Scientific journal
PHYSICAL AND MATHEMATICAL EDUCATION
Has been issued since 2013.

Науковий журнал
ФІЗИКО-МАТЕМАТИЧНА ОСВІТА
Видається з 2013.

ISSN 2413-158X (online)
ISSN 2413-1571 (print)

http://fmo-journal.fizmatsspu.sumy.ua/

Васильєва Л.В. Методика розв'язання задачі групування багатомірних об'єктів за допомогою кластерного аналізу // Фізико-математична освіта : науковий журнал. – 2017. – Випуск 3(13). – С. 31-34.

Vasilyeva L. The Method Of Solving The Problem Of Grouping Multidimensional Objects By Cluster Analysis // Physical and Mathematical Education : scientific journal. – 2017. – Issue 3(13). – P. 31-34.

УДК 004.67:378.147

Л.В. Васильєва
Донбаська державна машинобудівна академія, Україна
vasileva.dgma@gmail.com

МЕТОДИКА РОЗВ'ЯЗАННЯ ЗАДАЧІ ГРУПУВАННЯ БАГАТОМІРНИХ ОБ'ЄКТІВ ЗА ДОПОМОГОЮ КЛАСТЕРНОГО АНАЛІЗУ

Анотація. *У статті розглянуто питання використання інформаційних технологій у вивченні курсу «Багатомірний статистичний аналіз». Обґрунтована важливість навчання студентів вміння застосовувати математичний апарат та спеціальні програмні засоби для аналізу статистичних даних у майбутній професійній діяльності. У статті наведений алгоритм розв'язання задачі кластерного аналізу та розроблена методика проведення лабораторного заняття для студентів економічного напрямку навчання. Приведені основні відомості про найбільш поширені алгоритми кластерізації: критерій ближнього сусіда, критерій далекого сусіда, критерій центроїда у двох модифікаціях – критерій, розрахований без урахування статистичного ваги поєднуваних груп та критерій середнього сусіда, розрахований з урахуванням числа об'єктів поєднуваних груп. Показані засоби візуалізації процесу вирішення задачі, такі, як діаграма дерева класифікації, таблиця та графік послідовності агломерації. Багатомірне групування статистичних даних виконане на даних з предметної області.*

Ключові слова: *кластерний аналіз,групування багатомірних об'єктів, методика викладання, статистичні дані.*

Постановка проблеми. Дослідження зв'язку між експериментальними даними у більшості практичних випадків зумовлює потребу у побудуванні складних моделей, що, в свою чергу, потребує використання сучасних програмних засобів. При аналізі діяльності промислових підприємств, банків, фермерських господарств необхідно обов'язково дотримуватися вимоги однорідності. Загальноприйняте, що будь яку обробку даних можна проводити тільки для близьких за атрибутивними та кількісними признаками груп спостережень, але первісні статистичні дані не завжди задовольняють цій вимозі. Проблема групування об'єктів може бути вирішена за допомогою формальних процедур кластерного аналізу.

Аналіз актуальних досліджень. Аналіз науково-методичних джерел щодо використання вказаної процедури показує наявність значної кількості наукових робіт з даної тематики. Теоретичні проблеми розглянуті в роботах Б. Дюрана [1], Соколенко С. І. [2], Айвазяна С. А.[3]. Практичне використання метод групування багатомірних об'єктів знаходить в психології, політології, соціально-економічних дослідженнях. У якості програмного забезпечення для проведення кластерного аналізу найчастіше пропонуються універсальні пакети Statistica, SPSS, StatGraphics, є також спеціалізовані програми, такі, як вільно розповсюджувана програма HCE 3.5, або спеціально написані за допомогою VBA додатки до Excel. Здебільшого, автори наукових досліджень зосереджують увагу на самому методі або на його використанні в подальших наукових дослідженнях.

Метою цієї статті є обґрунтування методики проведення кластерного аналізу груп підприємств на підібраному статистичному матеріалі з предметної області студентами економічних спеціальностей за допомогою спеціальних програмних засобів.

Завдання цієї статті полягає у розробці алгоритму й методики проведення лабораторного заняття для студентів економічних спеціальностей з використанням модулю Cluster Analysis програми Statistica6 [4] для

аналізу даних на прикладах з предметної області. Диференційований підхід передбачає розгляд питань обов'язкового рівня засвоєння матеріалу, формування поняття про кластер, метрику, однорідні групи. Бажано, щоб студенти вже мали уявлення про побудову регресійних моделей, що розглядається у курсі «Економіко-математичні методи та моделі» [5].

Виклад основного матеріалу. Теоретичні питання, що виносяться на лабораторну роботу, включають в себе основні визначення та розуміння алгоритму класифікації.

Розбиття на кластери виконують, реалізуючи один з ієрархічних алгоритмів класифікації. На «нульовому» кроці алгоритму кожний об'єкт сукупності вважається окремим кластером, на останньому вся сукупність об'єктів – один кластер. На якому етапі кластеризації зупинитися і «зняти» інформацію про розбиття, дослідник вирішує сам. Найбільш поширені алгоритми кластеризації:

1) критерій ближнього сусіда (Single linkage). На кожному кроці об'єднуються кластери K_P і K_S, відстань між найближчими об'єктами p і s яких мінімальна: $d(K_P, K_S) = \min [\min d(z_p, z_s)]$. При його використанні на першому кроці поєднуються два найближчих між собою об'єкти, далі об'єднання відбуваються за мінімальною відстанню між двома найближчими сусідами.

2) критерій далекого сусіда (Complete linkage). На кожному кроці об'єднуються кластери K_P и K_S, відстань між найбільш віддаленими об'єктами p і s яких мінімальна: $d(K_P, K_S) = \min [\max d(z_p, z_s)]$. При його використанні на першому кроці поєднуються два найближчих між собою об'єкти, на другому кроці – кластери з мінімальною відстанню між двома далекими сусідами і т. д.

– критерій центроїда. На кожному кроці об'єднуються кластери K_P и K_S, відстань між центрами тяжіння яких мінімальна: $d(K_P, K_S) = \min \quad d(\bar{z}_p, \bar{z}_s)$. Даний критерій має дві модифікації в залежності від способу обліку чисельності кожного кластера: критерій центроїда, розрахований без урахування числа об'єктів (статистичної ваги) поєднуваних груп (Unweighted pair-group centroid); критерій середнього сусіда (центроїда), розрахований з урахуванням числа об'єктів (статистичної ваги) поєднуваних груп (Weighted pair-group centroid). При застосуванні другої модифікації при інших рівних умовах більш дрібні кластери приєднуються до більш великих.

Розглянемо методику вивчення деталей процесу кластеризації в пакеті Statistica6 на прикладі.

Завдання: отримані дані з 10 підприємств, які за рік характеризуються трьома економічними показниками: Var1 – продуктивність праці, грн.; Var2 – фондовіддача, грн.; Var3 – рівень рентабельності, %. Статистична інформація наведена на рис. 1.

За наведеними статистичними даними треба здійснити багатомірне групування, скориставшись модулем Cluster Analisys за двома критеріями, результати порівняти та зробити висновки.

В якості метрики відстані між багатовимірними об'єктами ознакового простору будемо використовувати евклідову відстань, а в ролі критерію об'єднання кластерів – критерій ближнього сусіда (Single linkage). На початковому етапі створюємо файл вихідних даних та здійснюємо їх стандартизацію по стовпцях: виділити дані (Var1-Var3), викликати контекстне меню, вибрати команду Fill/standardize block – Standartize Columns. Вихідні дані заміняться стандартизованими (рис. 2).

	1 Var1	2 Var2	3 Var3
1	854(	1,24	38,34
2	2911	0,63	44,69
3	663(	1,18	39,4
4	734:	1,12	20,1
5	3991	1,05	20
6	576(	0,99	18
7	3842,9	5,46	37,6
8	3457,7	5,53	37,9
9	3066,4	7,05	32,1
10	3011,9	7,29	32,1

Рис. 1. *Таблиця даних*

	1 Var1	2 Var2	3 Var3
1	1,79701:	-0,6842:	0,66843
2	-0,9482	-0,9022	1,3403
3	0,86549:	-0,7056	0,78060
4	1,21322	-0,7271:	-1,2616
5	-0,4215	-0,7521	-1,2722
6	0,44118	-0,7735	-1,4838
7	-0,493	0,82435	0,59013
8	-0,6816	0,84937	0,6218
9	-0,872	1,39275	0,00814
10	-0,8990	1,47854	0,00814

Рис. 2. *Таблиця стандартизованих даних*

Послідовність подальших дій для кластерного аналізу: Statistics – Multivariate Exploratory Techniques – Cluster Analysis – Joining (tree clustering) – Ok – Advanced – Variable – выделить Var1-Var3 (Ok) – Advanced – Input file: Raw data – Cluster: Cases (rows) – Amalgamation (linkage) rule: Single linkage – Distance measure : Euclidean distances – Ok – Vertical plot. Отримаємо діаграму дерева класифікації, за якою вже можна судити про кластери (рис. 3).

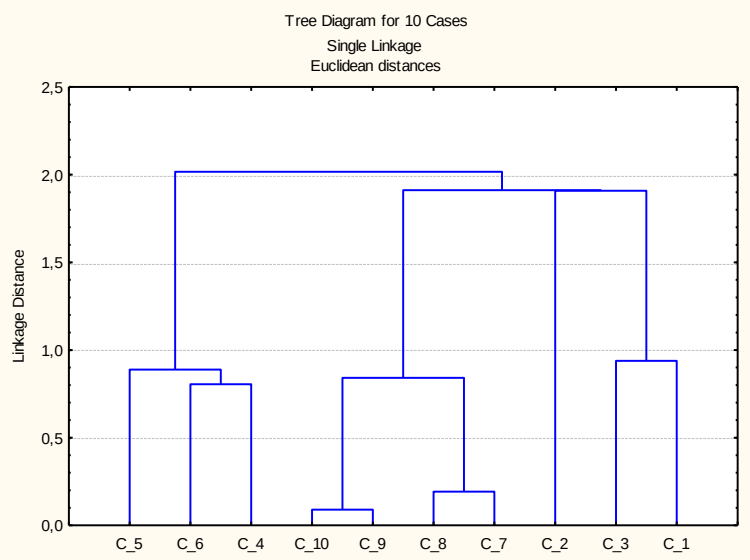

Рис. 3. *Діаграма дерева класифікації*

Більш докладно утворення кластерів можна простежити, вибравши пункти Amalgamation schedule (послідовність агломерації) (рис. 4) або Graph of Amalgamation schedule (графік послідовності агломерації) (рис. 5).

	Amalgamation Schedule (Spreadsheet1) Single Linkage Euclidean distances				
linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
,0898189	C_9	C_10			
,1921645	C_7	C_8			
,8047277	C_4	C_6			
,8416308	C_7	C_8	C_9	C_10	
,8885913	C_4	C_6	C_5		
,9384958	C_1	C_3			
1,908354	C_1	C_3	C_2		
1,911976	C_1	C_3	C_2	C_7	C_8
2,016894	C_1	C_3	C_2	C_7	C_8

Рис. 4. *Послідовність агломерації*

За дендограмою (діаграмою дерева класифікації) робимо висновок, що на дев'ятому кроці всі кластери об'єднуються, і в результаті виходить тривіальне рішення – вся сукупність підприємств об'єднується в один кластер. Очевидно, що таке рішення не має практичної цінності. Необхідно зупинитися на якомусь кроці, це зазвичай роблять при різкому стрибку мінімальної кластерної відстані. Візуальне дослідження дендограми показує наявність такого стрибка на третьому і сьомому кроках процесу кластеризації. Таким чином, в результаті проведення кластерного аналізу багатовимірного угруповання підприємств, робимо висновок, що досліджувану сукупність можна розбити на три групи – (1, 2, 3); (4, 5, 6); (7, 8, 9, 10).

Слід зазначити, що на практиці дослідник спочатку для дослідження вибирає дані, які «інтуїтивно» вважає однорідними. Якщо його інтуїція (або кваліфікація) не підводять, то, провівши процедуру кластеризації, він отримає графік послідовності агломерації, який порівняно монотонно і плавно зростає. Це свідчить про те, що досліджувана сукупність об'єктів достатньо однорідна і можна приступати до регресійного аналізу.

Висновки. За результатами дослідження можна зробити наступні висновки. Наведена вище методика може бути використана під час лабораторної та самостійної роботи студентів економічних спеціальностей при вивченні курсу «Багатомірний статистичний аналіз». Використання спеціальних програм із візуалізацією процесу розв'язання допоможе студентам краще засвоїти досить складний матеріал. Для більш ефективної роботи необхідний підібраний статистичний матеріал з предметної області.

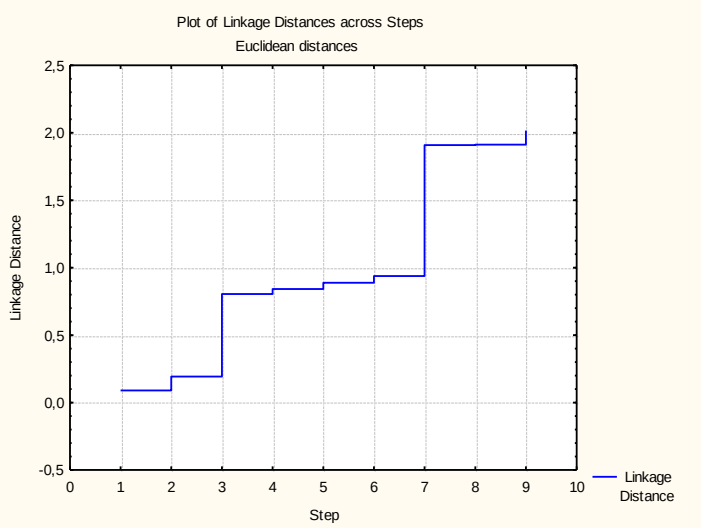

Рис. 5. *Графік послідовності агломерації*

Список використаних джерел

1. Дюран Б. Кластерный анализ / Б. Дюран, П. Одел [пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского]. – М. : Статистика, 1977. – 64 с.
2. Соколенко С. І. Кластери в глобальній економіці / С. І. Соколенко. – К. : Логос, 2004. – 848 с.
3. Айвазян С.А. Классификация многомерных наблюдений / С. А. Айвазян, З.И. Бежаева, О.В. Староверов/. – М. : Статистика, 1974. – 240 с.
4. Боровиков В.П. STATISTICA/ В.П. Боровиков, И.П. Боровиков – М.: Информационно-издательский дом “Филинъ”, 1997. – 592 с.
5. Топтунова Л. М. Дослідження однофакторної і багатофакторної регресії, аналіз часових рядів у системі STATISTICA 6 : навч. посіб. [для студ. екон. спец. вищ. навч. закл.] / Л. М. Топтунова, Л. В. Васильєва, О. А. Кльованік. – Краматорськ : ДДМА, 2008. – 122 с.

References

1. Dyuran B. Cluster analysis / B. Dyuran, P. Odel [per. s anhl. E. Z. Demydenko. Pod red. A. Ya. Boyarskoho]. – M. : Statystyka, 1977. – 64 s. (in Russian)
2. Sokolenko S. I. Clusters in the global economy / S. I. Sokolenko. – K. : Lohos, 2004. – 848 s. (in Ukrainian)
3. Ayvazyan S.A. Classification of multidimensional observations / S. A. Ayvazyan, Z.Y. Bezhaeva, O.V. Staroverov/. – M. : Statystyka, 1974. – 240 s. (in Russian)
4. Borovykov V.P. STATISTICA/ V.P. Borovykov, Y.P. Borovykov – M.: Ynformatsyonno-yzdatel'skyy dom “Fylyn”, 1997. – 592 s. (in Russian)
5. Toptunova L. M. Research single-factor and multifactorial regression, the analysis of time series in the system STATISTICA 6 : navch. posib. [dlya stud. ekon. spets. vyshch. navch. zakl.] / L. M. Toptunova, L. V. Vasilyeva, O. A. Kl'ovanik. – Kramatorsk : DDMA, 2008. – 122 s. (in Ukrainian)

THE METHOD OF SOLVING THE PROBLEM OF GROUPING MULTIDIMENSIONAL OBJECTS BY CLUSTER ANALYSIS

Lyudmila Vasilyeva

Donbass State Engineering Academy

Abstract. *In the article the questions of using information technologies in the study of the course "Multivariate statistical analysis". Explains the importance of teaching students the ability Provides basic information about the most common clustering algorithms: nearest-neighbor criterion, the criterion of farthest neighbor, the centroid criterion in two versions - criterion, calculated without taking into account the statistical weight of the merged groups and the criterion of the average neighbor, calculated based on the number of objects in the merged groups. Shows a visualization of the process of solving the problem, such as a chart of a classification tree, the table and sequence chart of agglomeration. Multidimensional group statistics performed on the data from the subject area. To apply mathematical tools and special software for analysis of statistics in their future professional activities. The article presents the algorithm for solving the cluster analysis and developed methodology for conducting laboratory classes for students of economic fields of study.*

Key words: *cluster analysis, grouping of multidimensional objects, teaching methods, statistical data.*