МЕТОДОЛОГИЯ КЛАССИФИКАЦИИ ПО ОБУЧАЮЩИМ ВЫБОРКАМ В СОЦИАЛЬНО-ЭКОНОМИЧЕСКОЙ СФЕРЕ

А. Ю. Козлова
Научный руководитель А. Ю. Трусова
Самарский национальный исследовательский университет
имени академика С.П. Королева

Актуальность работы обосновывается тем, что формирование обучающих выборок является важным при классификации объектов по изучаемым социально-экономическим показателям, которые на данном историческом этапе являются наиболее важными для изучения.

Научная новизна исследования состоит в применении аппарата классификации с обучением и без.

Практическая значимость работы заключается в том, что, используя классификацию с обучением, всегда можно понять, как комплексно по всем показателям субъект может быть охарактеризован, описан и соотнесен с какими-то эталонными субъектами.

Целью работы является формирование обучающих выборок и проведение классификации с обучением на примере социально-экономических показателей.

Для достижения заданной цели необходимо решить следующий ряд задач:

1) изучить теорию классификацию с обучением и без;
2) сформировать две обучающие выборки;
3) провести дискриминантный анализ;
4) получить устойчивое разбиение на две группы.

Для анализа были взяты социально-экономические показатели субъектов Приволжского федерального округа (ПФО) (Таблица 1), где x_1 - среднегодовая численность занятых (тысяч человек); x_2 - нагрузка незанятого населения, состоящего на регистрационном учете в органах службы занятости населения (человек); x_3 - среднедушевые денежные доходы (в месяц; рублей); x_4 - потребительские расходы (в месяц; рублей); x_5 - среднемесячная номинальная начисленная заработная плата работников организаций (рублей); x_6 - степень износа основных фондов (%) [1].

Таблица 1. Исходный массив данных.

№	Субъект РФ	x_1	x_2	x_3	x_4	x_5	x_6
1	Республика Башкортостан	1 650,90	0,5	32 621	28 048	42 848	59,4
2	Республика Марий Эл	270,9	0,4	23 185	18 274	35 497	71,1
3	Республика Мордовия	374	0,5	22 906	17 823	34 499	66,9
4	Республика Татарстан	1 985,80	0,2	39 679	33 152	45 800	53,5
5	Удмуртская Республика	693,7	0,4	27 650	21 592	39 791	68,7
6	Чувашская Республика	494,1	0,4	23 619	19 484	35 799	67,7
7	Пермский край	1 153,50	0,6	32 747	27 570	46 267	64,5
8	Кировская область	557,4	0,5	26 649	22 267	36 143	55,0
9	Нижегородская область	1 637,90	0,2	37 524	31 065	41 369	58,7
10	Оренбургская область	853	1,0	26 518	22 375	38 357	65,7
11	Пензенская область	579,3	0,5	26 415	22 508	36 031	53,9
12	Самарская область	1 620,90	0,3	32 663	27 373	42 771	60,4
13	Саратовская область	1 024,40	0,3	26 228	21 770	37 408	60,2

14	Ульяновская область	541,1	0,3	26 849	21 788	36 126	58,1

С помощью метода «ближнего соседа» сформируем две обучающие выборки (таблица 2), а оставшиеся объекты будем классифицировать [2].

Таблица 2. Формирование обучающих выборок

№	x_1	x_2	x_3	x_4	x_5	x_6
1-ая группа						
2	270,9	0,4	23 185	18 274	35 497	71,1
3	374	0,5	22 906	17 823	34 499	66,9
6	494,1	0,4	23 619	19 484	35 799	67,7
8	557,4	0,5	26 649	22 267	36 143	55
14	541,1	0,3	26 849	21 788	36 126	58,1
2-ая группа						
1	1 650,90	0,5	32 621	28 048	42 848	59,4
7	1 153,50	0,6	32 747	27 570	46 267	64,5
9	1 637,90	0,2	37 524	31 065	41 369	58,7
12	1 620,90	0,3	32 663	27 373	42 771	60,4

Перейдём к дискриминантному анализу [3]. Найдём совместную ковариационную матрицу (таблица 3) для двух групп по формуле

(1) $S_* = \frac{1}{n_1+n_1-2}(S_1 + S_2)$,

где n_i – количество объектов в *i*-ой выборке, S_i – ковариационная матрица *i*-ой выборки.

Таблица 3. Совместная ковариационная матрица

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	33581,15	-14,23	188827,94	189698,31	-172869,47	-666,29
x_2	-14,23	0,02	-164,74	-113,43	125,37	0,15
x_3	188827,94	-164,74	4672894,28	3951206,91	-725322,38	-8548,90
x_4	189698,31	-113,43	3951206,91	3601738,40	-323501,26	-8465,39
x_5	-172869,47	125,37	-725322,38	-323501,26	2121193,65	673,89
x_6	-666,29	0,15	-8548,90	-8465,39	673,89	29,75

Далее вычислим вектор коэффициентов дискриминантной функции по формуле

(2) $A = S_*^{-1}(\overline{X}_1 - \overline{X}_2)$,

где S_*^{-1} – обратная матрица S_*, $\overline{X}_i$ – вектор средних i-ой группы.

Так, значения коэффициентов, следующие: $a_1 = -4{,}84$, $a_2 = -2849{,}47$, $a_3 = -0{,}82$, $a_4 = 0{,}83$, $a_5 = -0{,}36$, $a_6 = -84{,}93$.

Отсюда по формуле

(3) $f(x) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$,

получим такие значения дискриминантных функций двух групп: $f_1 = -25119{,}50$, $f_2 = -33171{,}52$.

И определим константу дискриминации по формуле

(4) $C = \frac{1}{2}(\overline{f_1} + \overline{f_2})$,

то есть $C = -29145{,}51$.

Теперь перейдём к классификации объектов. Для этого по аналогии найдём значения дискриминантной функции (таблица 4).

Таблица 4. Классификация объектов.

№	x_1	x_2	x_3	x_4	x_5	x_6	f
4	1 985,8	0,2	39 679	33 152	45 800	53,5	-58064605,97
5	693,7	0,4	27 650	21 592	39 791	68,7	-9973473,08
10	853,0	1,0	26 518	22 375	38 357	65,7	-16851073,20
11	579,3	0,5	26 415	22 508	36 031	53,9	-8190467,71
13	1 024,4	0,3	26 228	21 770	37 408	60,2	-14720044,63

Значения дискриминантной функции у всех объектов больше константы дискриминации, поэтому они относятся ко второй обучающей выборке, которая является для них эталонной. Сравнив средние значения каждого показателя классифицируемых объектов с эталонными, можно сделать следующие выводы:

1) среднегодовая численность занятых в 1,5 раза меньше;
2) нагрузка незанятого населения, состоящего на регистрационном учете в органах службы занятости населения в 1,2 раза больше;
3) среднедушевые денежные доходы в 1,16 раза меньше;
4) потребительские расходы в 1,17 раза меньше;
5) среднемесячная номинальная начисленная заработная плата работников организаций в 1,1 раза меньше;
6) степень износа основных фондов в 1,01 раза меньше.

Поэтому Республике Татарстан, Удмуртской Республике, Оренбургской области, Пензенской области и Саратовской области необходимо повысить среднегодовую численность занятых, среднедушевые денежные доходы, потребительские расходы, среднемесячную номинальную начисленную заработную плату работников организаций хотя бы до уровня второй выборки.

Таким образом, в ходе исследования был изучен математический аппарат теории классификации с обучением и без, сформированы две обучающие выборки, проведён дискриминантный анализ, получено разбиение на две группы и сформулированы рекомендации для развития субъектов.

Список использованных источников

1. Статистические издания [Электронный ресурс]. – URL: https://rosstat.gov.ru/folder/210/document/13204 (дата обращения 26.11.2023)
2. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман, Ульман Дж.: ДМК-Пресс, 2016. — 500 с.
3. Сошникова, Л.А. Многомерный статистический анализ в экономике / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер. Москва : ЮНИТИ, 1999. — 598 с.