УДК 621.019

Э.М. Фархадзаде, А.З. Мурадалиев доктора техн. наук,
Ю.З. Фарзалиев, канд. техн. наук
Азербайджанский научно-исследовательский
и проектно-изыскательский ин-т энергетики
(Азербайджанская Республика, Az 1012 Баку, пр. Зардаби, 94,
тел. (+99412) 4316407, e-mail: fem1939@rambler.ru)

Оценка целесообразности классификации многомерных данных по заданному признаку

Показано, что оценка целесообразности классификации статистических данных о надежности сводится к сопоставлению статистических функций распределения совокупности многомерных данных и выборки. Эффективность критерия сравнения обратно пропорциональна ошибке первого рода, вычисляемой согласно распределениям статистик.

Показано, що оцінка доцільності класифікації статистичних даних про надійність полягає у зіставленні статистичних функцій розподілу сукупності багатовимірних даних і вибірки. Ефективність критерію порівняння обернено пропорціональна похибці першого роду, яка обчислюється згідно розподілу статистик.

К л ю ч е в ы е с л о в а: классификация, многомерные данные, критерий, ошибка.

Классификация оборудования и устройств (объектов) электроэнергетических систем по надежности их работы является одной из наиболее важных и трудных задач. Актуальность решения этой задачи обусловлена возможностью снижения эксплуатационных затрат вследствие дифференцирования правил технического обслуживания и ремонта. Трудность ее решения заключается в том, что информация о надежности объектов на самом деле имеет многомерный характер, т.е. статистические данные об отказах, дефектах, результатах испытания и восстановления износа зависят от большого числа признаков и их разновидностей, которые должны быть предварительно заданы в паспортных данных и перечне условий эксплуатации. При оценке усредненных показателей надежности многомерный характер статистических данных не учитывается. Ранжирование разновидностей признаков по значимости требует, прежде всего, разработки соответствующих методов, алгоритмов и программных средств, отражающих как природу этих данных (непрерывных или дискретных), так и особенности их признаков.

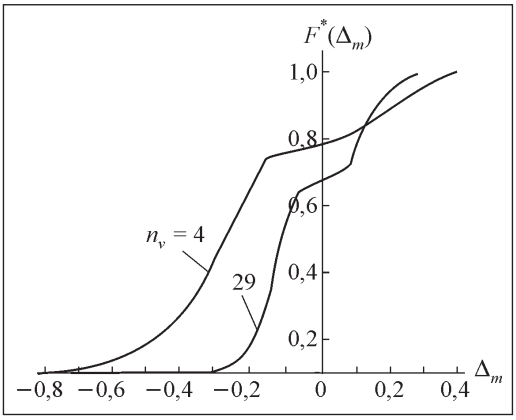

Рис. 1. Кривые СФР $F^*(\Delta_m)$ при числе итераций $N = 25000$ для различных значений n_v

Рассмотрим результаты исследований одного из основных вопросов — как оценить целесообразность классификации многомерных данных по заданному признаку.

Постановка задачи. Предположим, что известна база данных силовых трансформаторов, включающая перечень нерабочих, в том числе аварийных, состояний. Обозначим множество данных о длительности аварийного простоя τ_a как $\{\tau_a\}_m$, где m — число случайных величин τ_a. Это множество в математике принято называть конечной совокупностью многомерных данных. Статистическую функцию распределения (СФР) этих данных обозначим $F^*_\Sigma(\tau_a)$. Пусть заданной разновидностью признака будет «класс напряжения 110кВ.». СФР выборки из конечной совокупности многомерных данных обозначим $F^*_v(\tau_a)$. Необходимо проверить предположение о том, что конечная совокупность многомерных данных и выборка однородны, т.е. их СФР $F^*_\Sigma(\tau_a)$ и $F^*_v(\tau_a)$ различны случайно (гипотеза H_1).

На практике для статистической проверки этого предположения чаще всего применяется критерий Колмогорова, основанный на статистике D_n [1] и относящийся к группе непараметрических. Этот критерий с успехом может быть использован для сравнения как $F^*_\Sigma(\tau_a)$ и $F^*_v(\tau_a)$. Формулы и таблицы для применения этого критерия приведены во многих монографиях и учебных пособиях, и практически везде отмечена ошибочность нахождения величины наибольшего вертикального расхождения функций распределения $F_\Sigma(X)$ случайной величины X и СФР выборки $F^*_v(X)$ как максимального значения абсолютных величин наблюдаемых значений Δ. Однако ни в одном из многих руководств по математической статистике не указана причина этой ошибки.

Статистика Δ_m характеризует вертикальное отклонение $F_\Sigma(X)$ от $F^*_v(X)$. Реализация этой статистики выполняется по следующему алгоритму:

расчет $\Delta_i = (X_i - i/n_v),\ i = 1, n_v$;

определение $\Delta_{m,1} = \min\{\Delta_1, \Delta_2, ..., \Delta_i, \Delta_{nv}\}$, $\Delta_{m,1} = \max\{\Delta_1, \Delta_2, ..., \Delta_i, \Delta_{nv}\}$; если $|\Delta_{m,1}| > |\Delta_{m,2}|$, то $\Delta_m = \Delta_{m,1}$, иначе $\Delta_m = \Delta_{m,2}$. Здесь X — случайные числа с равномерным распределением в интервале [0, 1], имитирующие истинные значения квантилей равномерного распределения.

СФР $F^*(\Delta_m)$, построенные по 25000 реализациям Δ_m для ряда n_v, приведены на рис. 1. Важность этих исследований заключается в том, что с достаточной для практики точностью установлены границы изменения значения Δ_m, характеризующего наибольшее вертикальное расхождение $F^*_\Sigma(\xi)$ и $F^*_v(\xi)$ с заданным уровнем значимости. Установлено, что квантили распределения $F^*(\Delta_m)=\alpha \le 0{,}1$ при $n \ge 2$ равны по величине и противоположны по знаку квантилям распределения $F(D_n)=2\alpha$. Это означает, что, характеризуя вертикальное расхождение распределений, D_n все же не являются наибольшим вертикальным расхождением.

Распределение Δ_m имеет свои особенности. В данном случае Δ_m рассматривается как некоторое числовое значение, ранжированное в порядке возрастания. Если величину Δ_m рассматривать как наибольшее вертикальное расхождение СФР $F_\Sigma(X)$ и $F^*_v(X)$, то функция $F^*(\Delta_m)$ — не есть СФР. Причиной этого является наличие положительных и отрицательных значений Δ_m. Закономерности изменения распределения положительных и отрицательных значений Δ_m при $n_v = 4$ представлены на рис. 2, а соотношение их числа для ряда n_v приведено в табл. 1. Как следует из табл. 1, с увеличением значения n_v соотношение отрицательных и положительных значений Δ_m убывает, но и при $n_v = 150$ оно еще не равно единице.

В табл. 2 приведены экспериментальные и расчетные значения квантилей распределения $F^*_v(\Delta_m)$ для ряда n_v и вероятностей $R^*_v(\Delta_m)=$ $=\left\lfloor 1-F^*_v(\Delta_m)\right\rfloor=\alpha$. Экспериментальные значения $(\Delta^{э}_{m;\,0,\,5\alpha})$ получены с помощью имитационного моделирования на ЭВМ [2], а расчетные $(\Delta^{р}_{m;\,(1-0,\,5\alpha)})$ — по формуле

$$\Delta^{р}_{m;(1-0,\,5\alpha)} = -[\Delta^{э}_{m;0,\,5\alpha} + n_v^{-1}]. \qquad (1)$$

Таблица 1

Число случайных величин в выборке n_v	Относительное число отрицательных значений Δ_m	Соотношение отрицательных и положительных значений Δ_m
2	0,87	6,7
4	0,79	3,8
7	0,73	2,7
11	0,68	2,1
16	0,65	1,9
22	0,63	1,7
29	0,61	1,6
150	0,55	1,2

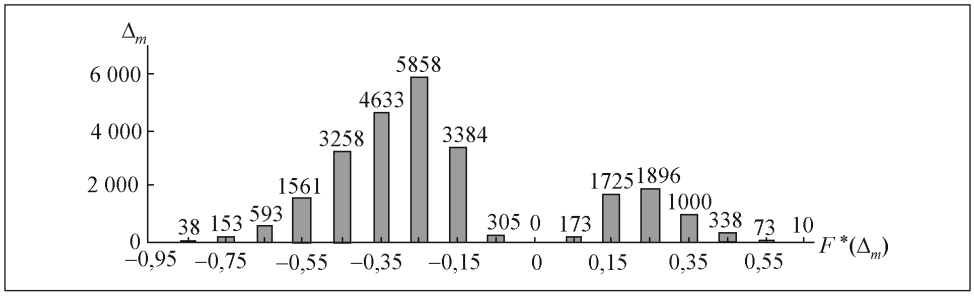

Рис. 2. Гистограмма наибольшего вертикального распределения $F_{\Sigma}^{*}(X)$ и $F_{v}^{*}(X)$

Данные, приведенные в табл. 2, свидетельствуют о том, что формула (1) достаточно точно отображает взаимосвязь граничных значений интервала изменения статистики Δ_m при условии $\alpha \le 0{,}25$.

Статистики B_v. В табл. 3 приведены квантили распределения B_V для заданного числа элементов выборки n_v и вероятностей $F^{*}(B_v)$, характеризующие наибольшее абсолютное расхождение. Алгоритм расчета следующий:

вычисление

$$\Delta_i = (X_i - i/n_v), \quad i = 1, n_v; \tag{2}$$

определение

$$B_v = \max\{|\Delta_1|, |\Delta_2|, ..., |\Delta_i|, |\Delta_{nv}|\}. \tag{3}$$

На рис. 3 приведены кривые СФР $R^{*}(B_v) = 1 - F^{*}(B_v)$ для различных значений n_v, построенные по данным табл. 3. Как и следовало ожидать, с

Таблица 2

$R_v^{*}(\Delta_m)$	Значение $F_v^{*}(\Delta_m)$ при n_v						
	2	4	6	11	40	90	150
0,025	0,343 0,342	0,377 0,373	0,358 0,356	0,302 0,298	0,185 0,183	0,131 0,130	0,104 0,103
0,05	0,285 0,285	0,319 0,317	0,303 0,302	0,260 0,262	0,164 0,162	0,116 0,116	0,092 0,092
0,1	0,184 0,184	0,240 0,244	0,244 0,244	0,216 0,218	0,140 0,139	0,100 0,100	0,079 0,079
0,2	0,060 0,061	0,160 0,165	0,171 0,171	0,160 0,164	0,112 0,116	0,091 0,081	0,065 0,064
0,3	–0,239 –0,027	–0,173 0,105	–0,127 0,125	–0,097 0,128	0,089 0,094	0,067 0,068	0,053 0,055

Примечание: над чертой указаны экспериментальные значения, под чертой — значения, полученные в результате расчета.

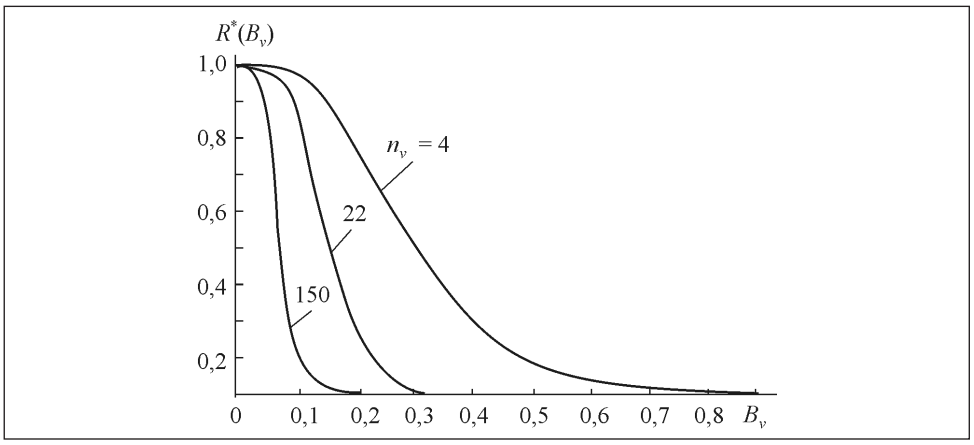

Рис. 3. Кривые СФР $R^*(B_v) = 1 - F^*(B_v)$ для различных значений n_v

Таблица 3

N	$F^*(B_v)$	Значение B_v для заданного числа n_v								
		2	4	7	11	22	29	40	90	150
1	0,05	0,112	0,127	0,116	0,104	0,083	0,075	0,067	0,048	0,038
2	0,1	0,157	0,154	0,136	0,120	0,094	0,084	0,075	0,053	0,042
3	0,15	0,193	0,175	0,151	0,131	0,103	0,092	0,081	0,057	0,045
4	0,2	0,223	0,191	0,164	0,142	0,110	0,098	0,087	0,061	0,048
5	0,25	0,249	0,208	0,177	0,152	0,117	0,104	0,092	0,064	0,051
6	0,3	0,274	0,222	0,189	0,160	0,124	0,110	0,097	0,067	0,053
7	0,35	0,300	0,236	0,201	0,170	0,130	0,115	0,101	0,071	0,056
8	0,4	0,324	0,250	0,213	0,179	0,136	0,121	0,106	0,074	0,058
9	0,45	0,348	0,268	0,225	0,189	0,143	0,127	0,111	0,077	0,061
10	0,5	0,376	0,286	0,236	0,198	0,150	0,132	0,116	0,080	0,063
11	0,55	0,401	0,306	0,249	0,209	0,157	0,139	0,121	0,083	0,066
12	0,6	0,426	0,326	0,262	0,219	0,164	0,145	0,127	0,087	0,069
13	0,65	0,449	0,348	0,277	0,231	0,172	0,152	0,133	0,091	0,072
14	0,7	0,473	0,370	0,294	0,244	0,181	0,160	0,139	0,095	0,075
15	0,75	0,499	0,393	0,313	0,258	0,191	0,169	0,147	0,100	0,079
16	0,8	0,548	0,421	0,334	0,276	0,203	0,179	0,155	0,106	0,083
17	0,85	0,620	0,454	0,358	0,295	0,217	0,191	0,166	0,112	0,088
18	0,9	0,683	0,497	0,391	0,322	0,235	0,206	0,179	0,122	0,096
19	0,95	0,778	0,568	0,442	0,362	0,263	0,232	0,201	0,136	0,107
20	0,99	0,902	0,689	0,538	0,440	0,320	0,283	0,240	0,164	0,129

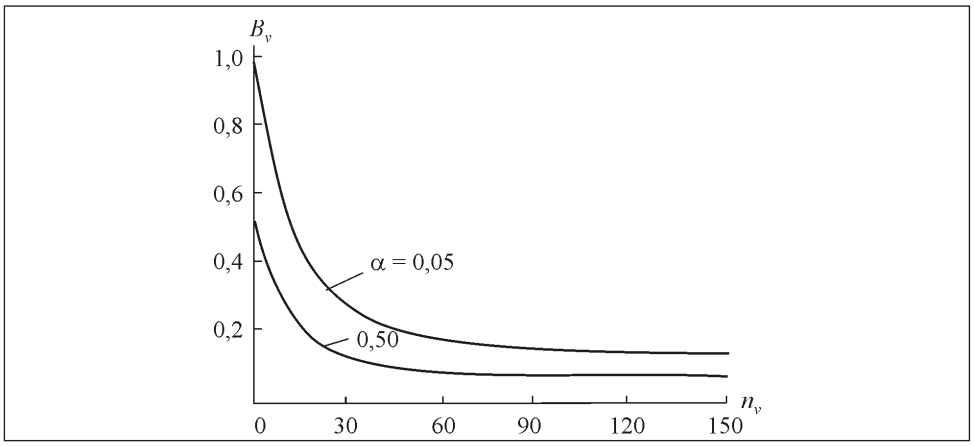

Рис. 4. Кривые зависимости изменения абсолютной величины наибольшего вертикального отклонения от числа элементов выборки n_v при различных значениях α

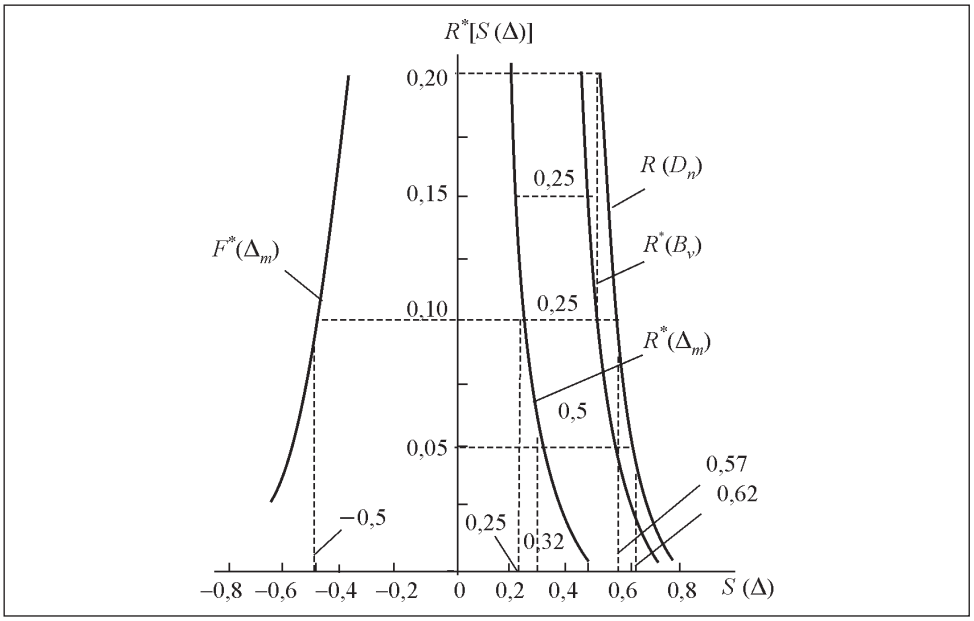

Рис. 5. Особенности распределения статистик D_n, Δ_m, и B_v при $n_v = 4$

увеличением значения n_v критические значения B_v уменьшаются. Изменяется и характер распределения $R(B_v)$.

В табл. 4 приведены коэффициенты уравнения $B_v = An_v^{-b}$ и коэффициент детерминизации R^2, рассчитанные по данным табл. 3. В качестве примера на рис. 4 представлены закономерности изменения кривой $B_v = An_v^{-b}$ при различных значениях α.

Таблица 4

$F^*(B_v)$	A	b	R^2
0,9	1,079	0,459	0,9998
0,9	0,942	0,453	0,9997
0,8	0,774	0,439	0,9986
0,7	0,668	0,430	0,9982
0,6	0,590	0,422	0,9985
0,5	0,518	0,412	0,9975
0,4	0,447	0,396	0,9956
0,3	0,384	0,382	0,9922
0,2	0,317	0,360	0,9862
0,1	0,236	0,321	0,9829

Анализ соотношения статистик D_n, Δ_m, и B_v. Распределение $F^*(B_v)$ имеет существенное преимущество по сравнению $F^*(\Delta_m)$. Оно характеризует распределение величины наибольшего отклонения СФР $F^*_\Sigma(X)$ и $F^*_v(X)$ без учета знака отклонения, т.е. считаются равнозначными положительное и отрицательное отклонения Δ_m. Если сопоставить данные табл. 2 и 3, нетрудно заметить существенное различие критических значений. Так, при $R^*(D_m) = R^*(B_v) = R^*(\Delta_m) = 0{,}05$ и $n_v = 4$ соответствующие квантили составляют $D_n = 0{,}62$, $B_v = 0{,}57$ и $\Delta_m = 0{,}32$.

На рис. 5. приведены участки функций $R(D_n) = 1 - F(D_n)$, $R^*(\Delta_m) = 1 - F^*(\Delta_m)$ и $R^*(B_v) = 1 - F^*(B_v)$ при $n_v = 4$, соответствующие следующим соотношениям:

$$R(D_n) = 2R^*(B_v = D_n), \tag{4}$$

$$\overline{\Delta_m} = -\left[\underline{\Delta_m} + \frac{1}{n_v}\right]. \tag{5}$$

Например, для (4) при $D_n = 0{,}5$ $R(D_n) = 0{,}2$, а $R^*(B_v = 0{,}5) = 0{,}1$; при $D_n = 0{,}57$ $R(D_n = 0{,}57) = 0{,}1$, а $R^*(B_v = 0{,}57) = 0{,}05$. Для (5), например, при $\underline{\Delta_m} = -0{,}5$ и $n_v = 4$ $\overline{\Delta_m} = 0{,}32$.

Таким образом, при практическом применения критерия Колмогорова суть ошибки часто заключается в том, что вычисляемая по формулам (2) и (3) статистика B_v сопоставляется не с критическим значением B_k при уровне значимости $\alpha = R(B_k)$, а с критическим значением статистики D_n.

Выводы

1. Каждую из вводимых в рассмотрение статистик, например D_n, B_v или Δ_m, при проверке предположения о случайном расхождении распределений $F_{\Sigma}^{*}(X)$ и $F_{v}^{*}(X)$ следует сопоставлять лишь со своими критическими значениями, вычисляемыми по распределениям соответственно $F^{*}(D_n)$, $F^{*}(B_v)$ и $F^{*}(\Delta_m)$.

2. Сопоставление критериев проверки предположений сводится к сопоставлению значимости особенностей случайных величин, определяемых статистиками D_n, Δ_m, и B_v.

СПИСОК ЛИТЕРАТУРЫ

1. *Рябинин И.А.* Основы теории и расчета надежности судовых электроэнергетических систем. — Л.: «Судостроение», 1971. — 453 с.
2. *Farhadzade E.M., Muradaliyev A.Z., Farzaliyev Y.Z.* Decrease in risk erroneous classification the multivariate statistical data describing the technical condition of the equipment of power supply system // J. Reliability: Theory&applications. (R&RATA). — 2013. — Vol. 8, No. 2 (29). — P. 55—64.

*E.M. Farhadzade, A.Z. Muradaliyev, Y.Z. Farzaliye*v

ESTIMATION OF EXPEDIENCY OF CLASSIFICATION
OF MULTIVARIATE DATA TO THE SET ATTRIBUTE

It shown, that the estimation of expediency of classification of statistical data about reliability is reduced to comparison of statistical functions of distribution of totality of multivariate data and samples. Efficiency of the criterion of comparison is inversely proportional to the first kind error calculated according to statistics distributions.

K e y w o r d s: classification, multivariate data, criterion, error.

REFERENCES

1. Ryabinin, I.A. (1971) *Osnovy teorii I raschota nadyozhnosti sudovykh elektroenergeticheskikh sistem*. [Bases of the theory and calculation of reliability of ship electric power systems], Sudostroenie, Leningrad, Russia.
2. Farhadzade, E.M., Muradaliyev, A.Z. and Farzaliyev, Y.Z. (2013) “Decrease in risk erroneous classification of the multivariate statistical data describing the technical condition of the equipment of power supply system”, *J. Reliability: Theory & Applications (R&RATA)*, Vol. 8, no. 2(29), pp. 55-64.

Поступила 04.10.13

ФАРХАДЗАДЕ Эльмар Мехти оглу, д-р техн. наук, профессор, руководитель лаборатории «Надежность оборудования энергосистем» Азербайджанского научно-исследовательского и проектно-изыскательского ин-та энергетики (г. Баку). В 1961 г. окончил Азербайджанский ин-т нефти и химии. Область научных исследований — надежность и эффективность электроэнергетических систем.

МУРАДАЛИЕВ Айдын Зураб оглу, д-р техн. наук, руководитель отдела «Надежность оборудования энергосистем» Азербайджанского научно-исследовательского и проектно-изыскательского ин-та энергетики (г. Баку). В 1982 г. окончил Азербайджанский ин-т нефти и химии. Область научных исследований — количественная оценка индивидуальной надежности оборудования и устройств электроэнергетических систем.

ФАРЗАЛИЕВ Юсиф Зейни оглу, канд. техн. наук, главный специалист лаборатории «Надежность оборудования энергосистем» Азербайджанского научно-исследовательского и проектно-изыскательского ин-та энергетики (г. Баку). В 1985 г. окончил Азербайджанский госуниверситет. Область научных исследований — точность и достоверность оценок показателей индивидуальной надежности оборудования и устройств энергетических систем.