

Дискримінантний аналіз

Дискримінантний аналіз призначений для з'ясування того, до якого з відомих класів потрібно віднести новий об'єкт, що характеризується певною множиною значень кількісних ознак. Передбачається, що для кожного класу вже складені характерні вибірки об'єктів (навчальні підсукупності). Ці вибірки зазвичай складаються кваліфікованими спеціалістами-експертами, які для діагностики (віднесення об'єкта до того чи іншого класу) можуть використовувати всі доступні їм знання, в тому числі й неформалізовані.

Проблема процедури класифікації. Класичний дискримінантний аналіз складається з формулювання деяких правил, критеріїв класифікації, «інформантів», «дискримінантних функцій», які дозволяють оцінити ймовірність належності нового об'єкта до кожного класу. Зазвичай у якості таких критеріїв класифікації вибирають відстань Махаланобіса від об'єкта до центрів кожного класу. За цим критерієм об'єкт відноситься до найближчого класу. Цю стандартну методику (ядро, основу дискримінантного аналізу) можна доповнити. По-перше, деякі класи (групи) об'єктів більш поширені, розповсюджені, ніж інші, й за однакових умов мають деякі переваги. Якщо ми володіємо апріорною інформацією про кількісне співвідношення (про обсяг) класів, тобто про ймовірності q_i належності об'єктів до кожного класу ($q_1 + q_2 + \dots + q_m = 1$), то цю додаткову інформацію слід врахувати у правилах класифікації (інформантах). По-друге, можна оцінити наслідки від можливих помилок класифікації. Якщо експерти складуть матрицю збитків c_{ij} (від неправильного віднесення об'єкта з класу i до класу j), то і цю додаткову інформацію можна врахувати в класифікаційних правилах, які тепер будуть оцінювати величину можливих збитків від віднесення об'єкта до того чи іншого класу. Об'єкт потрібно відносити до того класу, для якого ризик (величина можливих збитків) буде найменшим.

Теоретично можливо було б врахувати особливості багатовимірною розподілу об'єктів в кожному класі та прийняти замість відстані Махаланобіса

обґрунтованішу міру. Проте ця можливість залишається практично не реалізованою, оскільки з багатовимірних розподілів достатньо добре вивчений лише багатовимірний нормальний закон, що і визначає вибір в якості міри саме відстань Махаланобіса.

Відстань Махаланобіса є квадратичним інформантом – квадратичною функцією виміряних ознак об'єктів. Проте на практиці вводять додаткове спростовувальне припущення про рівність всіх коваріаційних матриць для кожного класу (класи тепер відрізняються тільки центрами групувань). Зазвичай навчальні підсукупності (характерні вибірки) надто малі, щоб для кожної вибірки достатньо надійно можна було визначити дисперсії і коваріації ознак – саме тому і доводиться використовувати загальну (об'єднану) коваріаційну матрицю. Вимушене прийняття припущення про рівність коваріаційних матриць для всіх класів істотно спростує (та полегшує) класифікаційні правила – тепер замість квадратичних інформантів одержали зручніші лінійні інформанти. Викладене вище в загальних рисах описує методи класичного дискримінантного аналізу.

Останнім часом класичний дискримінантний аналіз доповнений теорією «канонічних дискримінантних функцій» або просто «дискримінантних функцій». У зв'язку з цим пропонується за класифікаційними правилами залишити назву «інформанти» та не використовувати для них назву «дискримінантні функції», що тепер будуть позначати дещо інше.

Дискримінантними функціями за новим змістом є деякі узагальнюючі (агреговані) змінні, в просторі яких найчіткіше видно відмінності між класами (відомими підсукупностями). Ці узагальнюючі змінні складаються у вигляді лінійних комбінацій початкових ознак, а коефіцієнти цих лінійних комбінацій визначаються з умови максимуму певного функціонала, що характеризує відмінності між класами. Декілька перших (головних) дискримінантних функцій вичерпно описують практично всі відмінності між класами.

Попередній перехід до скороченого простору головних дискримінантних функцій усуває багато проблем класичного дискримінантного аналізу –

результати стають стійкішими й надійнішими. Справа в тому, що відстань Махаланобіса враховує кореляції між ознаками, що вважається позитивною властивістю даної міри. Але, з іншого боку, за високих кореляцій ця міра стає нестійкою, а за точної мультиколінеарності ознак – виродженою. Слід зазначити, що спеціалісти не зобов'язані перевіряти (та й не перевіряють) незалежність ознак у системі, що визначає об'єкт. У скороченому просторі (головних) дискримінантних функцій виродженість виключається, оскільки дискримінантні функції ортогональні. Крім того, часто для опису відмінностей між класами достатньо всього двох перших дискримінантних функцій і тоді з'являється можливість візуального відображення об'єктів кожного класу (і нового спірного об'єкта) на площині, що є надзвичайно цінним надбанням.

Теорія канонічних дискримінантних функцій дуже схожа за методами і цілями на теорію головних компонент факторного аналізу.

Проблема наочності (візуалізації). Впровадження канонічних дискримінантних функцій здійснюється через створення упорядкованих за значущістю функцій, аналогічних до головних компонентів факторного аналізу. Перша дискримінантна функція обирається так, щоб її середні значення для різних класів відрізнялись найістотніше. Друга дискримінантна функція добирається за ідентичних умов, але вона не повинна корелювати з першою. Третя функція не повинна корелювати з упровадженими раніше функціями. Тоді кілька перших дискримінантних функцій (головних) несуть усю інформацію про відмінність між класами.

Після завершення аналізу можна виконувати зворотний перехід у початковий простір реальних змінних.

Позначимо k – кількість класів; m – кількість дискримінантних ознак; h_i – кількість об'єктів (спостережень) класу p ; n – загальна кількість об'єктів усіх класів. У дискримінантному аналізі приймають наступні математичні допущення:

- 1) існує два чи більше класів $k \geq 2$;
- 2) у кожному класі маємо принаймні два об'єкти $h_p \geq 2$;

- 3) кількість дискримінантних ознак необмежена, але не перевищує загальну кількість об'єктів мінус два, $0 < m < (n - 2)$;
- 4) вимірювання дискримінантних ознак повинне бути здійснене за інтервальною шкалою;
- 5) дискримінантні ознаки лінійно незалежні;
- 6) передбачається приблизна рівність коваріаційних матриць для кожного класу (якщо не використовуються спеціальні прийоми);
- 7) приймається гіпотеза про багатовимірну нормальність закону розподілу об'єктів для кожного класу.

Проблема оцінки дискримінантних функцій та їх послідовного відбору. Дискримінантним методом передбачається обчислення таких числових характеристик:

1. Процентний вміст $\frac{\lambda_j}{\sum_{j=1}^m \lambda_j} \cdot 100\%$ показує, наскільки важливіша та чи

інша дискримінантна функція.

2. Квадрат коефіцієнта канонічної кореляції (кореляційного відношення)

$\eta_i^2 = \frac{\lambda_i}{1 + \lambda_i}$ показує, яка частка повної мінливості дискримінантної функції

пояснюється відмінністю груп.

3. Критерій Фішера $F = \lambda \frac{n - k}{k - 1}$, який можна порівняти з табличними

значеннями $F_{0,05}(k - 1, n - k)$ і $F_{0,01}(k - 1, n - k)$ та визначити, чи є значущою одержана дискримінантна функція.

4. Статистика Уїлкса. Слід розглянути її детально. Нехай $l = \min\{k - 1, m\}$

– загальна кількість дискримінантних функцій з ненульовими λ_i . Тоді

$$\Delta_0 = \frac{1}{1 + \lambda_1} \cdot \frac{1}{1 + \lambda_2} \cdot \frac{1}{1 + \lambda_3} \cdots \frac{1}{1 + \lambda_l}$$

є мірою остаточної мінливості, якщо врахувати всі дискримінантні функції; тобто Δ_0 оцінює дискримінантну здатність усієї системи функцій. Тепер потрібно оцінити дискримінантну здатність системи без першої, найважливішої функції:

$$\Delta_1 = \frac{1}{1 + \lambda_2} \cdot \frac{1}{1 + \lambda_3} \dots \frac{1}{1 + \lambda_l}.$$

Ця величина вже перевищує Δ_0 . Що ближче Δ_1 до одиниці, тим слабшою є дискримінантна здатність інших функцій системи, які можна враховувати.

Потім слід обчислити $\Delta_2 = \frac{1}{1 + \lambda_3} \dots \frac{1}{1 + \lambda_l}$, Δ_z , ... до Δ_{l-1} . Значущість послідовних значень Δ_j оцінюється за допомогою критерію Пірсона:

$$\chi_j^2 = - \left[n - \frac{m+k}{2} - 1 \right] \ln \Delta_j,$$

який слід порівнювати з табличними значеннями $\chi_{0,05}^2(v)$ і $\chi_{0,01}^2(v)$, де $v_j = (m-j)(k-j-1)$ – кількість степенів свободи.

На певному етапі система функцій, що залишилась, відкидається, і можна отримати систему інформативних показників – систему головних дискримінантних функцій.

Дискримінантні функції корисні для візуальної наочності розподілу об'єктів за класами, а також для одержання стабільних інформантів, які слід складати з дискримінантних функцій, а не вихідних ознак. Отже, переходимо до створення класифікаційних функцій (інформантів), за чисельними значеннями яких можна визначати, до якого класу найімовірніше віднести той чи інший об'єкт.

Приклад застосування дискримінантного аналізу.

Для визначення однорідності сукупності об'єктів рекомендовано

застосовувати кластерний аналіз за загальними складними ознаками СЕС, виміряними в кожному спостереженні. Для обчислень кластерного аналізу в ППП Statgraphics потрібно активізувати меню **Describe/Multivariate Methods/Cluster Analysis**. У результаті чого маємо розшарування підприємств на групи за системою загальних складних ознак, що відображає дендрограма на рис. 1.

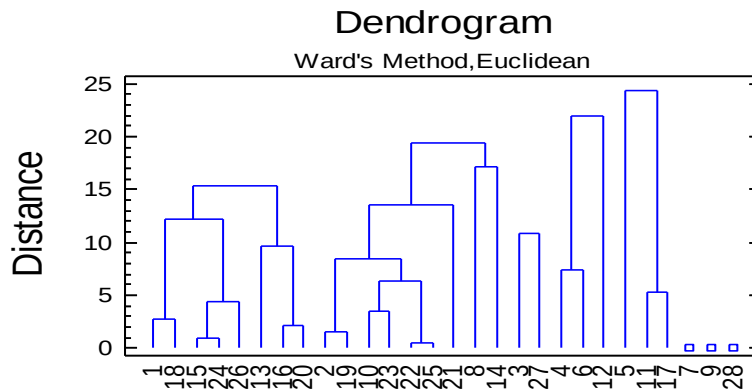


Рис. 1. Дендрограма кластеризації 28 промислових підприємств Харківського регіону за критерієм загальних латентних складних ознак виробничо-господарської діяльності у 2003 році

Будь-яку модель, що рекомендується до практичного застосування, необхідно дослідити на стійкість відносно коливань передумов та початкових даних. І далі – найрадикальнішим засобом перевірки стійкості розв’язків вважається побудова іншої моделі об’єкта. Якщо в новій моделі розв’язки виявляються близькими до розв’язків початкової, то найбільш вірогідно, що обидві вони достатньо адекватно відображають реальність, ніж у процесі розгляду тільки однієї моделі. На основі цієї думки був застосований дискримінантний аналіз для перевірки класифікації об’єктів за 8-ма групами, одержаної за допомогою кластерного аналізу в скороченому просторі загальних складних ознак. Для обчислень дискримінантного аналізу в ППП Statgraphics потрібно активізувати меню **Describe/Multivariate Methods/Discriminant Analysis**. Для досліджуваних 8 груп (класів) розраховані дискримінантні функції, наведені в табл. 3; в ній же наведені характеристики перших 7

функцій.

Таблиця 3

Результати обчислення дискримінантних функцій (ДФ)

і їх статистичний аналіз

№ ДФ	Узагальнені власні числа	Власні числа у процентах	Кореляційні відношення	Статистика Уїлкса	Критерій Пірсона	Число ступенів свободи	Рівень значущості α
1	106,134	61,03	0,99532	4,56328E-8	295,7962	77	0,0000
2	37,8139	21,74	0,98703	0,00000488	213,9998	60	0,0000
3	10,7827	6,20	0,95662	0,00018975	149,9712	45	0,0000
4	10,0363	5,77	0,95362	0,00223581	106,8051	32	0,0000
5	6,34244	3,65	0,92941	0,024675	64,7844	21	0,0000
6	1,86797	1,07	0,80704	0,181175	29,8952	12	0,0029
7	0,924548	0,53	0,69311	0,519602	11,4571	5	0,0430

Виокремлено сім (головних) дискримінантних функцій (ДФ), що впорядковані послідовно за значеннями узагальнених власних чисел. Ця послідовність збігається і з власними числами у відсотках від їх загальної суми та з даними в наступній колонці таблиці, де обчислені кореляційні відношення. При піднесенні кореляційних відношень до квадрату можна отримати індекси детермінації, які також підтверджують, що перша дискримінантна функція вагоміша для розділення сукупності об'єктів на класи. Статистики критерію Уїлкса демонструють відносний внесок перешкоди в загальну мінливість з урахуванням усіх функцій і при почерговому виключенні кожної наступної. За допомогою критерію Пірсона можна оцінити значущість статистик Уїлкса. Усі функції значущі, крім останньої, хоча для неї $\alpha > 0,01$, але її слід також вважати значущою, якщо прийняти рівень значущості $\alpha = 0,05$.

Таким чином, ознаковий 11-вимірний простір ознак скоротився до 7- вимірного, а можливо, і до 2-вимірного, оскільки перші 2 функції визначають загальний внесок (майже 83%). Коефіцієнти дискримінантних функцій обчислені у стандартизованій (табл. 4) і нестандартизованій формі.

Коефіцієнти в нестандартизованій формі несуть інформацію про абсолютний внесок даної складної ознаки в значення дискримінантної функції,

а коефіцієнти в стандартизованій формі – про відносний внесок ознаки. За даними табл. 4 визначається перша дискримінантна функція від стандартизованих змінних (складних ознак):

$$L_1 = -2,50Z_{F_1} - 0,80Z_{F_2} - 0,52Z_{F_3} - 0,09Z_{F_4} + 0,73Z_{F_5} + 2,26Z_{F_6} + 0,17Z_{F_7} - 0,84Z_{F_8} + 0,39Z_{F_9} - 0,22Z_{F_{10}} - 0,09Z_{F_{11}}.$$

Таблиця 4

Значення стандартизованих β - коефіцієнтів дискримінантних функцій

Ознаки (фактори)	Стандартизовані β - коефіцієнти						
F1	-2,50037	-0,407685	1,08752	0,526164	0,152259	1,86148	0,40472
F2	-0,795044	1,23193	0,720436	-1,01413	-0,0530502	0,214707	-0,036864
F3	-0,518632	1,41863	0,181346	0,82298	-0,268686	-0,066616	0,248815
F4	-0,0897036	-0,932923	-0,510947	-0,0129002	0,0475044	-1,28796	-0,432817
F5	0,722934	-0,229954	0,154992	-0,840602	-0,165299	-0,114081	0,861639
F6	2,25938	0,582309	-1,06868	0,540365	0,159043	-1,18266	-0,195323
F7	0,170045	-0,512609	1,13815	0,273807	0,479308	-0,498877	-0,140826
F8	-0,841862	-0,967958	-0,157215	0,408536	0,979613	0,266041	-0,244616
F9	0,392519	0,846252	0,500343	0,308441	-1,53069	-0,149921	0,10317
F10	-0,201835	-1,27124	0,733416	-0,305875	0,360367	-0,375734	0,223119
F11	-0,0875112	0,785047	0,020938	-0,148257	-0,675002	-0,288981	-0,404306

Таким же чином можна записати і всі 7 дискримінантних функцій. Порівнюючи внески кожної змінної F у формування дискримінантних функцій, слід зазначити, що перша функція утворена завдяки F_1 та F_6 ; друга – F_2 і F_3 , F_{10} ; третя – F_1 , F_6 , F_7 ; четверта – F_2 ; п'ята – F_9 і F_8 ; шоста – F_1 , F_4 , F_6 ; сьома – можна вважати, що це є F_5 . Таким чином, ознаки F_1 та F_6 є найважливішими у формуванні всіх 7 дискримінантних функцій, а отже, вони основні дискримінантні ознаки даної класифікації.

Таблиця 5

Коефіцієнти розкладу класифікаційних функцій

Змінні (фактори)	Групи об'єктів							
	1	2	3	4	5	6	7	8
F1	-4,1232	0,84100	9,1987	-1,3873	-2,0522	-6,14791	37,8466	-6,2356
F2	-2,6389	-0,7706	15,282	-1,5999	-5,9659	-6,58575	7,27225	0,34547
F3	-1,5873	0,00743	5,5177	-2,6632	-8,8959	-3,49504	2,16902	0,78797
F4	3,33387	6,99643	-8,368	15,1025	17,2596	5,84979	19,4026	1,10874
F5	1,16822	-1,0786	-1,749	0,43047	2,55944	1,56529	-7,5846	1,73652
F6	3,93159	-0,8955	-11,28	-0,6938	-1,5882	6,81424	-38,455	4,2031
F7	2,67939	-0,4132	-1,973	2,18639	6,30752	19,29	4,53485	1,39896
F8	-0,8482	2,47989	-0,032	3,20124	4,32736	0,160481	23,6132	-8,7713
F9	-0,0316	-1,7449	0,5804	-4,3748	-6,2920	0,320612	-12,331	11,969
F10	2,27373	1,07965	-2,507	6,8641	15,1353	11,7921	21,2033	-1,2756
F11	-1,3003	-0,0140	6,0165	-1,6775	-5,7874	-4,55526	-4,0508	5,64396
Const	-13,147	-13,010	-156,6	-35,182	-104,56	-156,187	-838,78	-79,641

Таким же чином можна записати і всі 7 дискримінантних функцій. Порівнюючи внески кожної змінної F у формування дискримінантних функцій, слід зазначити, що перша функція утворена завдяки F_1 та F_6 ; друга – F_2 і F_3 , F_{10} ; третя – F_1 , F_6 , F_7 ; четверта – F_2 ; п'ята – F_9 і F_8 ; шоста – F_1 , F_4 , F_6 ; сьома – можна вважати, що це є F_5 . Таким чином, ознаки F_1 та F_6 є найважливішими у формуванні всіх 7 дискримінантних функцій, а отже, вони основні дискримінантні ознаки даної класифікації.

Далі визначені класифікаційні функції, коефіцієнти яких наведені в табл. 3.13. Так, класифікаційна функція для першої групи об'єктів має вигляд:

$$y_1 = -13,15 - 4,12F_1 - 2,64F_2 - 1,59F_3 + 3,33F_4 + 1,17F_5 + 3,93F_6 + 2,68F_7 - 0,85F_8 - 0,03F_9 + 2,27F_{10} - 1,3F_{11}.$$

Аналогічно записуються всі інші класифікаційні функції – інформанти. Знайдені функції дозволяють побудувати класифікаційну таблицю об'єктів, яка показує ступінь підтвердження того початкового розшарування в сукупності, що отримано за допомогою кластерного аналізу. Склад кожного кластера відомий.

Результат зіставлення складу груп, отриманих за кластерним аналізом (КА), зі складом груп, одержаних за дискримінантним аналізом (ДА), відмічений процентом співпаданя (табл. 6).

Зміст табл. 6 переконливо свідчить, що наявна цілком стійка та достовірна класифікація об'єктів за системою загальних складних ознак, оскільки в кожній групі дискримінантний аналіз підтвердив склад груп на 100%. Отже, як безпосередня міра точності передбачення цей процентний зміст підтверджує правильність складання кластерів за методом Уорда. У результаті обчислень за допомогою дискримінантного аналізу отримано інформацію про окремі об'єкти, відмінності між класами, про здатність наявної системи загальних складних ознак як цілого точно відрізняти класи.

Можна продовжити аналіз вирішенням проблеми послідовного відбору дискримінантних ознак, але це доцільно робити, коли є система елементарних ознак об'єкта. У випадку коли дискримінантні ознаки – латентні складні ознаки, такий аналіз підтверджує значущість складних ознак, але з позиції їх відмінностей за об'єктами сукупності, тобто з метою класифікації.

Класифікаційна таблиця об'єктів

[illegible]