Кластерний аналіз

Розбудова моделі складної метричної ознаки передбачає застосування кластерного аналізу. Об’єкти в економіці, як правило, мають складну структуру, а отже, сукупності значень ознак, тобто значень показників, перш ніж обробляти, потрібно завжди перевіряти на однорідність, тоді результати обчислень будуть надійними. За допомогою методів автоматичної класифікації можна здійснювати ефективні перевірки. Різні застосування кластерного аналізу зводяться до чотирьох основних задач: розробки типології чи класифікації; дослідження корисних концептуальних схем групування об’єктів; породження гіпотез на основі дослідження даних; перевірки гіпотез або дослідження для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

Існування різних обчислювальних алгоритмів кластерного аналізу пояснюється вимогами завдань, що вирішуються на практиці, а також тим, що теорія кластерного аналізу ще далека від завершення. Фахівці з даного методу застерігають, що, по-перше, багато методів кластерного аналізу не мають достатнього статистичного обґрунтування; по-друге, дані методи розроблялися для багатьох наукових дисциплін і тому несуть у собі їх відбиток; по-третє, різні кластерні методи можуть породжувати й породжують різні рішення для одних і тих же даних; по-четверте, сама процедура кластеризації може привнести певну структуру, якої насправді немає.

Сучасними різновидами кластерного аналізу є детерміновані процедури, проте глибоке вивчення можливостей розшарування на кластери приводить до висновку, що даний метод має бути заснований на ймовірнісних підходах з обчисленням ймовірностей належності об’єкта до того чи іншого класу. Техніка кластерного аналізу базується на поняттях схожості об’єктів (типів об’єктів). Підбором найподібніших об’єктів виконується розподіл сукупності на групи (кластери, класи, таксони). На відміну від звичайних групувань за однією ознакою кластерний аналіз виявляє природне розшарування величин на групи з урахуванням всіх ознак одночасно. Проте строго аналітично не визначаються

чіткі межі кожної групи (це право залишається за ОУР – особою, що ухвалює рішення). У комп'ютерних реалізаціях методу також не визначається, на яку кількість груп розподіляється сукупність, – це також має зробити ОУР. Такий стан у кластерному аналізі свідчить про незавершеність теорії і є його недоліком, що не заважає його широкому застосуванню.

Умови задачі в економіці, яка так чи інакше зводиться до задачі кластеризації, визначають перелік проблем, які необхідно вирішити для її розв'язання. Незважаючи на те що існує багато обчислювальних процедур, у результаті яких отримуються різні види дендрограмм, тобто декілька різних розшарувань об'єктів, наразі існують уже надійні методи кластерного аналізу, що пройшли численні перевірки. Однією з таких процедур, що заслуговує на довіру, на думку автора, є метод Уорда. Звичайно, результати класифікації, в тому числі і за методом Уорда, потребують незалежної перевірки. Було перевірено усталеність отриманої класифікації за методом Уорда за допомогою дискримінантного аналізу, заснованого на відстані Махаланобіса (це буде обговорюватися нижче). Якість класифікації за методом Уорда також доведена в процесі розвязання задачі про розділення суміші на однорідні групи об'єктів.

Обчислювальні процедури кластерного аналізу передбачають вирішення наступних математичних проблем.

Проблема вибору типу кластерного аналізу. Методи кластерного аналізу підрозділяються на сім великих груп: ієрархічні агломеративні методи; ієрархічні дивізимні методи; ітеративні методи групувань; методи пошуку модальних значень щільності; факторні методи; методи згущень; методи, що використовують теорію графів. Але на практиці, особливо в процесі розв'язування задач в економіці, найчастіше використовують дві великі групи: агломеративні (об'єднавчі) та дивізимні (розподільчі). Агломеративні методи послідовно об'єднують окремі об'єкти в групи (кластери), а дивізимні послідовно розділяють великі групи на менші – однорідніші групи. У свою чергу, кожен метод як об'єднавчого, так і розподільчого типу може бути реалізований за допомогою різних алгоритмів. Більшість із цих алгоритмів була

перевірена на відомих класифікаціях. Слід врахувати існуючу в аналітиків думку про те, що ніякою кількістю успішних прикладів не можна підтвердити математичну методику, але одного невдалого прикладу достатньо, щоб відбраковувати неправильний алгоритм. Ці пробні обчислення були використані для відбраковування явно неефективних методів кластерного аналізу. У цілому вважається, що ієрархічний агломеративний кластерний аналіз найдоцільніший та найдоступніший у використанні для великих сукупностей значень величин ознак.

Ієрархічні процедури майже не допускають помилок під час початкового об'єднання (перші ітерації під час об'єднання початкових об'єктів), і тільки далі зі збільшенням кількості ітерацій можуть об'єднуватися не досить схожі об'єкти.

Основна ідея ієрархічних алгоритмів полягає в наступному. На першому кроці кожен об'єкт вважається окремим кластером. На наступному кроці об'єднуються два найближчі об'єкти, які утворюють новий кластер. Далі ця процедура повторюється, доки всі об'єкти не об'єднаються в один кластер.

Результати роботи ієрархічної процедури зазвичай оформляються у вигляді дендрограми, в якій наведені номери об'єднаних об'єктів та значення міри схожості (S_{ij}), за якою ці об'єкти були об'єднані. Об'єкти об'єднуються в кластери за мірою схожості між ними. Протилежне поняття до міри схожості – це відстань між об'єктами (d_{ij}). Міра схожості корисна на заключній стадії для кращої інтерпретації результатів. Однакові об'єкти мають схожість $S = 100$ (100%), абсолютно різнорідні об'єкти – схожість $S = 0$. На якомусь критичному рівні схожості дендрограма розрізається та розпадається на окремі кластери, виявлення яких є метою будь-якого економічного аналізу. Найчастіше дендрограму розрізають за рівня схожості $S = (50 \div 60)\%$. Отже, якщо реалізовано метод з урахуванням міри схожості, автоматично розв'язується проблема визначення кількості кластерів.

Проблема нормування значень показників ознак об’єкта. Оцінка відстані між об’єктами надзвичайно залежить від абсолютного значення величини ознаки та від ступеня її варіації в сукупності значень. Щоб ліквідувати подібний вплив на процедуру класифікації, вихідні значення показників нормуються. Найпопулярнішою є стандартизація всіх показників. У процесі стандартизації як норма приймається стандартне відхилення (середньоквадратичне відхилення S_x):

$$z_i = \frac{x_i - \bar{x}_i}{S_{x_i}},$$

де $\overline{x_i}$ – середнє значення показника.

Звичайно, даний спосіб нормування не позбавлений недоліків: за стандартизації всі обчислення прив’язуються до конкретної вибірки спостережень, яка може бути будь-якою, в тому числі такою, що викликає сумніви щодо своєї репрезентативності. Різні показники по-різному впливають на класифікацію. Зокрема, маловаріабельні величини ознак практично не впливають на процес, але після стандартизації всі показники стають рівноправними за дисперсією, що дорівнює одиниці, в тому числі й маловаріабельні ознаки. Тому стандартизацію необхідно доповнювати процедурою попереднього відбраковування, що відноситься до перевірки робастності величини ознак. Відбраковувати потрібно ті ознаки, які мають коефіцієнт варіації величини, менший за 2% чи навіть 5% (про це йшла мова в попередньому розділі). У деяких випадках стандартизацію проводять з урахуванням вагових коефіцієнтів, які задаються пропорційно до ступеня важливості ознак, а отже, їх показників.

Альтернативне нормування, що враховує відмінності у варіації окремих показників:

$$z_i = \frac{x_i}{\bar{x}_i},$$

позбавлене багатьох недоліків попередньої, але воно залишилось прив'язаним до конкретної вибірки. Модифікацією альтернативного нормування є:

$$z_j = \frac{x_j}{x_{j\ max} - x_{j\ min}}.$$

У деяких умовах управління діяльністю підприємств має сенс ще один вид нормування

$$z_i = \frac{x_i}{a_i},$$

де a_i – визначене (економічно обґрунтоване) еталонне значення величини даної ознаки (якщо це значення відоме).

Проблема усталеності класифікації (за ознаками). Результати класифікації в скороченому просторі усталеніші та надійніші, ніж у багатовимірному; мала кількість параметрів легше піддається змістовному сприйняттю та подальшому аналізу, ніж велика кількість; у випадку скорочення простору до розмірності 1 – 3 дані стають візуально спостережуваними, а наочність корисна у всіх відношеннях; скорочення кількості ознак приводить до спрощення обчислювальних процедур класифікації.

Для скорочення простору ознак доцільно спочатку використовувати факторний аналіз, який на основі об'єктивно існуючих залежностей між показниками ознак дозволяє виконувати частину редукції простору ознак.

Проблема відстані між об'єктами. Вибір метрики є вузловим моментом дослідження, від якого залежить кінцевий варіант розподілу сукупності об'єктів на класи. Схожість чи відмінність між об'єктами, що класифікуються, встановлюється залежно від метричної відстані між ними. Якщо об'єкт сукупності описується m показниками, то він може бути представлений точкою

в m-вимірному просторі і схожість з іншими об'єктами буде визначатися відстанню до інших об'єктів сукупності. У кластерному аналізі використовуються різні міри відстаней між об'єктами:

1) евклідова відстань:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(z_{ik} - z_{jk}\right)^2}, \; i, j = \overline{1,n}, \; k = \overline{1,m};$$

2) зважена евклідова відстань:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k \left(z_{ik} - z_{jk}\right)^2};$$

3) відстань city-block:

$$d_{ij} = \sum_{k=1}^{m} \left| z_{ik} - z_{jk} \right|;$$

4) узагальнена відстань Мінковського:

$$d_{ij} = \left(\sum_{k=1}^{n} \left| z_{ik} - z_{jk} \right|^{p} \right)^{1/p};$$

5) відстань Махаланобіса:

$$d_{ij}{}^2 = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)' \times \boldsymbol{S}^{-1} \times \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right),$$

де d_{ij} – відстань між i-м та j-м об'єктами;

w_k – вагові коефіцієнти, щодо k-го показника;

x_{ki}, x_{kj} – значення k-го показника відповідно для i-го та j-го об'єктів;

z_{ki}, z_{kj} – те ж для стандартизованих величин;

$\boldsymbol{x}_i, \boldsymbol{x}_j$ – вектори значень показників для i-го та j-го об'єктів;

$\boldsymbol{S}$ – загальна коваріаційна матриця.

Щодо відстані Махаланобіса, яка, на думку деяких фахівців, є найбільш доцільною в реалізації процедур кластерного аналізу, то для її застосування потрібно обчислити матрицю коваріацій за всіма показниками, знайти обернену матрицю і скласти квадратну форму з цієї матрицею. Як на позитивну якість відстані Махаланобіса вказують на її інваріантність відносно геометричних лінійних перетворень. За відсутності кореляції між показниками відстань Махаланобіса зводиться до евклідової відстані з попередньою стандартизацією всіх показників. Проте ця відстань має багато негативних якостей, у тому числі всі негативи, що обумовлені стандартизацією (про що було сказано раніше). Крім того, у разі зростання кореляції ці ефекти нормування не зникають, а, навпаки, набувають нових, ускладнених форм. Якщо кореляції близькі до одиниці, визначник коваріаційної матриці наближається до нуля, а сама матриця є близькою до виродженої. Урахування коваріацій обумовлює залежність відстані між двома точками від відстаней між іншими точками. Наявність таких негативних рис може спотворити всю матрицю цих відстаней. Тому громіздка міра Махаланобіса, на думку автора, не має жодних переваг і не повинна рекомендуватися до широкого застосування.

Однією з найкращих мір відстаней між окремими об'єктами є так звана «дивергенція»:

$$d_{ij} = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{x_{ki} - x_{kj}}{x_{ki} + x_{kj}}\right|.$$

Дана міра не потребує ніякого нормування та враховує відносні внески окремих членів. За умов однакової різниці $\left|x_{ki} - x_{kj}\right|$ внесок тим менший, що більші абсолютні значення $\left|x_{ki} + x_{kj}\right|$.

Проблема відстаней між кластерами. Одним з основних моментів в обчисленнях кластерного аналізу є вибір відстані між групами об’єктів (між кластерами). Відомо близько дванадцяти різних методів, чотири з яких найпоширеніші: близького, далекого, середнього зв’язків, центроїдного зв’язку та метод Уорда. Відмінності між цими відстанями випливають із самих назв принципів. Наприклад, принцип центроїдного зв’язку свідчить, що відстань між групами вважається такою, що дорівнює відстані між їх центрами. Вибір виду відстані між кластерами обумовлюється умовами конкретної задачі.

Принцип Уорда (Ward) слід вважати найкращим принципом відстані між кластерами, згідно з яким два кластери будуть найближчими, якщо в разі їх об’єднання мінімізується приріст загальної дисперсії. Цільова функція подається як середньогрупова сума квадратів чи сума квадратів відхилень (СКВ), що обчислюються за формулою:

$$СКВ = x_j^2 - \frac{\left(\sum_j x_j\right)^2}{n},$$

де x_j – значення величини ознаки j-го об’єкта.

Спочатку, коли кожен кластер складається з одного об’єкта, СКВ дорівнює 0. За методом Уорда об’єднуються ті групи чи об’єкти, для яких СКВ отримує мінімальний приріст. Метод передбачає знаходження кластерів майже однакових розмірів, які мають гіперсферичну форму. Необхідно вважати це перевагою використання саме методу Уорда для класифікації об’єктів у вимірюванні величини його ознак.

Відстань між об’єктами в методі Уорда передбачається середньою евклідовою:

$$d_{ij} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(z_{ki} - z_{kj}\right)^2}.$$

Використання цієї відстані теоретично обґрунтоване у випадках, коли:

спостереження вибирають з генеральних сукупностей, що мають багатовимірний нормальний розподіл з коваріаційною матрицею виду $\sigma^2 E_k$, тобто складові X взаємно незалежні та мають одну й ту ж дисперсію;

складові вектора спостережень X однорідні за фізичним змістом (якщо різнорідні, то вони нормуються) та однаково важливі для класифікації.

Проблема міри схожості. Часто відстань між об'єктами має шкалу від нуля до нескінченності, тому міра схожості є визначеним функціональним перетворенням відстані, наприклад, у вигляді так званої потенційної функції:

$$S_{ij} = \frac{100}{1 + d_{ij}^2} \text{ або } S_{ij} = \frac{100}{1 + \frac{1}{2} d_{ij}^2}$$

(при $d_{ij} = 0$ отримується $S_{ij} = 100$, при $d_{ij} \to \infty$ – $S_{ij} \to 0$).

Міра відстані між об'єктами та міра відстані між кластерами – основні поняття в кластерному аналізі. Відстань між об'єктами на сукупності ознак є багатокритеріальною функцією мети, зведеним показником якості, нехай не в абсолютних, а тільки у відносних одиницях. Проблеми обґрунтованої побудови багатокритеріальної функції мети виділені в окрему задачу, і з цього приводу щодо різних запропонованих конструкцій існують суперечливі думки.

Проблема стійкості класифікації. Для перевірки достовірності рішень кластерного аналізу рекомендується використовувати: тести значущості ознак, повторну вибірку, тести значущості для незалежних ознак, метод Монте-Карло.

Сучасний рівень розвитку обчислювальної техніки та програмних засобів дозволяє успішно вирішувати вказані проблеми, залишається лише питання формування адекватної концептуальної схеми технології реалізації кластерного аналізу в проведенні класифікації.

Приклад застосування кластерного аналізу

Для дослідження заробітної плати по регіонам України спочатку доцільно їх кластеризувати за системою шести показників, які визначають стан заробітної плати в кожному регіоні: x_1- середньомісячна заробітна плата; x_2- середня заробітна плата; x_3- заборгованість з виплат заробітної плати; x_4-

фонд оплати праці; x_5 - фонд основної заробітної плати; x_6 - фонд додаткової заробітної плати. Дану проблему слід вирішити за допомогою кластерного аналізу.

Об'єкти в економіці, як правило, мають складну структуру, а отже, сукупності значень ознак, тобто значень показників, перш ніж обробляти, потрібно завжди перевіряти на однорідність, тоді результати обчислень будуть надійними. За допомогою методів автоматичної класифікації можна здійснювати ефективні перевірки. Різні застосування кластерного аналізу зводяться до чотирьох основних задач: розробки типології чи класифікації; дослідження корисних концептуальних схем групування об'єктів; породження гіпотез на основі дослідження даних; перевірки гіпотез або дослідження для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

Існування різних обчислювальних алгоритмів кластерного аналізу пояснюється вимогами завдань, що вирішуються на практиці, а також тим, що теорія кластерного аналізу ще далека від завершення. Фахівці з даного методу застерігають, що, по-перше, багато методів кластерного аналізу не мають достатнього статистичного обґрунтування; по-друге, дані методи розроблялися для багатьох наукових дисциплін і тому несуть у собі їх відбиток; по-третє, різні кластерні методи можуть породжувати й породжують різні рішення для одних і тих же даних; по-четверте, сама процедура кластеризації може привнести певну структуру, якої насправді немає.

Умови задачі в економіці, яка так чи інакше зводиться до задачі кластеризації, визначають перелік проблем, які необхідно вирішити для її розв'язання. Незважаючи на те що існує багато обчислювальних процедур, у результаті яких отримуються різні види дендрограмм, тобто декілька різних розшарувань об'єктів, наразі існують уже надійні методи кластерного аналізу, що пройшли численні перевірки. Однією з таких процедур, що заслуговує на довіру, на думку автора, є метод Уорда. Звичайно, результати класифікації, в тому числі і за методом Уорда, потребують незалежної перевірки.

Принцип Уорда (Ward) слід вважати найкращим принципом відстані між кластерами, згідно з яким два кластери будуть найближчими, якщо в разі їх об'єднання мінімізується приріст загальної дисперсії. Цільова функція подається як середньогрупова сума квадратів чи сума квадратів відхилень. За методом Уорда об'єднуються ті групи чи об'єкти, для яких сума квадратів відхилень отримує мінімальний приріст. Відстань між об'єктами в методі Уорда передбачається середньою евклідовою:

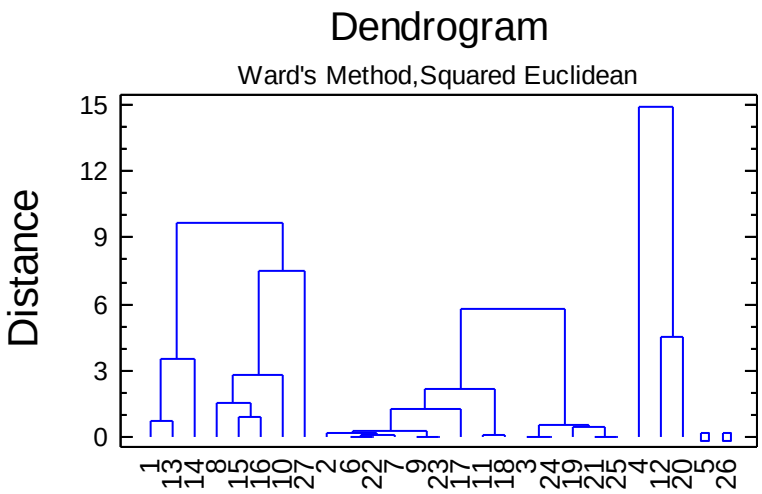
$$d_{ij} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(z_{ki} - z_{kj}\right)^2} .$$

Використання цієї відстані теоретично обґрунтоване у випадках, коли:

спостереження вибирають з генеральних сукупностей, що мають багатовимірний нормальний розподіл з коваріаційною матрицею виду $\sigma^2 E_k$, тобто складові X взаємно незалежні та мають одну й ту ж дисперсію;

складові вектора спостережень X однорідні за фізичним змістом (якщо різнорідні, то вони нормуються) та однаково важливі для класифікації.

Для обчислень кластерного аналізу регіонів України за системою 6 показниками заробітної плати був використаний ППП Statgraphics. У результаті чого маємо розшарування регіонів України на п'ять груп(рис. 3).



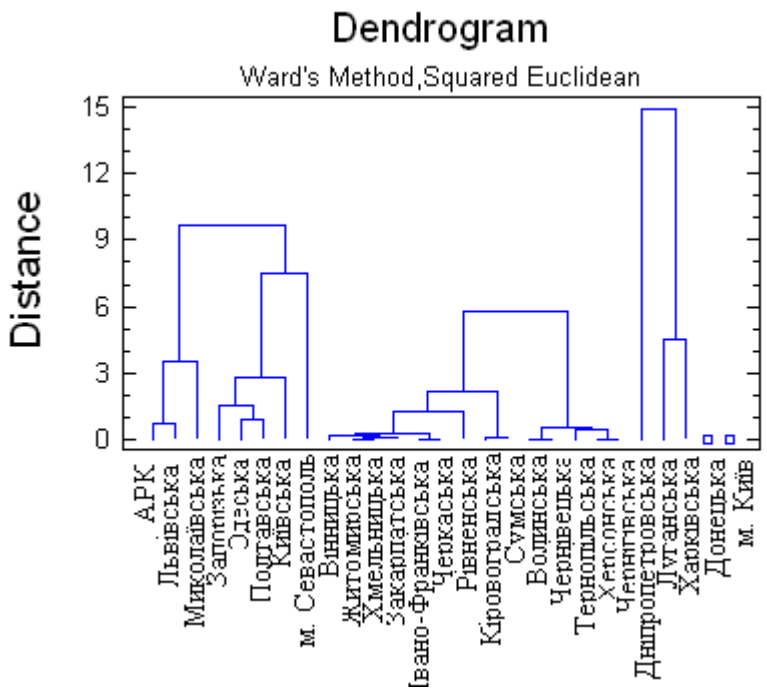


Рис. 3 Дендрограма регіонів України за критерієм заробітної плати в 2011 році

Таким чином, до першої групи увійшли регіони: АРК, Львівська, Миколаївська, Запорізька, Одеська, Полтавська, Київська та м. Севастополь; до другої групи: Вінницька, Житомирська, Хмельницька, Закарпатська, Івано-Франківська, Черкаська, Рівненська, Кіровоградська, Сумська, Волинська, Чернівецька, Тернопільська, Херсонська; до третьої групи: Дніпропетровська, Луганська, Харківська; 4 групу утворили Донецька і 5 групу – м. Київ.