Розділ 5. Методи класифікації

До задач класифікації відносяться задачі, в яких необхідно поділити деякі об'єкти або явища на однорідні групи (класи) за наявності сукупності властивостей, що описують ці об'єкти (властивість не одна – їх багато).

Ці задачі поділяють на дві групи:

– визначення належності об'єкту до однієї з груп, які задані навчальними вибірками (дискримінантний аналіз). Це є класифікацією „з вчителем”;

– розбиття множини об'єктів на однорідні групи за відсутності навчальних вибірок (кластерний аналіз). Ця група ще носить назву автоматичної класифікації.

Дуже часто важливою супутньою задачею є визначення мінімальної інформативної підмножини змінних, що описують об'єкт, достатньої для поділу об'єктів на однорідні групи.

До задач класифікації відносяться, наприклад, встановлення діагнозу за результатами аналізу і обстеження (дискримінантний аналіз), а також – визначення груп хворих, які можна вважати однорідними за сукупністю їх індивідуальних особливостей (включаючи перебіг хвороби і процес лікування), що дозволяє краще підбирати курс лікування (кластерний аналіз).

На сьогодні для розв’язання задач класифікації (в сенсі дискримінації) використовують штучні нейронні мережі. Але їх використання потребує достатньо великої за об’ємом навчальної вибірки.

5.1. Кластерний аналіз (КА)

5.1.1. Теоретичні відомості

Для поділу вибірки на однорідні підвибірки використовуються методи кластерного аналізу. КА – це сукупність багатовимірних статистичних процедур, яка дозволяє упорядкувати об'єкти за однорідними групами.

Призначення.

За допомогою кластерного аналізу досліджувану сукупність об'єктів, що представлена у вигляді матриці „об'єкти-властивості”, розбивають на невелику кількість однорідних груп (їх кількість наперед може бути відома або невідома). Навчальної вибірки в КА не існує.

Матриця „об'єкти-властивості” має такий вигляд:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix},$$

де x_{ij} – значення j-ї властивості для об'єкту за номером i. Тобто маємо n об'єктів і m властивостей, що описують ці об'єкти.

Класичним прикладом такого поділу на кластери є якісно різна ефективність лікарського препарату (одужання або погіршення стану) залежно від статі хворого. Розрив простору існування фактора може виникати також при певній комбінації деякої сукупності незалежних змінних. Проте побудувати якісну і таку, що має фізичний сенс, модель для розривної області не можливо за об'єктивними причинами. Потрібно знайти фактор або комбінацію факторів, що відповідають за розрив, і побудувати моделі для кожної з виділених підобластей окремо. Тобто одна модель описує вплив препарату на організм чоловіка, інша – на організм жінки і тому подібне.

На рис. 5.1 показана діаграма розсіювання, на якій видно, що дані розпадаються на дві розділені в просторі підвибірки (кластери), що непов'язані між собою.

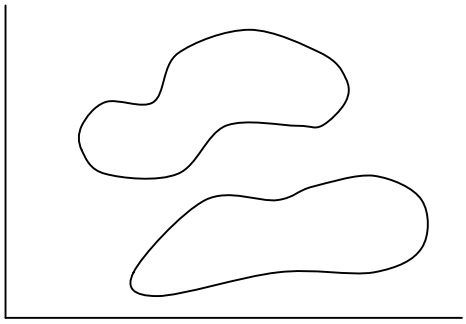

Рис. 5.1. Приклад структуризації даних за кластерами.

Загальна схема розв’язання задачі кластерного аналізу даних

Розв’язання задачі КА виконується в такій послідовності:

– формування вибірки для аналізу;
– вибір сукупності ознак, що характеризують об'єкти;
– вибір міри схожості (відстані) між об'єктами та їх розрахунок;
– формування кластерів;
– аналіз отриманих результатів.

Більшість алгоритмів КА відносять до так званих агломеративних процедур, за допомогою яких спочатку об'єднують в групи найближчі об'єкти, а потім до них приєднують більш далекі.

Можливі міри схожості (відстані між об'єктами)

Для визначення схожості між об'єктами користуються поняттям відстані $d_{ij}(O_i,O_j)$ між об'єктами O_i і O_j. Чим менше відстань між ними, тим більше схожими вважаються об'єкти. Щоб бути придатною для визначення відстані, пропонована міра повинна володіти такими властивостями:

– симетрією – $d_{ij}(O_i,O_j) = d_{ji}(O_j,O_i)$;
– мінімальною відстанню об'єкту до самого себе – $d_{ii}(O_i,O_i)$=0;

– монотонною зміною d_{ij} в описуваному просторі;

– змістовною інтерпретацією міри (бажано).

Найчастіше в КА використовують міри, що базуються на узагальненій відстані Махаланобіса, яку задають формулою:

$$d_{ij} = \sqrt{\left(X_i - X_j\right)^T \Lambda^T \Sigma^{-1} \left(X_i - X_j\right)}, \qquad (5.1)$$

де X_i – i-й вектор спостережень; Λ – симетрична невід'ємно визначена матриця вагових коефіцієнтів (зазвичай діагональна); Σ – коваріаційна матриця сукупності, з якої вибрані спостереження.

Реально використовуються такі часткові види відстаней:

1. Евклідова відстань. Ця відстань використовується при виконанні наступних умов:

– компоненти X взаємно незалежні, мають одну і ту ж дисперсію;

– компоненти є однорідними за фізичним змістом.

Якщо різні властивості, що характеризують об'єкт, мають різну важливість і її можна оцінити (хоча би за допомогою експертів) використовують взважену евклідову відстань:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \omega_k \left(x_{ik} - x_{jk}\right)^2}, \qquad (5.3)$$

де ω_k – ваговий коефіцієнт для k-ї властивості. При цьому зазвичай приймається $0 \le \omega_k \le 1$ для всіх k.

2. Хеммінгова відстань. Інколи його називають відстанню міських кварталів (тобто шлях від перехрестя до перехрестя не безпосередньо, а тільки по вулицях). Її визначають за формулою:

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|; \qquad (5.4)$$

3. Відстань між класами. Наведені вище міри визначають відстань між об'єктами. Для виконання кластерного аналізу необхідно встановити, що вважати за відстань між кластерами. Зазвичай використовують такі міри близькості груп (відстаней між кластерами):

– відстань, що обчислюється за принципом „найближчого сусіда". Це є мінімальною відстанню між парою об'єктів, кожен з яких знаходиться в іншому кластері. Його обчислюють за формулою:

$$d(S_l, S_k) = \min d(X_i, X_j), \quad \text{де } X_i \in S_l,\ X_j \in S_k; \qquad (5.5)$$

– відстань, що обчислюється за принципом „далекого сусіда". Це є максимальною відстанню між парою об'єктів, кожен з яких знаходиться в іншому кластері. Його обчислюють за формулою:

$$d(S_l, S_k) = \max d(X_i, X_j), \quad \text{де } X_i \in S_l,\ X_j \in S_k; \qquad (5.6)$$

– відстань, що обчислюється за „центрами тяжіння” кластерів за формулою:

$$d(S_l, S_k) = d\left(\overline{X}(l), \overline{X}(k)\right), \tag{5.7}$$

де $\overline{X}(l)$ – середнє арифметичне векторних спостережень, які входять в кластер S_l. Таким чином, це відстань між „центрами тяжіння” відповідних кластерів;

– відстань, що обчислюється за принципом „середнього зв'язку”. Це є арифметичне середнє всіх можливих пар комбінацій між об'єктами, що входять в різні кластери. Ця відстань обчислюється за формулою:

$$d(S_l, S_k) = \frac{1}{n_l n_k} \sum\sum d\left(X_i, X_j\right),\ X_i \in S_l,\ X_j \in S_k. \tag{5.8}$$

Якість розбиття на класи

Існує досить велика кількість різних процедур КА. Для порівняння якості розбиття на класи використовується ряд функціоналів якості. Найбільш вживаними з них є:

– сума внутрішньокласових дисперсій відстаней, що визначається за формулою:

$$Q = \sum_{k=1}^{p} \sum d(X_i, \overline{X}_k),\ X_j \in S_k \tag{5.9}$$

де p – кількість кластерів;

– сума попарних внутрішньокласових відстаней, що визначається за формулою :

$$Q = \sum_{k=1}^{p} \sum_{X_j \in S_k} d(X_j, X_j) \tag{5.10}$$

де p – кількість кластерів.

Примітка. Мета КА – пошук існуючих реальних структур даних. Різні процедури КА для одних і тих же даних можуть давати різне розбиття на кластери (як за їх кількістю, так і за складом). Більшість методів КА не мають суворо статистичного обґрунтування.

5.1.2. Опис використовуваного алгоритму КА

Розглянемо процедуру КА, що запропоновано для обробки даних. Існують два різновиди, які можуть давати різне розбиття на кластери. Вибрати найбільш відповідний різновид треба виходячи із постановки задачі. Якщо це неможливо, необхідно провести розбиття двома способами і спробувати визначити, який з них більше відповідає фактично існуючим структурам даних.

При ізотонічному розбитті групи об'єктів складаються з однорідних за рівнем значень.

При ізоморфному розбитті в групи включаються об'єкти, що є близькими за структурою, тобто ті, в яких пропорції ознак мало відрізняються.

Це означає, що різноманітні способи розбиття можуть давати різне розбиття за групами. Наприклад, є дані, які характеризують розподіл прибуток фірм на розширення виробництва, наукові дослідження, соціальні виплати тощо. Тоді при ізотонічному розбитті групи будуть складатись із фірм, в яких рівні прибутку є близькими, а при ізоморфному – в однорідні групи будуть включатись ті компанії, в яких структура розподілу прибутків подібна.

В обох випадках ознаки спочатку перетворюють таким чином, щоб не було одиниць виміру і розмах шкали був однаковим.

Ізотонічне розбиття

Для нормування шкал тут необхідно виконати такі перетворення.

Спочатку кожне значення ознаки замінюється на обчислене за формулою:

$$V_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \tag{5.11}$$

де x_{ij} – значення j-ї ознаки для i-го об'єкту.

Після цього кожному об'єкту ставиться у відповідність одне число, що обчислене за формулою:

$$\omega_i = \sum_{j=1}^{m} V_{ij} \ . \tag{5.12}$$

Відстань між двома об'єктами визначають за формулою:

$$d_{ij} = \left| \omega_i - \omega_j \right| . \tag{5.13}$$

Ізоморфне розбиття

Спочатку виконують нормування шкал за формулою:

$$Z_{ij} = \frac{\dfrac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}}{\sum_{j=1}^{m} \dfrac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}}, \tag{5.14}$$

де x_{ij} – значення j-ї ознаки для i-го об'єкту.

Відстань між двома об'єктами визначають за формулою:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(Z_{ij} - Z_{ik}\right)^2} \qquad (5.15)$$

У ізоморфному перетворенні відстань буде мінімальною в тому випадку, якщо вектори колініарні, і максимальною – якщо вони перпендикулярні.

Розбиття на кластери

Після визначення відстаней можливе розбиття на групи за допомогою методу куль. Потім, побудувавши дендриты, можна визначити форму сліду даних. Під слідом розуміють просторову форму, яку приймає сукупність експериментальних точок. У методі куль критичний радіус (відстань, яка визначає чи належить об'єкт даному кластеру) обчислюють за формулою:

$$r = \max_i \min_j d_{ij} \qquad (5.16)$$

Тобто, по кожному об'єкту вибирають мінімальну відстань до найближчого до нього об'єкту. Потім з цих мінімальних відстаней вибирають максимальну. Тоді об'єкти, відстані між якими менше критичної, належать до одного кластера. На рис.5.2 показано розбиття на кластери з використання методів куль. В один кластер включають об'єкти, відстань між якими менше критичної. Недолік методу полягає в тому, що кластери отримують дещо штучними (кулястої форми).

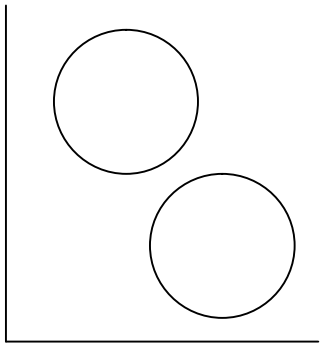

Рис.5.2. Розбиття на кластери
з використанням методу куль.

Побудова дендритів і визначення зв'язності

Первинне розбиття на кластери дозволяє отримати кластери кулястої або еліпсовидної форми. Оскільки при виконанні практичних задач такі кластери зустрічаються не завжди (бувають випадки, коли слід має С-, S-образну і складнішу форму, то на наступному етапі звичай виконують побудову дендритів і визначення зв'язності в системі кластерів. Це дозволяє об'єднати

первинні кластери в складніші структури, які в більшій мірі відповідають їх реальній формі.

На цьому етапі визначають, чи є виділені кластери повністю незв'язаними або утворюють яку-небудь зв'язну структуру.

Для цього визначають відстань між кластерами, яка дорівнює мінімальній відстані між двома об'єктами, що входять в ці кластери:

$$C_{lk} = \min_{p \in g_l} \min_{q \in g_k} C_{lk}(p,q) . \qquad (5.17)$$

Критична відстань – це відстань, при перевищенні якої кластери вважаються за незв'язні. Критичну відстань приймають рівною максимальній відстані між двома сусідніми елементами в одному кластері. Його розраховують за формулою:

$$C_l(p) = \frac{1}{k} \sum_{l=1}^{G} \sum_{p=1}^{P_l} C_l(p) \qquad (5.18)$$

де $C_l(p) = \min_{q \in g_l} C_{ll}(p,q)$, p=1,2,...,P_l, $k = \sum_{l=1}^{G} P_l$,

$C_{ll}(p,q)$ – відстань між елементами p і q, що належать до l-ї групи (кластера);

$C_l(p)$ – відстань від елемента p до сусіднього елемента в групі l;

P_l – кількість елементів в l-ій групі;

G – кількість груп.

Після визначення відстаней кластери послідовно об’єднують в групи таким чином, щоб зрештою отримати дендрит, який об'єднує всю сукупність даних (кожен кластер зв'язується з найближчим до нього). Зазвичай отриманий дендрит має форму ланцюжка, але можливі і більш складні форми: „дерево”, „розетка”, „амеба” тощо.

Після побудови дендриту, що об'єднує всі дані, на основі аналізу критичної відстані розривають зв'язки між кластерами, відстань між якими більше критичного. В результаті отримують набір дендритів, які незв'язані між собою, але кластери, що входять в кожен дендрит, утворюють зв'язану підвибірку. Дані, що входять в таку підвибірку, можна апроксимувати за допомогою регресійного аналізу.

Якщо заключний зв'язковий дендрит має форму „розетка” або „амеба” (рис. 5.3), його треба поділити на більш прості, оскільки отримати в такій області регресійну модель складно. Зазвичай в таких випадках маємо справу не з одним слідом, а з перетином декількох різних слідів. Область перетину у них спільна, а в іншому – це окремі області простору (рис. 5.4).

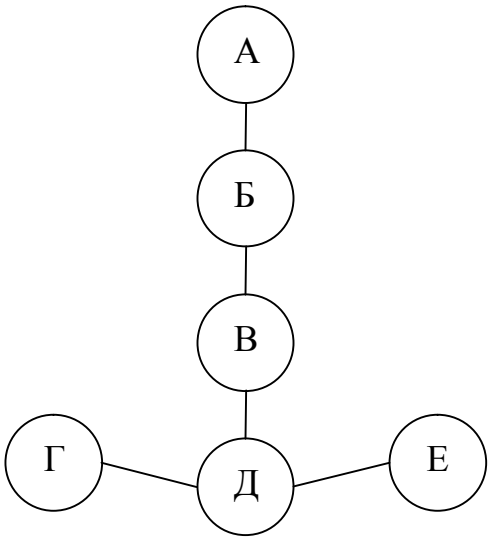

Рис. 5.3. Типовий вид дендриту (числа – відстань між відповідними кластерами).

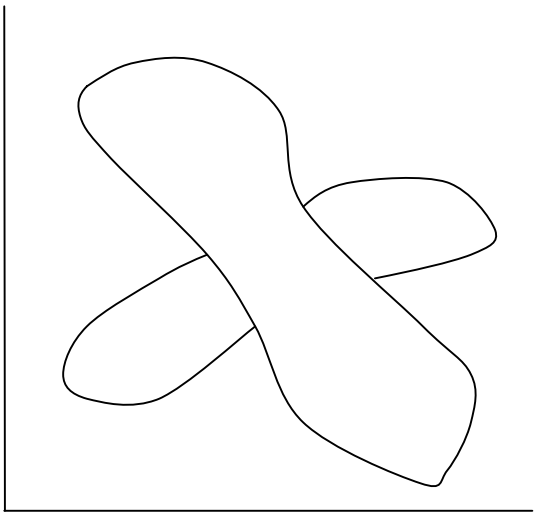

Рис.5.4. Приклад кластерів, що перетинаються.

5.2. Дискримінантний аналіз (ДА)

Теоретичні відомості

Дискримінантний аналіз дозволяє вивчати відмінності між двома і більш групами об'єктів за декількома змінними одночасно. Кожна група об'єктів називається класом. Класи розглядуються як значення деякої класифікуючої змінної, виміряної за шкалою найменувань.

Якщо класифікуюча змінна залежить від дискримінантних змінних, то в цьому випадку дискримінантний аналіз є аналогом багатофакторного регресійного аналізу, коли відгук вимірюється за шкалою найменувань. Наприклад, дискримінантні змінні описують стан хворого і метод лікування, а змінна, що класифікує показник виду: видужав – не видужав.

У випадку, коли дискримінантні змінні залежать від класифікуючої змінної, тобто приналежність об'єкту до певного класу викликає зміни одночасно в декількох змінних, дискримінантний аналіз є аналогом узагальненого багатовимірного дисперсійного аналізу. Наприклад, класифікуюча змінна – захворювання, а дискримінантні змінні, на підставі яких потрібно поставити діагноз, характеризують стан хворого.

В дискримінантному аналізі розрізняють дві задачі: інтерпретації і класифікації.

Задача інтерпретації – визначення кількості, значущості дискримінантних функцій і їх значень для пояснення відмінностей між класами.

Задача класифікації – визначення класу, до якого належить новий об'єкт.

В дискримінантному аналізі, на відміну від кластерного аналізу, є навчальна вибірка, в якій відомо, до яких класів відносяться об'єкти. За

навчальною вибіркою треба побудувати правила, які надалі дозволять визначати, до якого класу відносяться нові об'єкти.

Призначення.

Вивчення відмінностей між двома і більш групами об'єктів одночасно за декількома описовими змінними. ДА дозволяє за навчальною вибіркою отримати правила (формули), за якими визначають належність об'єкту до певної групи.

Передумови.

1. Об'єкти (спостереження) належать до двох або більшої кількості класів.
2. У кожному класі є як мінімум два об'єкти.
3. Кількість дискримінантних змінних не має бути більше кількості об'єктів мінус два.
4. Дискримінантні змінні вимірюють в шкалі інтервалів або шкалі відношень.
5. Дискримінантні змінні мають бути лінійно незалежні.
6. Дискримінантні змінні мають бути розподілені за багатовимірним нормальним законом розподілу (тобто кожна змінна має нормальний закон розподілу при фіксованих решти змінних).
7. Коваріаційні матриці класів приблизно однакові між собою.

Розв’язання задачі ДА для випадку двох класів

Найбільш поширеним (простим для розв’язання), є випадок двох класів.

Є два класи, об'єкти характеризуються *p* змінними. Для першого класу сформована вибірка *X*, об'ємом n_1, для другого – *Y* об'ємом n_2.

1. Розраховують середні значення по кожній змінній для кожної вибірки (класу):

$$\bar{\bar{x}}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij} \quad \text{і} \quad \bar{\bar{y}}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij} .$$

2. Визначають оцінки коваріаційних матриць для кожного класу S_x і S_y:

$$S_{kj}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad \text{і} \quad S_{kj}(y) = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_{ij} - \bar{y}_j)(y_{ik} - _k) .$$

3. Розраховують незміщену оцінку об'єднаної коваріаційної матриці:

$$S = \frac{1}{n_1 + n_2 - 2}(n_1 S_x + n_2 S_y) . \qquad (5.19)$$

4. Знаходять матрицю S^{-1}, що є оберненою *S*.

5. Розраховують вектор оцінок коефіцієнтів дискримінантної функції $A = S^{-1}(\overline{X} - \overline{Y})$ /

6. Визначають оцінки векторів дискримінантних функцій для початкових змінних $U_x=XA$ і $U_y=YA$.

7. Обчислюють середні значення оцінок дискримінантних функцій:

$$\overline{\overline{u}}_x = \frac{1}{n_1}\sum_{i=1}^{n_1} u_{xi} \text{ і } \overline{\overline{u}}_y = \frac{1}{n_2}\sum_{i=1}^{n_2} u_{yi} .$$

8. Визначають дискримінантну константу

$$C = \frac{1}{2}(\overline{\overline{u}}_x + \overline{\overline{u}}_y) .$$

Для того, щоб визначити, до якого класу належить яке-небудь спостереження Z (об'єкт), необхідно спочатку обчислити для нього оцінку дискримінантної функції: $u_z = \sum_{i=1}^{p} a_i z_i$. Якщо це значення більше або дорівнює константі C, то новий об'єкт відноситься до класу X, якщо менше – до класу Y (при $\overline{\overline{u}}_x > \overline{\overline{u}}_y$).

Послідовність розв'язання задачі ДА для загального випадку k класів

Кінцева дискримінантна функція в цьому випадку має такий вигляд:

$$f_{ki} = u_0 + \sum_{j=1}^{p} u_j X_{jki} \qquad (5.20)$$

де f_{ki} – значення канонічної дискримінантної функції для i-го об'єкту в k-му класі; u_j – шукані коефіцієнти дискримінантної функції; X_{jki} – значення дискримінантної змінної X_j для i-го об'єкту в класі k. Функцію будують так, щоб її середні значення для різних класів якомога більше відрізнялися. При цьому сукупність функцій повинна утворювати ортогональний простір, тобто функції повинні бути незалежними один від одного. З цього виходить, що кількість функцій не може бути більше кількості класів мінус 1 або кількості дискримінантних змінних (залежно від того, яка з цих величин менша).

Задачу вирішують у декілька етапів.

1. Побудова матриці T, елементи якої визначають за формулою:

$$t_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{jki} - \overline{\overline{X}}_j\right)\left(X_{jki} - \overline{\overline{X}}_i\right), \qquad (5.21)$$

де X_{jki} – значення дискримінантної змінної X_j для i-го об'єкту в класі k; $\overline{\overline{X}}_j$ – середнє значення для змінної j за всіма класами; n – загальна кількість спостережень за всіма класами; K – кількість класів.

2. Обчислення матриці W, яка визначає ступінь розкиду всередині класів. Елементи цієї матриці знаходять за формулою:

$$W_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{jki} - \overline{\overline{X}}_{jk} \right)\left(X_{jki} - \overline{\overline{X}}_{ik} \right), \qquad (5.22)$$

де $\overline{\overline{X}}_j$ – середнє значення для змінної j в k-ому класі. Решта позначень аналогічна позначенням з попередньої формули.

3. Обчислення матриці $B=T-W$, елементи якої визначають як

$$b_{jl}=t_{jl}-W_{jl}.$$

4. Розв'язання системи рівнянь:

$$\begin{aligned} &\sum_{j=1}^{p} b_{1j}v_j = \lambda\sum_{j=1}^{p} w_{1j}v_j \\ &\sum_{j=1}^{p} b_{2j}v_j = \lambda\sum_{j=1}^{p} w_{2j}v_j \\ &\text{.....} \\ &\sum b_{pj}v_j = \lambda\sum_{j=1}^{p} w_{pj}v_j \end{aligned}, \qquad (5.23)$$

де λ – власне значення (чим воно більше, тим більше груп розділятиме відповідна дискримінантна функція). Побудована на його базі канонічна кореляція:

$$r_i^{\bullet} = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

відображає ступінь залежності між дискримінантною функцією і класами, а квадрат кореляції показує частку дисперсії дискримінантної функції, яка пояснюється розподілом на класи. Для отримання єдиного рішення вводять додаткове обмеження: $\sum_{j=1}^{p} v_j^2 = 1$.

5. Отримані коефіцієнти нормують за формулою:

$$u_j = v_j\sqrt{n-K}\ .$$

При цьому $\overline{\overline{x}}_i = -\sum_{j=1}^{p} u_j \overline{\overline{X}}_j$.

Коефіцієнти u_j називають нестандартними, оскільки залежать від одиниць виміру змінних, тому часто переходять до стандартних коефіцієнтів, які відображають відносний вклад змінної, що незалежить від шкали виміру. Перехід до стандартних коефіцієнтів обчислюють за формулою:

$$c_i = u_i\sqrt{\frac{W_{ii}}{n-K}}, \qquad (5.24)$$

де n – загальна кількість спостережень; K – кількість класів (груп); W_{ii} – діагональний елемент матриці оцінки розсіювання, обчисленої за формулою (5.22).

Величина стандартного коефіцієнта пропорційна його вкладу у дискримінантну функцію. При цьому слід мати на увазі, що якщо змінні закорельовані, стандартні коефіцієнти не відображатимуть дійсного вкладу.

Кількість дискримінантних функцій

Задачу дискримінації бажано вирішувати з використанням мінімальної кількості функцій. Кількість функцій у кожному конкретному випадку залежить від конфігурації класів в багатовимірному просторі дискримінантних змінних. Чим складніше конфігурація, тим більше функцій необхідно для їх поділу. Щоб визначити, скільки необхідно функцій, використовують перевірку їх на значущість. Для характеристики, наскільки одна функція слабкіша за іншу, використовують відносний процентний вміст, який визначають для однієї функції за формулою:

$$\frac{\lambda_i}{\sum \lambda_i} 100\% .$$

Але цей показник не може служити підставою для відкидання частини функцій. Для оцінки значущості використовують Λ –статистику Уїлкса. Критеріальне значення обчислюють за формулою:

$$\Lambda_k = \prod_{i=k+1}^{K} \frac{1}{1+\lambda_i}, \qquad (5.25)$$

де K – кількість класів, а k – кількість вже обчислених дискримінантних функцій. Чим ближче значення Λ до 0, тим краща відмінність класів. А чим ближче Λ до 1, тим відмінність гірша (класи співпадають).

Можлива перевірка значущості за критерієм χ^2 з використанням статистики Уїлкса. Для цього необхідно розрахувати критеріальне значення за формулою:

$$\chi^2 = -\left[n - \frac{p+K}{2} \right] \ln \Lambda_k . \qquad (5.26)$$

Якщо це значення більше критичного χ^2 із заданим рівнем значущості і числом степенів свободи $(p\text{-}k)(K\text{-}k\text{-}1)$, то значущість підтверджується.

Класифікація при кількості груп більше двох

Розраховані значення канонічної дискримінантної функції f_{ki} (див. формулу (5.20)) розглядають як точки в деякому просторі. Для кожної групи можна розрахувати центр групування (середнє). Тому в цій новій системі координат для нового об'єкту розраховують відстань від нього до кожної точки групування. Зазвичай для цього використовують квадрат відстані Махаланобіса $(D^2(X,G_k)$ – відстань від об'єкту X, який необхідно класифікувати, до центру класу G_k). Об'єкт зараховують до групи, відстань до якої $(D^2(X, G_k))$ найменша.

Покроковий дискримінантний аналіз

Він дозволяє здійснити послідовний відбір дискримінантних змінних з метою формування дискримінантних функцій з мінімальною кількістю аргументів при забезпеченні надійної класифікації.

Ймовірність (вірогідність) належності до класу

Зазвичай передбачається однакова апріорна ймовірність належності об'єкту до певного класу. Бувають ситуації, в яких класи за всією природою мають різну кількість елементів. Наприклад, пацієнти з деякими специфічними особливостями конкретного захворювання складають тільки 5% від загальної кількості. В таких випадках для правильного розв'язання задачі класифікації необхідне врахування цієї ймовірності. Тому відстань D^2 модифікують таким чином:

$$D'^2(X,G_k)=D^2(X,G_k)-2lnP_{apriori,k}, \tag{5.27}$$

де $P_{apriori,k}$ – апріорна ймовірність належності об'єкту до групи k.

Класифікація без інтерпретації

При використанні дискримінантних функцій, окрім задачі класифікації, вирішують і задачу інтерпретації. В деяких випадках немає необхідності вирішувати задачу інтерпретації. В цьому випадку використовують прості класифікаційні функції (в ряді робіт саме їх називають дискримінантами, а представлені раніше – канонічними дискримінантними функціями), що основані безпосередньо на дискримінантних змінних:

$$h_k = b_{k0} + \sum_{i=1}^{p} b_{ki} x_i \,, \tag{5.28}$$

де h_k – значення функції для k-го класу, x_i – значення дискримінантних змінних.

Значення b_k визначають за формулами:

$$b_{ki} = (n-K)\sum_{i=1}^{p} a_{ij} X_{jk} \,, \tag{5.29}$$

$$b_{k0} = -0.5\sum_{i=1}^{p} b_{ki} X_{jk} \,, \tag{5.30}$$

де a_{ij} – елемент матриці, що є оберненою W, яку розраховують за формулою (5.22).

Для класифікації використовують квадрат відстані Махаланобіса, який обчислюють таким чином:

$$D^2(X,G_k) = (n-K)\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}\left(X_i - \overline{\overline{X}}_{ik}\right)\left(X_j - \overline{\overline{X}}_{jk}\right). \tag{5.31}$$

В. Плюта рекомендує використовувати скореговану незміщену оцінку цієї величини:

$$D'^2 = \frac{n-p-3}{n-2} D^2 - \left(\sum_{i=1}^{K} \frac{p}{n_1}\right). \tag{5.32}$$

Об'єкт X зараховують до групи, відстань до якої є $(D^2(X,G_k))$ найменшою. Врахування апріорної ймовірності виконують за формулою (5.27).

Ймовірність (вірогідність) належності до класу

В випадку, коли має місце значне перекриття класів і, отже, слабка їх відмінність, бажано окрім відстані розраховувати ще і ймовірність належності до класу, яку обчислюють за формулою:

$$p = \frac{1}{\sum_{k=1}^{K} e^{(f_{kx} - f_{\max})}}, \qquad (5.33)$$

де f_{kx} – значення дискримінантної функції для об'єкту X в k -ому класі, f_{max} – значення дискримінантної функції для об'єкту X для класу, відстань до якого мінімальна. Якщо ймовірність мала, то класифікувати об'єкт при даному способі розподілу не можна.