Тема 10. Методи кластерного аналізу. Ієрархічні методи

План

1. Кластерний аналіз.
2. Методи кластерного аналізу.
3. Ієрархічний кластерний аналіз.

Мета вивчення теми: вивчити методи кластерного аналізу; засвоїти особливості ієрархічних методів.

Перелік ключових слів та понять із теми

Кластерний аналіз, кластер, ієрархічні методи, дивизимні методи, евклідова відстань, дентдрограма

Теоретичні відомості з теми

1. Кластерний аналіз

Поняття кластеризації розглянуто в п’ятій темі курсу. У цій лекції опишемо поняття «кластер» із математичної точки зору, а також розглянемо методи розв’язання задач кластеризації – методи кластерного аналізу.

Термін кластерний аналіз, уперше введений Тріоном (Tryon) у 1939 році, містить більш 100 різних алгоритмів.

На відміну від задач класифікації, кластерний аналіз не вимагає апріорних припущень про набір даних, не накладає обмеження на показ досліджуваних об’єктів, дозволяє аналізувати показники різних типів даних (інтервальні дані, частоти, бінарні дані). При цьому необхідно пам’ятати, що змінні повинні вимірюватися в порівнюваних шкалах.

Кластерний аналіз дозволяє скорочувати розмірність даних, робити їх наглядними.

Кластерний аналіз може застосовуватися до сукупностей тимчасових рядів, тут можуть виділятися періоди схожості деяких показників і визначатися групи тимчасових рядів зі схожою динамікою.

Кластерний аналіз паралельно розбудовувався в декількох напрямках, таких як біологія, психологія, ін., тому більшість методів мають по дві й більш назв. Це суттєво ускладнює роботу при використанні кластерного аналізу.

Задачі кластерного аналізу можна об’єднати в такі групи:

1. Розробка типології або класифікації.
2. Дослідження корисних концептуальних схем групування об’єктів.
3. Представлення гіпотез на основі дослідження даних.
4. Перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

Як правило, при практичному використанні кластерного аналізу одночасно розв’язуються декілька із зазначених задач.

Розглянемо приклад процедури кластерного аналізу.

Допустимо, маємо набір даних А, що складається з 14-ти прикладів, у яких є по дві ознаки *X* і *Y*. Дані в табличній формі не носять інформативний характер. Представимо змінні *X* і *Y* у вигляді діаграми розсіювання (рис. 10.1).

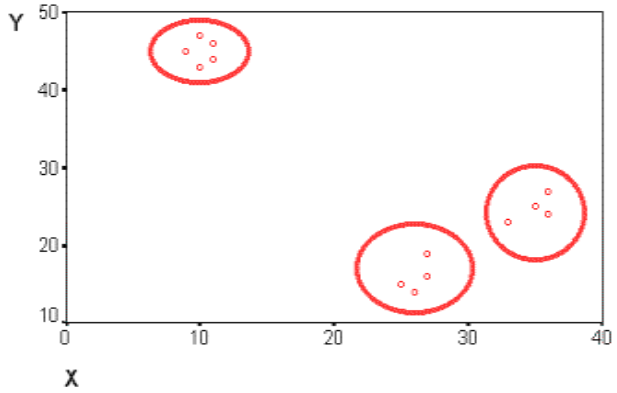

Рисунок 10.1 – Діаграма розсіювання змінних X і Y

На рисунку бачимо кілька груп «схожих» прикладів. Приклади (об'єкти), які за значеннями *X* і *Y* «схожі» один на одного, належать до однієї групи (кластеру); об'єкти з різних кластерів не схожі один на одного.

Критерієм для визначення схожості й відмінності кластерів є відстань між точками на діаграмі розсіювання. Цю подібність можна «виміряти», вона дорівнює відстані між точками на графіку. Способів визначення міри відстані між кластерами, яку називають ще мірою близькості, існує небагато. Найпоширеніший спосіб – обчислення евклідової відстані між двома точками *i* та *j* на площині, коли відомі їхні координати *X* і *Y*:

$$D_{ij} = \sqrt{\left(\left[x_i - x_j\right]\right]^2 + \left(\left[y_i - y_j\right]\right]^2}, \qquad (10.1)$$

Якщо нам потрібно знайти відстань між двома точками в просторі трьох вимірів (рис. 10.2), формула (10.1) набуває вигляду:

$$D = \sqrt{\left(\left[x_1 - x_2\right]\right]^2 + \left(\left[y_1 - y_2\right]\right]^2 + \left(\left[z_1 - z_2\right]\right]^2}, \qquad (10.2)$$

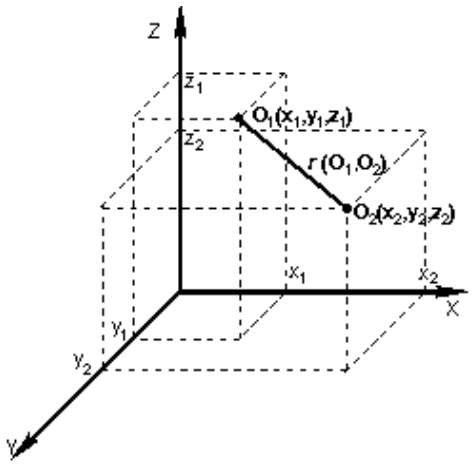

Рисунок 10.2 – Відстань між двома точками в просторі трьох вимірів

Кластер має такі математичні характеристики: центр, радіус, середньоквадратичне відхилення, розмір кластера.

Центр кластера – це середнє геометричне місце точок у просторі змінних.

Радіус кластера – максимальна відстань точок від центру кластера.

Як було відзначено в одній із попередніх тем, кластери можуть бути, такими, що перекриваються. Така ситуація виникає, коли виявляється перекриття кластерів. У цьому випадку неможливо за допомогою математичних процедур однозначно віднести об'єкт до одного з двох кластерів. Такі об'єкти називають спірними.

Спірний об'єкт – це об'єкт, який у міру подібності може бути віднесений до декільком кластерам.

Розмір кластера може бути визначений або за радіусом кластера, або за середньоквадратичним відхиленням об'єктів для цього кластера. Об'єкт належить до кластера, якщо відстань від об'єкта до центру кластера менше радіуса кластера. Якщо ця умова виконується для двох і більш кластерів, об'єкт є спірним. Неоднозначність може бути усунута експертом або аналітиком.

Робота кластерного аналізу опирається на два припущення. Перше припущення – розглянуті ознаки об'єкта в принципі допускають бажане розбиття сукупності об'єктів на кластери. Друге припущення – правильність вибору масштабу або одиниці вимірювання ознак.

Вибір масштабу в кластерному аналізі має велике значення. Розглянемо приклад. Уявимо собі, що дані ознаки *x* у наборі даних А на два порядки більші даних ознаки *y*: значення змінної *x* перебувають в діапазоні від 100 до 700, а значення змінної *y* – у діапазоні від 0 до 1.

Тоді, при розрахунках величини відстані між точками, що відображають положення об'єктів у просторі їх властивостей, змінна, що має більші значення, тобто змінна *x*, буде практично повністю домінувати над змінною з малими значеннями, тобто змінної *y*. У такий спосіб через неоднорідність одиниць виміру ознак стає неможливим коректно розрахувати відстані між точками.

Ця проблема вирішується за допомогою попередньої стандартизації змінних. Стандартизація (standardization) або нормування (normalization) приводить значення всіх перетворених змінних до єдиного діапазону значень шляхом вираження через відношення цих значень до якоїсь величини, що відображає певні властивості конкретної ознаки. Існують різні способи нормування вихідних даних.

Два найпоширеніші способи:

- розподіл вихідних даних на середньоквадратичне відхилення відповідних змінних;
- обчислення Z-внеску або стандартизованого внеску.

Поряд зі стандартизацією змінних, існує варіант додавання до кожної з них певного коефіцієнта важливості, або ваги, яка би відображала значимість відповідної змінної. За ваги можуть виступати експертні оцінки, отримані в ході опитування експертів – фахівців предметної області. Отримані добутки нормованих змінних на відповідні ваги дозволяють одержувати відстані між точками в багатомірному просторі з урахуванням неоднакової ваги змінних.

У ході експериментів можливе порівняння результатів, отриманих з урахуванням експертних оцінок і без них, і вибір якіснішого з них.

2. Методи кластерного аналізу

Методи кластерного аналізу можна розділити на дві групи:

- ієрархічні;
- неієрархічні.

Кожна із груп включає безліч підходів і алгоритмів.Використовуючи різні методи кластерного аналізу, аналітик може одержати різні розв’язки для тих самих даних. Це вважається нормальним явищем.

Розглянемо ієрархічні й неієрархічні методи докладно.

Ієрархічні методи кластерного аналізу. Суть ієрархічної кластеризації полягає в послідовному об’єднанні менших кластерів у більші або поділі більших кластерів на менші.

Ієрархічні агломеративні методи (Agglomerative Nesting, AGNES). Ця група методів характеризується послідовним об’єднанням вихідних елементів і відповідним зменшенням числа кластерів.

На початку роботи алгоритму всі об’єкти є окремими кластерами. На першому кроці найбільш схожі об’єкти поєднуються в кластер. На наступних кроках об’єднання триває доти, поки всі об’єкти не будуть становити один кластер.

Ієрархічні дивизимні (ділені) методи (Divisive Analysis, DIANA). Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об’єкти належать одному кластеру, який на наступних кроках ділиться на менші кластери, у результаті утворюється послідовність груп, що розщеплюються. Принцип роботи описаних вище груп методів у вигляді дендрограми показаний на рис. 10.3.

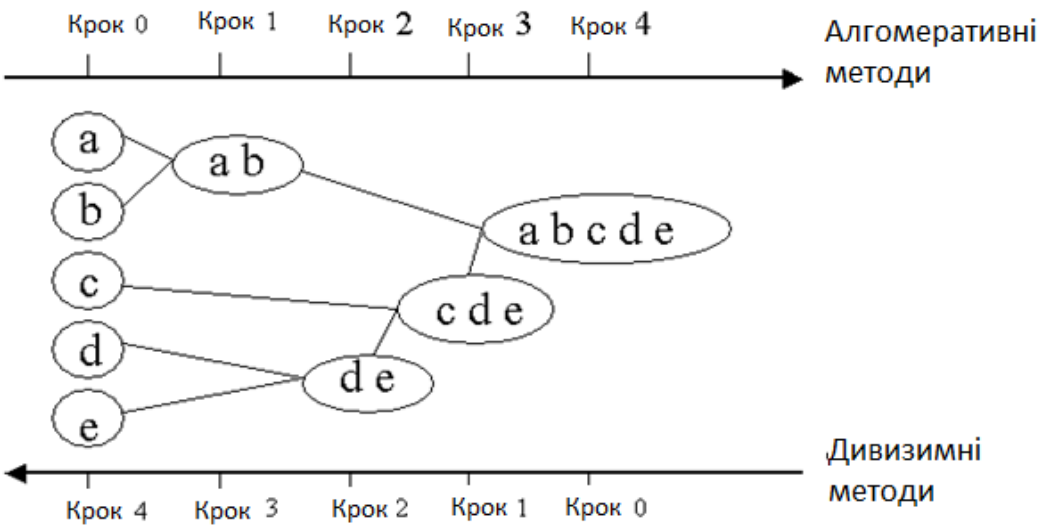

Рисунок 10.3 – Дендрограма агломеративних і дивизимних методів

Програмна реалізація алгоритмів кластерного аналізу широко представлена в різних інструментах Data Mining, які дозволяють вирішувати завдання досить великої розмірності. Наприклад, агломеративні методи реалізовані в пакеті SPSS, дивизимні методи – у пакеті Statgraf.

Ієрархічні методи кластеризації різняться правилами побудови кластерів. За правила виступають критерії, які використовуються при вирішенні питання про «схожість» об'єктів при об'єднанні їх в групу (агломеративні методи) або поділу на групи (дивизимні методи).

Ієрархічні методи кластерного аналізу використовуються при невеликих обсягах наборів даних.

Перевагою ієрархічних методів кластеризації є їхня наочність.

Ієрархічні алгоритми пов'язані з побудовою дендрограм (від грецького dendron – «дерево»), які є результатом ієрархічного кластерного аналізу.

Дендрограма описує близькість окремих точок і кластерів один до одного, представляє в графічному вигляді послідовність об'єднання (поділу) кластерів.

Дендрограма (dendrogram) – деревоподібна діаграма, що містить *n* рівнів, кожний з яких відповідає одному з кроків процесу послідовного укрупнення кластерів. Дендрограму також називають деревоподібною схемою, деревом об'єднання кластерів, деревом ієрархічної структури.

Дендрограма являє собою вкладене угруповання об'єктів, яке змінюється на різних рівнях ієрархії.

Існує багато способів побудови дендограмм. У дендограмі об'єкти можуть розташовуватися вертикально або горизонтально. Приклад вертикальної дендрограми наведений на рис. 10.4.

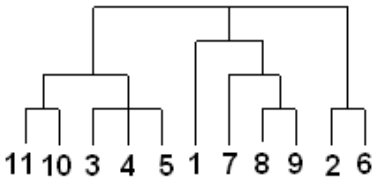

Рисунок 10.4 – Приклад дендрограмми

Числа 11, 10, 3 і т.д. відповідають номерам об’єктів або спостережень вихідної вибірки. Бачимо, що на першому кроці кожне спостереження представляє один кластер (вертикальна лінія), на другому кроці спостерігаємо об’єднання таких спостережень: 11 і 10; 3, 4 і 5; 8 і 9; 2 і 6. На другому кроці триває об’єднання в кластери: спостереження 11, 10, 3, 4, 5 і 7, 8, 9. Цей процес триває доти, поки всі спостереження не об’єднаються в один кластер.

Міри подібності. Для обчислення відстані між об’єктами використовуються різні міри подібності, їх називають також метриками або функціями відстаней. На початку теми розглянуто евклідову відстань, це найбільш популярна міра подібності.

Квадрат евклідової відстані. Для надання більшої ваги більш віддаленим один від одного об’єктів можемо скористатися квадратом евклідової відстані шляхом піднесення у квадрат стандартної евклідової відстані.

Манхеттенска відстань (відстань міських кварталів), також називається «хемінговою» або «сіті-блок» відстанню. Ця відстань розраховується як середня різниць по координатах. У більшості випадків ця міра відстані приводить до результатів, подібних розрахункам відстані евкліда. Однак, для цієї міри вплив окремих викидів менший, ніж при використанні евклидової відстані, оскільки тут координати не підносяться до квадру.

Відстань Чебишева. Цю відстань варто використовувати, коли необхідно визначити два об’єкти як «різні», якщо вони відрізняються за якимось одним виміром.

Відсоток незгоди. Ця відстань обчислюється, якщо дані є категоріальними.

Методи об’єднання або зв’язки. Коли кожний об’єкт являє собою окремий кластер, відстані між цими об’єктами визначаються обраною мірою. Виникає таке питання – як визначити відстані між кластерами? Існують різні правила – методи об’єднання або зв’язки для двох кластерів.

Метод найближчого сусіда або одиночний зв’язок. Тут відстань між двома кластерами визначається відстанню між двома найбільш близькими об’єктами (найближчими сусідами) у різних кластерах. Цей метод дозволяє виділяти кластери як завгодно складної форми за умови, що різні частини таких кластерів з’єднані ланцюжками близьких один до одного елементів. У результаті роботи цього методу кластери представляються довгими «ланцюжками» або «волокнистими» кластерами, «зчепленими разом» тільки окремими елементами, які випадково виявилися ближче інших один до одного.

Метод найбільш віддалених сусідів або повний зв’язок. Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об’єктами в різних кластерах (тобто «найбільш віддаленими сусідами»).

Метод добре використовувати, коли об'єкти дійсно походять із різних «ділянок». Якщо ж кластери мають до певної міри подовжену форму або їх природній тип є «ланцюговим», то цей метод не слід використовувати.

Метод Варда (Ward's method). За відстань між кластерами береться приріст суми квадратів відстаней об'єктів до центрів кластерів, одержуваний у результаті їх об'єднання (Ward, 1963). На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами, тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму поєднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньо групової суми квадратів. Цей метод спрямований на об'єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру.

Метод незваженого попарного середнього (метод незваженого попарного арифметичного середнього – unweighted pair-group method using arithmetic averages, UPGMA (Sneath, Sokal, 1973)). За відстань між двома кластерами береться середня відстань між усіма парами об'єктів у них. Цей метод слід використовувати, якщо об'єкти дійсно походять із різних «ділянок», у випадках присутності кластерів «ланцюгового» типу, при припущенні нерівних розмірів кластерів.

Метод зваженого попарного середнього (метод зваженого попарного арифметичного середнього – weighted pair-group method using arithmetic averages, WPGM A (Sneath, Sokal, 1973)). Цей метод схожий на метод незваженого попарного середнього, різниця полягає лише в тому, що тут як ваговий коефіцієнт використовується розмір кластера (число об'єктів, що втримуються в кластері). Рекомендується використовувати саме при наявності припущення про кластери різних розмірів.

Незважений центроїдний метод (метод незваженого попарного центроїдного усереднення – unweighted pair-group method using the centroid average (Sneath and Sokal, 1973)). За відстань між двома кластерами в цьому методі береться відстань між їхніми центрами ваги.

Зважений центроїдний метод (метод зваженого попарного центроїдного усереднення – weighted pair-group method using the centroid average, WPGMC (Sneath, Sokal 1973)). Цей метод схожий на попередній, різниця полягає в тому, що для обліку різниці між розмірами кластерів (числа об'єктів у них), використовуються ваги. Використовують переважно у випадках, якщо є припущення щодо істотних відмінностей у розмірах кластерів.

3. Ієрархічний кластерний аналіз

Розглянемо процедуру ієрархічного кластерного аналізу в пакеті SPSS (SPSS), в якому ця процедура передбачає угруповання як об'єктів (рядків матриці даних), так і змінних (стовпців). Можна вважати, що в останньому випадку роль об'єктів відіграють змінні, а роль змінних – стовпці.

У цьому методі реалізується ієрархічний агломеративний алгоритм, зміст якого полягає в такому. Перед початком кластеризації всі об'єкти

вважаються окремими кластерами, у ході алгоритму вони поєднуються. Спочатку вибирається пара найближчих кластерів, які поєднуються в один кластер. У результаті кількість кластерів стає рівним *N-1*. Процедура повторюється, поки всі класи не об'єднаються. На будь-якому етапі об'єднання можна перервати, одержавши потрібне число кластерів. Отже, результат роботи алгоритму агрегування залежить від способів обчислення відстані між об'єктами й визначення близькості між кластерами.

Для визначення відстані між парою кластерів можуть бути сформульовані різні підходи. З урахуванням цього в SPSS передбачені такі методи:

- Середня відстань між кластерами (Between-groups linkage), установлюється за замовчуванням.
- Середня відстань між усіма об'єктами пари кластерів з урахуванням відстаней усередині кластерів (Within-groups linkage).
- Відстань між найближчими сусідами – найближчими об'єктами кластерів (Nearest neighbor).
- Відстань між самими далекими сусідами (Furthest neighbor).
- Відстань між центрами кластерів (Centroid clustering) або центроидный метод. Недоліком цього методу є те, що центр об'єднаного кластера обчислюється як середнє центрів поєднуваних кластерів, без обліку їх обсягу.
- Метод Варда.
- Метод медіан – той же центроїдний метод, але центр об'єднаного кластера обчислюється як середнє всіх об'єктів (Median clustering).

Приклад ієрархічного кластерного аналізу. Порядок агломерації (протокол об'єднання кластерів) представлених раніше даних наведено в таблиці 10.1. У протоколі зазначені такі позиції:

- Stage – стадії об'єднання (крок);
- Cluster Combined – поєднувані кластери (після об'єднання кластер ухвалює мінімальний номер з номерів поєднуваних кластерів);
- Coefficients – коефіцієнти.

Так, у колонку Cluster Combined можна побачити порядок об'єднання в кластери: на першому кроці були об'єднані спостереження 9 і 10, вони утворюють кластер під номером 9, кластер 10 в оглядовій таблиці більше не з'являється. На наступному кроці відбувається об'єднання кластерів 2 і 14, далі 3 і 9, і т.д.

Таблиця 10.1 − Порядок агломерації

Stage	Cluster Combined	Coefficients	Результат
1	1	-	-
2	2	14	1,461E-02
3	3	9	1,461E-02
4	5	8	1,461E-02
5	6	7	1,461E-02
6	3	13	3,490E-02

Stage	Cluster Combined	Coefficients	Результат
7	2	11	3,651E-02
8	4	5	4,144E-02
9	2	6	5,118E-02
10	4	12	0,105
11	1	3	0,120
12	1	4	1,217
13	1	2	7,516

У колонку Coefficients наведена кількість кластерів, яку варто було б уважати оптимальною; під значенням цього показника мається на увазі відстань між двома кластерами, визначене на підставі обранооїміри відстані. У нашому випадку це квадрат відстані, обчислений із використанням стандартизованих значень. Процедура стандартизації використовується для виключення ймовірності того, що класифікацію будуть визначати зміни, що мають найбільший розкид значень. У SPSS застосовуються такі види стандартизації:

- Z-Шкали (Z-Scores). Зі значень змінних віднімається їхнє середнє, і ці значення діляться на стандартне відхилення.
- Розкид від -1 до 1. Лінійним перетворенням змінних домагаються розкиду значень від -1 до 1.
- Розкид від 0 до 1. Лінійним перетворенням змінних домагаються розкиду значень від 0 до 1.
- Максимум 1. Значення змінних діляться на їхній максимум.
- Середнє 1. Значення змінних діляться на їхнє середнє.
- Стандартне відхилення 1. Значення змінних діляться на стандартне відхилення.

Крім того, можливі перетворення самих відстаней, зокрема, можна відстані замінити їхніми абсолютними значеннями, це актуально для коефіцієнтів кореляції. Можна також усі відстані перетворити так, щоб вони змінювалися від 0 до 1.

Визначення кількості кластерів. Існує проблема визначення числа кластерів. Іноді можна апріорно визначити це число. Однак у більшості випадків число кластерів визначається в процесі агломерації/поділу безлічі об’єктів.

Процесу угруповання об’єктів в ієрархічному кластерному аналізі відповідає поступове зростання коефіцієнта, який називається критерієм E. Стрибкоподібне збільшення значення критерію E можна визначити як характеристику числа кластерів, які дійсно існують у досліджуваному наборі даних. Отже, цей спосіб зводиться до визначення стрибкоподібного збільшення деякого коефіцієнта, який характеризує перехід від сильно зв’язаного до слабко зв’язаного стану об’єктів.

У таблиці 10.1 ми бачимо, що значення поля Coefficients збільшується стрибкоподібно, отже, об’єднання в кластери слід зупинити, інакше буде відбуватися об’єднання кластерів, що перебувають на відносно великій відстані один від одного. У цьому прикладі це стрибок з 1,217 до 7,516.

Оптимальним уважається кількість кластерів, рівне різниці кількості спостережень (14) і кількості кроків до стрибкоподібного збільшення коефіцієнта (12).

Отже, після створення двох кластерів об’єднань більше проводити не слід, хоча візуально очікувалася поява трьох кластерів.

Агрегування даних може бути презентовано графічно у вигляді дендрограми. Вона визначає об’єднані кластери й значення коефіцієнтів на кожному кроці агломерації (відображені значення коефіцієнтів, наведені на шкалі від 0 до 25).

Дендрограма для цього прикладу наведена на рис. 10.5. Розріз дерева агрегування вертикальною рисою дав нам два кластери, що складаються із 9 і 5 об’єктів.

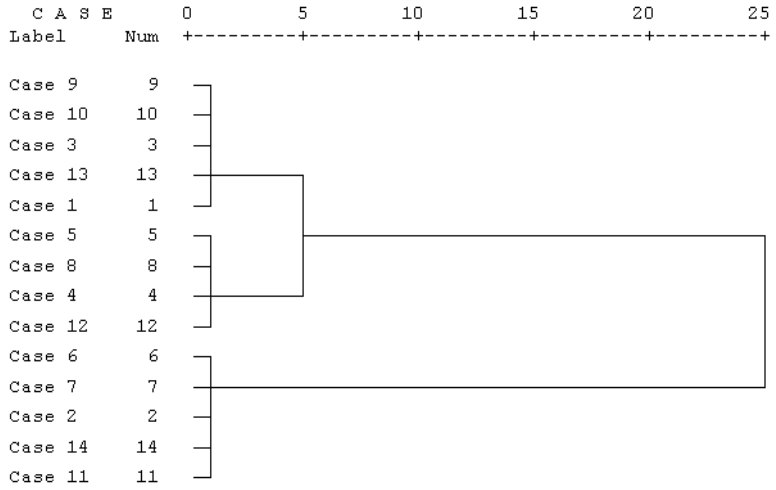

Рисунок 10.5 – Дендрограмма процесу злиття

На верхній лінії по горизонталі відзначені номери кроків алгоритму, усього алгоритму треба було 25 кроків для об’єднання всіх об’єктів в один кластер.

Питання для самоконтролю

1. Які існують групи задач кластерного аналізу?
2. Як вимірюється відстань між двома точками? Які функції відстані ви знаєте?
3. Дайте визначення математичним характеристикам кластеру: центр, радіус, середньоквадратичне відхилення, розмір кластера?
4. Які найпоширеніші способи нормування (normalization) змінних?
5. Які існують групи кластерного аналізу? Чим вони відрізняються?
6. Як відбувається визначення кількості кластерів?