МОДЕЛЬ ИНФОРМАТИВНОСТИ ДАННЫХ НА ОСНОВЕ ПАКЕТОВ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ НЕФТЕГАЗОВОЙ ОТРАСЛИ

И.О. Орлова, *доцент*
Е.Н. Даценко, *доцент*
Н.Н. Авакимян, *доцент*
В.С. Гнеуш, *магистрант*
Кубанский государственный технологический университет
(Россия, г. Краснодар)

DOI:10.24412/2500-1000-2023-1-2-94-99

Аннотация. *В статье рассматривается модель расчета информативности признаков и диагностических коэффициентов для отнесения нефтяной скважины к одному из множеств на примере пакетов математического моделирования MathCAD. В исследовании проведен метод расчета и статистический анализ данных. На основании данного анализа, выявлено, что данный расчет повысит эффективность и качество обработки информации в нефтегазовой отрасли.*

Ключевые слова: *модель, информативность, диагностический коэффициент, распознавание, прогноз, множество, нефтедобыча, скважина, пакет прикладных программ.*

Для повышения эффективности нефтедобычи предполагается, в том числе, повысить качество обработки информации, используемой в рассматриваемой отрасли. Качество обработки информации оценивается не только дисперсионным анализом [1], но и также на основе определения информативности и диагностических коэффициентов используемых факторов или признаков. Организация вычислительного процесса может быть реализована на базе пакетов прикладных программ пакетов математического моделирования, например *MathCAD , Maple* и др. [2].

Основная часть.

Допустим, два множества объектов располагают общим для них признаком. Если значения данного признака различны для каждого множества объектов, то признак считается информативным, поскольку отделяет одно множество объектов от другого множества объектов. В противном случае, признак не обладает информативностью, поскольку не различает объекты, относящиеся к различным множествам [3].

Чем больше множеств, групп или образов различает признак, тем выше его информативность и наоборот. При этом дисперсионный анализ, а именно, критерий Фишера [4] и иные статистические критерии не в состоянии рассчитать величину информативности признака. Количественно оценить информативность признаков в состоянии осуществить мера Кульбака [5].

Рассмотрим такой расчет на конкретной задаче.

Задача. Даны коэффициенты нефтеотдачи для 115 объектов нефтедобычи, которые описываются следующими факторами (признаками):

1) количеством закачанной воды $V_{зан}$ в объёмах нор;
2) темп разработки T;
3) проницаемостью пласта, $k, мД$;
4) плотностью сетки скважин $S, \text{га}/\text{скв}$;
5) содержание глины в коллекторах $C_г, \%\,\text{вес}$;
6) содержание смол в нефти $C_{см}, \%\,\text{вес}$;

Объекты распределяются на два множества A и B: у первого множества коэффициенты нефтеотдачи $\eta < 0{,}4$, у второго множества $\eta \geq 0{,}4$. Различия в значениях признаков для пары множеств заключаются в следующем.

Для первого из признаков -количество прокачанной воды, - имеется диапазон значений $[0;450]$. Указанный диапазон разбивается на 9 интервалов (Таблица 1),

количество интервалов может быть от 8 до 12 интервалов.

Расчёты в ячейках Таблицы 1 производятся с помощью пакетов прикладных программ *MathCAD*. (Рисунок 1).

В графы 3 и 4 помещаются данные по частоте попадания месторождений из множеств A и B в каждый из интервалов. Графы 5 и 6 содержат данные относительных частостей в процентах, при этом принимается за 100% сумма частостей применительно к A и B по всем диапазонам.

Таблица 1. Расчёт информативности для признака 1

Номер интервала	Интервал	Частота попадания в группы		Частость, %				Отношение сглаженных частостей $\bar{y}_A/\bar{y}_B$	ДК	$J_{расч}$
				вероятная		сглаженная				
		A	B	A	B	A	B			
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
-1	-	0	0	0	0	6	2	-	-	-
0	-	0	0	0	0	14	7	-	-	-
1	0,0-50	24	18	58	24	28	16	1,92	3	0,18
2	50,1-100	9	19	22	25	23	18	1,28	1	0,03
3	100,1-150	3	10	7	13,5	15	15	1	0	0
4	150,1-200	3	4	7	5	7	10	0,7	-2	0,03
5	200,1-250	2	7	5	9	4	8	0,5	-3	0,06
6	250,1-300	0	5	0	7	2	7	0,49	-5	0,12
7	300,1-350	0	3	0	4	0,5	6	0,08	-11	0,30
8	350,1-400	0	6	0	8	0	5	0	0	0
9	400,1-450	0	3	0	3	0	3	0	0	0
		41	**75**	**99**	**98,5**	**99,5**	**97**	-	-	**0,72**

В интервале $[0;50]$ для группы A имеется значение вероятной частости $(24:41)\cdot 100\% = 58\%$. Поскольку итоговые значения частости зависят от выбора границ интервалов, то снижения указанного влияния определяются средневзвешенные (сглаженные) значения частостей с учётом значений данного параметра в 4х смежных диапазонах следующим образом:

$$\bar{y}_2 = (y_1 + 2y_2 + 4y_3 + 2y_4 + y_5)/10 \qquad (1)$$

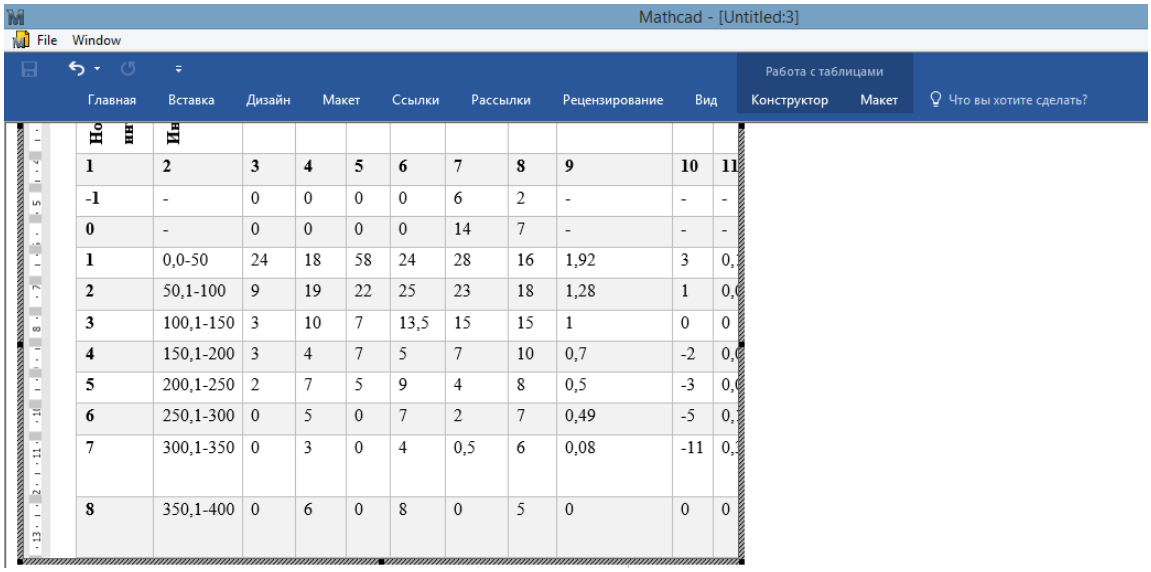

Рис. 1. Фрагмент ввода исходных данных в таблице для вычисления в Mathcad

Для первого интервала вводятся дополнительно несуществующие интервалы 0 и -1, в которых в связи с отсутствием наблюдений, частости в диапазонах нулевые: $y_0 = y_{-1} = 0$.

Сглаженная частость для первого и второго интервала для группы A рассчитывается так:

$$\bar{y}_{1A} = (0 + 0 + 4y_1 + 2y_2 + y_3)/10 = (0 + 0 + 4 \cdot 58 + 2 \cdot 22 + 7)/10 \approx 28$$
$$\bar{y}_{2A} = (0 + 2y_1 + 4y_2 + 2y_3 + y_4)/10 = (0 + 2 \cdot 58 + 4 \cdot 22 + 2 \cdot 7 + 7)/10 \approx 23$$

Сглаженные значения частостей в % округляются до целых значений, при значениях меньше 5% округление производится до 1 знака после запятой. В столбце 9 приведено отношение сглаженных частостей $\bar{y}_A / \bar{y}_B$. В столбце 10-диагностические коэффициенты (*ДК*), которые вычисляются следующим образом:

$$ДК = 10\lg(\bar{y}_A / \bar{y}_B) \quad (2)$$

Поскольку сглаженные значения частостей имеются в интервалах 0 и -1, то средневзвешенные величины $\bar{y}_1, \bar{y}_0$ и $\bar{y}_{-1}$ суммируются, а полученная сумма считается средневзвешенной частостью $\bar{y}_1$ признака для первого интервала:

$$\frac{\bar{y}_{A1}}{\bar{y}_{B1}} = \frac{48}{25} = 1{,}92; \qquad ДК = 10\lg 1{,}92 = 3.$$

Столбец 11 Таблицы 1 заполняются значениями информативности признака для всех диапазонов.

В соответствие с формулой Кульбака величина информативности J $i-го$ интервала $j-го$ признака рассчитывается следующим образом:

$$J(x_j^i) = ДК(x_j^i)\frac{1}{2}\left[P(\frac{x_j^i}{A}) - P(\frac{x_j^i}{B}) \right], \quad (3)$$

где $ДК(x_j^i)$-диагностический коэффициент $i-го$ интервала $j-го$ признака; $P(\frac{x_j^i}{A})$-вероятность (сглаженная частость) того, что в группе A $i-го$ интервала отмечено попадание $j-го$ признака, $\bar{y}_{A1}$; $P(\frac{x_j^i}{B}) = \bar{y}_{Bi}$.

В составе диагностической таблицы определяется информативность признака во всех интервалах и находится совокупная информативность признака x_j:

$$J(x_j) = \sum_i J(x_j^i) \quad (4)$$

Информативность показателя «количество закачанной воды» для первого интервала

$[0;50]$равна: $J = 3 \cdot \frac{1}{2}(0{,}28 - 0{,}16) = 0{,}18$, для второго интервала - $[50;100]$: $J = 1 \cdot \frac{1}{2}(0{,}23 - 0{,}18) = 0{,}025$. Информативность рассматриваемого признака вычисляется как сумма информативности в диапазонах $J_B = 0{,}72$. Таким же образом вычислены информативности остальных указанных вначале признаков (Таблица 2, 3, 4).

Таблица 2. Расчёт информативности для признака 2

Номер интервала	Интервал	Частота попадания в группы		Частость, %				Отношение сглаженных частостей $\bar{y}_A/\bar{y}_B$	ДК	$J_{расч}$
				вероятная		сглаженная				
		A	B	A	B	A	B			
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
-1	-	0	0	0	0	4	1,5	-	-	-
0	-	0	0	0	0	10	5	-	-	-
1	0,0-0,08	16	11	39	15	20	10	2,2	3	0,29
2	0,08-0,16	9	13	22	17	20	14	1,43	2	0,05
3	0,16-0,24	5	8	12	11	15	13	1,15	1	0,01
4	0,24-0,32	4	11	10	15	11	11	1	0	0
5	0,32-0,40	1	2	2,4	2,7	6	6	1	0	0
6	0,40-0,48	4	2	10	2,7	6	5	1,2	1	0,01
7	0,48-0,56	1	5	2,4	7	4	5	0,8	-1	0,01
8	0,56-0,64	1	3	2,4	4	2	5	0,4	-4	0,06
9	0,64-0,72	0	4	0	5	0,7	5	0,14	-8	0,18
10	0,72-0,80	0	4	0	5	0,2	6	0,03	-15	0,45
11	0,80-0,88	0	4	0	5	0	6	0	-	-
12	0,88-0,96	0	8	0	11	0	6	0	-	-
		41	**75**	**100,2**	**100,4**	**100,0**	**98,5**	-	-	**J=1,05**

Таблица 3. Расчёт информативности для признака 3

Номер интервала	Интервал	Частота попадания в группы		Частость, %				Отношение сглаженных частостей $\bar{y}_A/\bar{y}_B$	ДК	$J_{расч}$
				вероятная		сглаженная				
		A	B	A	B	A	B			
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
-1	-	0	0	0	0	2	0	-	-	-
0	-	0	0	0	0	6	1	-	-	-
1	0,0-50	7	0	17	0	14	3	5,50	7	0,63
2	50,1-100	12	8	29	11	20	8	2,50	4	0,24
3	100,1-150	7	11	17	15	17	12	1,40	1	0,02
4	150,1-200	5	7	12	9	12	12	1,00	0	0
5	200,1-250	2	12	5	16	7	13	0,54	-3	0,09
6	250,1-300	0	7	0	9	4	12	0,33	5	0,20
7	300,1-350	2	10	5	13	5	12	0,42	-4	0,14
8	350,1-400	4	9	10	11	5	9	0,56	-2,5	0,05
9	400,1-450	0	4	0	5	3	6	0,50	-3	0,05
10	450,1-500	1	0	2	0	2	3	0,67	-2	0,01
11	500,1-550	0	2	0	3	1	3	0,33	-5	0,05
12	550,1-600	1	3	2	4	1	2	0,50	-3	0,01
13	600,1-650	0	1	0	1	0,4	2	0,20	-7	0,06
14	650,1-700	0	1	0	1	0,2	1	0,20	-7	0,03
		41	**75**	**99,0**	**99,6**	**99,0**	**99,0**			**J=1,58**

Т

аблица 4. Расчёт информативности для признака 4

Номер интервала	Интервал	Частота попадания в группы		Частость, %				Отношение сглаженных частостей $\bar{y}_A/\bar{y}_B$	ДК	$J_{расч}$
				вероятная		сглаженная				
		A	B	A	B	A	B			
1	2	3	4	5	6	7	8	9	10	11
-1	-	0	0	0	0	1	2	-	-	-
0	-	0	0	0	0	5	9	-	-	-
1	0-7	4	19	10	25	12	20	0,58	-2	0,13
2	7-14	13	32	32	43	18	27	0,67	-2	0,09
3	14-21	5	14	12	19	16	21	0,76	-1	0,02
4	21-28	5	6	12	8	15	12	1,25	1	0,02
5	28-35	7	2	17	2,7	12	5	2,40	4	0,14
6	35-42	3	0	7	0	9	1,4	6,40	8	0,32
7	42-49	2	0	4,9	0	5	0,6	8,30	9	0,20
8	49-56	0	0	0	0	2	0,6	3,30	5	0,03
9	56-63	0	2	0	2,7	1	1,2	0,83	-1	0,00
10	63-70	0	0	0	0	1	0,5	2,00	3	0,01
11	70-77	2	0	49	0	2	0,3	6,67		0,07
		41	**75**	**99,8**	**100,4**	**99,0**	**100,6**	-	-	**J=1,03**

Значения информативности для признаков 5 и 6 равны: $J_{\Gamma} = 0{,}099; J_c = 0{,}089$. Ввиду малого значения информативности, рекомендуется их не учитывать и не приводить расчётные таблицы, а признаки считать не существенными. После расчёта информативности строится диагностическая таблица для распознавания образов A и B по всем признакам. Распознавание осуществляется по формуле:

$$порог\ A < \frac{P(x_1^1/A)}{P(x_1^1/B)}\frac{P(x_2^2/A)}{P(x_2^2/B)}\ldots\frac{P(x_I^I/A)}{P(x_I^I/B)} < порог\ B \qquad (5)$$

где $\frac{P(x_I^I/A)}{P(x_I^I/B)}$-отношение частости одинакового интервала одного признака одной и другой группы. Последовательно перемножая отношения частости добиваются достижения наиболее близкого порога одного из множеств A или B, что позволит сделать вывод о принадлежности изображения или объекта к такому множеству. Если при использовании всей признаковой информации порог не достигнут, то результат распознавания не получен. Повышение информативности признаков будет способствовать правильному распознаванию одного из множеств A или B. В результате были рассчитаны диагностические коэффициенты для 115 скважин, 39 скважин объединены в множество A, остальные-в множество B.Обучение проведено на 40 скважинах, поровну для каждого из множеств. Граница между множествами A и B определена как 0. Отрицательные суммы ДК (до -25) свидетельствуют в пользу множества A, положительные суммы ДК (до 25)-в пользу B. Таким образом, показано, как распознавание множеств может быть реализовано с привлечением пакетов прикладных программ *MathCAD*, что позволяет оперативно получать результаты по распознаванию и прогнозу скважин.

Заключение.

В данной статье показано, как учет информативности признаков и диагностической информации позволил с привлечением пакетов прикладных программ *MathCAD* оперативно распределить по двум множествам 115 нефтяных скважин с учётом значений информативности признаков и диагностических коэффициентов.

Библиографический список

1. Дисперсионный анализ. – [Электронный ресурс]. – Режим доступа: http://statsoft.ru/home/textbook/modules/stanman.html#basic (дата обращения 15.01.2023).
2. Таранчук В. Б. Основные функции систем компьютерной алгебры. – Минск: БГУ, 2013. – 59 с.
3. Фомин Я.А. Распознавание образов: теория и применения. – 2-е изд. – М.: ФАЗИС, 2012. – 429 с. – ISBN 978-5-7036-0130-4.
4. F-Test for Equality of Two Variances. – [Электронный ресурс]. – Режим доступа: https://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm (дата обращения 15.01.2023).
5. Kullback S. Information Theory and Statistics. – John Wiley & Sons, 1959.

DATA INFORMATIVENESS MODEL BASED ON APPLICATION SOFTWARE PACKAGES FOR THE OIL AND GAS INDUSTRY

I.O. Orlova, *Associate Professor*
E.N. Datsenko, *Associate Professor*
N.N. Avakimyan, *Associate Professor*
V.S. Gneush, *Graduate Student*
Kuban State Technological University
(Russia, Krasnodar)

Abstract. *The article considers a model for calculating the informative value of signs and diagnostic coefficients for assigning an oil well to one of the sets using the example of MathCAD mathematical modeling packages. The study carried out the method of calculation and statistical analysis of the data. Based on this analysis, it was revealed that this calculation will increase the efficiency and quality of information processing in the oil and gas industry.*

Keywords: *model, informativeness, diagnostic coefficient, recognition, forecast, set, oil production, well, application software package.*