экспертных оценок предпринята попытка классифицировать представленные характеристики, относящиеся, главным образом, к средствам испытаний на герметичность, что позволяет исследователю без серьезных временных (материальных) затрат в выборе данных средств выполнить поставленные перед ним задачи. Предлагаемая новая усовершенствованная классификационная структура, построенная на объединении в группы основных характеристик методов и средств испытаний на герметичность с учетом критериев их выбора, дает возможность сформулировать основные методологические подходы и требования к герметичности, ведущие к повышению качества и эффективности контроля.

ЛИТЕРАТУРА

- ГОСТ Р 51780–2001. Контроль неразрушающий. Методы и средства испытаний на герметичность. Порядок и критерии выбора. Госстандарт России. — М.: Изд-во стандартов, 2002. — 4 с.
- Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Исследование зависимостей. Справочное издание. — М.: Финансы и статистика, 1985. — 487 с.
- Штремель М. А., Кудря А. В., Иващенко А. В. Непараметрический дискриминантный анализ в задачах управления качеством / Заводская лаборатория. Диагностика материалов. 2006. Т. 72. № 5. С. 53 – 62.
- Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В. Непараметрические системы классификации. — Новосибирск: Наука, 2000. — 240 с.
- Орлов А. И. Прикладная статистика: Учебник. — М.: Экзамен, 2004. — 656 с.
- Петров П. К. Математико-статистическая обработка и графическое представление результатов педагогических исследований с использованием информационных технологий: учебное пособие. — Ижевск: Удмуртский университет, 2013. — 179 с.
- Стукач О. В. Программный комплекс STATISTICA в решении задач управления качеством: учебное пособие. — Томск: Изд-во Томского политехнического университета, 2011. — 163 с.
- Сидоренко Е. В. Математические методы обработки в психологии. — СПб.: ООО «Речь», 2002. — 350 с.
- Ермолаев О. Ю. Математическая статистика для психологов: Учебник. Изд. 2-е испр. — М.: Флинта, 2003. — 336 с.
- Громова Н. М., Громова Н. И. Основы экономического прогнозирования. Учебное пособие. — М., 2006. — 80 с.
- Буальская А. К., Клейменова Н. А., Горевая М. И. Маркетинговые исследования: Учебное пособие для вузов. — Новосибирск: Сибиринт, 2010. — 404 с.

12. **Дюран Б. и Оделл П.** Кластерный анализ / Пер. с англ. Е. З. Демиденко; под ред. Б. Я. Боярского. — М.: Статистика, 1977. — 128 с.
13. **Боровиков В.** Искусство анализа данных на компьютере: для профессионалов. Изд. 2-е. — СПб.: Питер, 2003. — 688 с.
14. **Кулаичев А. П.** Методы и средства анализа данных в среде Windows STADIA. Изд. 4-е. — М.: Информатика и компьютеры, 2002. — 341 с.
15. **Симчера В. М.** Методы многомерного анализа статистических данных: учебное пособие. — М.: Финансы и статистика, 2008. — 400 с.
16. **Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р. и др.** Факторный, дискриминантный и кластерный анализ / Пер. с англ.; под редакцией И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.

REFERENCES

1. RF State Standard GOST R 51780–2001. Non-destructive testing. Methods and means of leak testing. Order and selection criteria. GOSSTANDART of Russia. — Moscow: Izd. standartov, 2002. — 4 p. [in Russian].
2. **Aivazyan S. A., Enyukov I. S., Meshalkin L. D.** Applied statistics: Research of dependences. Reference book — Moscow: Finansy i statistika, 1985. — 487 p. [in Russian].
3. **Shtremel' M. A., Kudrya A. V., Ivashchenko A. V.** Nonparametric discriminant analysis in quality management tasks / Zavod. Lab. Diagn. Mater. 2006. Vol. 72. N 5. P 53 – 62.
4. **Lapko A. V., Lapko V. A., Sokolov M. I., Chentsov S. V.** Nonparametric classification system. — Novosibirsk: Nauka, 2000. — 240 p. [in Russian].
5. **Orlov A. I.** Applied statistics. Textbook. — Moscow: Ékzamen, 2004. — 656 p. [in Russian].

6. **Petrov P. K.** Mathematico-statistical processing and graphic representation of the results of pedagogical research using information technology: a tutorial. — Izhevsk: Udmurtskii universitet, 2013. — 179 p. [in Russian].
7. **Stukach O. V.** The program complex STATISTICA in the decision of tasks of quality management: a tutorial. — Tomsk: Izd. Tomsk. Politekh. Univ., 2011. — 163 p. [in Russian].
8. **Sidorenko E. V.** Mathematical methods of processing in psychology. — St. Petersburg: JSC “Rech’”, 2002. — 350 p. [in Russian].
9. **Ermolaev O. Yu.** Mathematical statistics for psychologists: Textbook. 2nd edition. — Moscow: Flinta, 2003. — 336 p. [in Russian].
10. **Gromova N. M., Gromova N. I.** The foundations of economic forecasting. Tutorial. — Moscow, 2006. — 80 p. [in Russian].
11. **Buyal'skaya A. K., Kleimenova N. A., Gorevaya M. I.** Marketing Research: textbook for universities. — Novosibirsk: Sibprint, 2010. — 404 p. [in Russian].
12. **Durant B., Odell P.** Cluster analysis / E. Z. Demidenko (transl.), B. Ya. Boyarskii (ed.). — Moscow: Statistika, 1977. — 128 p. [Russian translation].
13. **Borovikov V.** The art of data analysis on the computer: for professionals. 2nd edition. — St. Petersburg: Piter, 2003. — 688 p. [in Russian].
14. **Kulaichev A. P.** Methods and tools for data analysis in Windows STADIA. 4th edition. — Moscow: Informatika i komp'yutery, 2002. — 341 p. [in Russian].
15. **Simchera V. M.** Methods of multivariate analysis of statistical data: a tutorial. — Moscow: Finansy i statistika, 2008. — 400 p. [in Russian].
16. **Kim J.-O., Mueller Ch. W., Klecka W. R., et al.** Factorial, discriminant and clattery analysis / I. S. Enyukov (ed.). — Moscow: Finansy i statistika, 1989. — 215 p. [in Russian].