Determination of the probability and quantity of received information is in problems of informatin – analytial systems / V. V. Hrytsyk (sen), V. V. Hrytsyk (jun). // Інформаційні технології і системи. – 2007. – Т. 10, № 1. – С. 179–190. 12. Hrytsyk (sen) V. V. Information-quality assessment of data process and transfer in condition of noise and distortion of messages in Computer Vision tasks. The pro-bability (confidence) of transmission and processing of information method / V. V. Hrytsyk (sen), V. V. Hrytsyk (jun) // Інформаційні технології і системи. –2007. – Т. 10, № 2. – С. 161-170. 13. Hrytsyk V. V. (sen). Information-quality assessment of data process and transfer in condition of noise and distortion of messages in Computer Vision tasks. General expression for entropy $H_x(\tilde{Y})$ */ V. V. Hrytsyk (sen), V. V. Hrytsyk (jun) // Інформаційні технології і системи. – 2008. – Т. 11, № 2. – С. 165–174. 14. Complex software systems – heal thyself. Reasearch*eu results supplement/ - №25/ – June 2010. – P.28. 15. Find a digital partner to trust .- Reasearch*eu results supplement/ - №25/ - June 2010. – P.34. 16. Seeing understands – using artificial intelligence to analyse multimedia content. .- Reasearch*eu results supplement/ - №25/ - June 2010. – P.36. 17. Software: running commentary for smarter surveillance? .- Reasearch*eu results supplement. – №24. – May 2010. – P.29.*

УДК 519.238.8:331.546

Р.М. Камінський, Л.Я. Нич
Національний університет "Львівська політехніка",
кафедра інформаційних системи та мережі"

ІЄРАРХІЧНИЙ АГЛОМЕРАТИВНИЙ КЛАСТЕРНИЙ АНАЛІЗ ОДНОВИМІРНИХ АСИМЕТРИЧНО РОЗПОДІЛЕНИХ ДАНИХ У СЕРЕДОВИЩІ MS EXCEL

Наведено інформаційну технологію ієрархічного агломеративного кластерного аналізу об'єктів, поданих вибірками одновимірних даних різних обсягів. Ця технологія реалізована в середовищі MsExcel-2003. Вона містить: перетворення одновимірних даних на багатовимірні за допомогою показників описової статистики та параметрів індивідуальних розподілів, формування таблиці "об'єкт–властивість", побудову матриці близькостей, визначення структури дендрограми та інтерпретації кластерів.

Ключові слова: кластерний аналіз, таблиця "об'єкт-властивість", матриця близькостей, дендрограма, кластери.

Hierarchical anglomerative information technology cluster analysis of objects onedimensional data samples of different volumes is presented. This technology is implemented in an environment MsExcel-2003. It includes: the transformation of onedimensional data in multidimensional indexes, using descriptive statistics and distributions of individual parameters, the formation of the table "object- property", build proximity matrix, defining the structure and interpretation of the dendrogram clusters. Used as an example of individual data from 13 operators. As a result of the cluster analysis three clusters were identified and their average parameters were shown. Work is of practical importance in systems training personnel carrier.

Key words: cluster analysis , the table object-property, proximity matrix, dendrohrama, clusters.

Вступ

Використання наукових методів для добору операторського персоналу сприяє підвищенню загальної безпеки та психологічної стійкості як окремого оператора, так і цілого колективу.

Результатом кваліфікаційного добору є вміння персоналу ідентифікувати потенціальну небезпеку, визначати її просторові і часові координати, передбачати можливість прояву небезпечних та шкідливих чинників на людину [1]. Важливою функцією людини-оператора є опрацювання вхідної інформації. Тому такі функції оператора, як обсяг оперативної пам'яті, здатність до інтерполяції і корегування помилок, оперативність у виборі і прийняття рішень мають вельми важливе значення і в першу чергу визначаються психологами, основним технічним арсеналом яких є різноманітні тестові методики та спостереження.

У багатьох системах підвищення кваліфікації та атестації операторського персоналу застосовують комп'ютерні тренажери, в яких як тестові завдання фігурують сценарії – моделі реальних робочих ситуацій. Зміст таких сценаріїв переважно полягає в тому, що оператору-реципієнту на екрані монітора подається послідовність зображень проблемних ситуацій, кожну з яких він повинен оперативно ідентифікувати і прийняти правильне рішення. Основними оцінками такої його діяльності є дані щодо вибору відповідного рішення та оперативності (часу) його реалізації. Кількість ситуацій у сценарії переважно забезпечує достатню репрезентативність емпіричних даних. Оскільки неправильні рішення є рідкісними подіями – їх кількість недостатня для правомірної статистичної оцінки і вказує лише на якість роботи оператора, тому оцінюють його роботу лише за оперативністю прийняття рішень. Час опрацювання поданих ситуацій дає інформацію про індивідуальний характер оператора в сенсі уважності, сконцентрованості, психомоторних особливостей та певних набутих фахових навичок. Отримані в таких дослідженнях результати подаються у вигляді значень часу вибору і прийняття рішення і за своїм характером є індивідуальними вибірками одновимірних величин.

Очевидно, підібрати персонал чи сформувати операторські колективи на підставі одновимірних даних часу прийняття рішення не складно, наприклад, за шкалою вимірювань, розбивши її на відповідні фрагменти, проте такий підхід є не коректним, оскільки не враховує індивідуальних особливостей операторів, які проявляються в вибіркових показниках та параметрах функцій розподілів. Такі індивідуальні показники і параметри різнобічно характеризують операторів як багатомірні об'єкти і можуть бути використані як класифікаційні ознаки у задачах розбиття певної групи учасників тренажерного навчання на підгрупи за рівнем підготовки (кваліфікації) чи будь-яким іншим критерієм [2, 3]. Таке розбиття можна здійснити за допомогою ієрархічного кластерного аналізу, який фактично проводиться в рамках первинного аналізу даних, а його суттєвою перевагою є те, що розбивають множину об'єктів не за однією ознакою, а за цілим набором ознак і його також можна застосовувати в різних предметних областях незалежно від природи об'єктів.

Основною проблемою в більшості систем комп'ютерного навчання та тренінгу, незважаючи на те, що впровадження персонального комп'ютера в інформаційну сферу визначило новий етап розвитку інформаційних технологій, зокрема обробки даних, призначених для розв'язання добре структурованих задач, за відомими алгоритмами і з усіма необхідними вхідними даними, є відсутність ліцензованого програмного забезпечення. Серед найпоширеніших пакетів прикладних програм засоби для проведення кластерного аналізу є в Statistica і SPSS, проте вони не завжди є "під рукою". Наведений в [4] метод "k-середніх" в середовищі Ms Excel є ітераційною процедурою і потребує знання особливостей саме цього методу кластерного аналізу, тобто, визначення наперед кількості кластерів.

У ситуації поділу групи на підгрупи більш відповідною є технологія ієрархічного агломеративного кластерного аналізу, яку можна застосовувати на рівні користувачів невисокої кваліфікації з метою автоматизації деяких рутинних та постійно повторюваних операцій. Хоча вибір інформаційної технології для обробки даних переважно визначається самими дослідженнями, доволі часто дослідник використовує доступні засоби. Одним з таких програмних засобів для обробки даних є табличний процесор **Ms Excel.**

Проте, незважаючи на значну його багатофункціональність, безпосередня реалізація ієрархічного агломеративного аналізу в ньому відсутня. Тим не менш, такі його властивості, як:

автозаповнення та табулювання за формулою; проста і швидка заміна, виключення та транспонування стовпчиків і рядків таблиць; використання *посилань* – фіксації адреси комірки, тобто назви стовпчика і номера рядка за допомогою символу "$" – дають підстави для розроблення простої інформаційної технології проведення ієрархічного агломеративного кластерного аналізу для розв'язання задачі поділу групи багатовимірних об'єктів за скінченним набором індивідуальних класифікаційних ознак.

Метою роботи є розроблення інформаційної технології ієрархічного агломеративного кластерного аналізу групи об'єктів у середовищі табличного процесора Ms Excel, практичність якої має бути підтверджена розв'язанням задачі поділу групи учасників на підгрупи за їхніми індивідуальними вибірками, поданими одномірними асиметрично розподіленими даними.

Очевидно, що реалізація ієрархічного агломеративного кластерного аналізу за такою інформаційною технологією передбачає такі кроки.

1. Побудову таблиці "об'єкт–властивість" за вихідними даними.
2. Вибір та обґрунтування форми нормування класифікаційних характеристик-ознак.
3. Побудову матриці "відстаней" на підставі обґрунтованого вибору метрики.
4. Обчислення параметрів дендрограми та її побудова.
5. Вибір та інтерпретація отриманих кластерів за даною дендрограмою.

Вихідними даними є індивідуальні дані опрацювання тестових ситуацій, рішення щодо яких приймаються на підставі пошуку і виявлення об'єктів заданого класу на складному тлі. В дослідженні взяли участь тринадцять операторів-реципієнтів. Використовувався для всіх один і той самий сценарій, утворений послідовністю зображень з моделями ситуацій. Тривалість експозиції зображення ситуації постійна, а зміна зображень практично миттєва. Тривалість тестування була однакова, проте індивідуальні особливості операторів спричинили різний обсяг їхніх вибірок даних. Зміст даних це тривалість часу – з моменту появи на моніторі зображення ситуації до моменту реалізації рішення. Така тривалість часу є випадковою величиною, причому ця випадковість є суто індивідуальною в рамках даного тестування, а самі дані є одновимірними випадковими величинами з невідомими законами розподілів.

Формування таблиці "об'єкт–властивість"

У багатьох експериментальних дослідженнях низки явищ, ситуацій, об'єктів отримані дані є одновимірними випадковими величинами з невідомим розподілом, що суттєво ускладнює їх розуміння в сенсі індивідуальності та отримання щодо них об'єктивної класифікації.

У випадку індивідуальних вибірок одновимірних даних для повнішої характеристики об'єктів кластерного аналізу (операторів) використовують статистичні показники, отримані в результаті обчислень цих даних та побудови їх емпіричної функції розподілу. Для формування таблиці "об'єкт–властивість" як класифікаційні ознаки вибираємо найхарактерніший для операторської діяльності показники описової статистики та емпіричної функції розподілу варіант, використовуючи індивідуальні дані.

Вибір та обґрунтування класифікаційних ознак за описовою статистикою. Використання описової статистики до одновимірних даних дає близько двох десятків кількісних показників [5]. Деякі показники описової статистики опосередковано дублюються, а саме: стандартне відхилення і його квадрат – дисперсія та інтервал як різниця між максимальним і мінімальним значеннями.

Інструмент "Описова статистика" з *Пакета аналізу даних* у Ms Excel-2003 містить 14 показників, з яких до класифікаційних ознак відібрано такі:

- *середнє значення* – як узагальнюючий показник характеристики оперативності, навколо якого групуються елементи (варіанти) вибірки;
- с*ередньоквадратичне відхилення* – характеризує розкид варіант відносно середнього значення;
- *медіану* – значення вибірки, що відповідає середині впорядкованої сукупності;
- *мінімальне значення,* яке в даному випадку є більш індивідуальною характеристикою оперативності ніж розмах і максимальне значення і характеризує психомоторику оператора;

- *ексцес* – характеристика форми вершини кривої щільності емпіричного розподілу ймовірності випадкової величини;
- *асиметрія* – характеристики несиметричності розподілу варіант відносно середнього значення, власне з погляду скупченості варіант в області малих значень (висока оперативність в прийнятті рішення), хоча практично асиметрію треба визначати відносно моди.

Вибір медіани ґрунтується на тому, що вона є найбільш стійкою характеристикою розподілу вибірки, яка фактично не обчислюється, а має конкретне "місце" на осі значень, незалежно від виду розподілу. Асиметрія та ексцес є визначені через центральні моменти, проте вони є безрозмірними величинами і характеризують лише загальну форму невідомого розподілу цих даних.

Власне середнє стандартне відхилення і характеристики форми є визначальними класифікаційними показниками за описовою статистикою.

Вибір та обґрунтування класифікаційних ознак за емпіричною функцією розподілу. Як емпіричну функцію розподілу вибрано розподіл Вейбула з таких міркувань. По-перше, отримані дані підпадають під категорію даних типу часу життя [6], і по-друге, залежно від значення параметра форми цей розподіл добре апроксимує дані як з експоненційним, так і нормальним розподілом.

Як класифікаційні ознаки з функції розподілу вибрано такі:

- *мода* – значення варіанти, яке відповідає максимальному значенню функції щільності;
- *alfa* – параметр розподілу, що характеризує його форму;
- *beta* – параметр розподілу, який відповідає масштабу;
- *tau1* – значення границі ширини функції щільності розподілу на половині її висоти зліва від моди;
- *tau2* – значення границі ширини функції щільності розподілу на половині її висоти справа від моди.

Останні два параметри функції щільності розподілу визначають інтервал найбільш характерних значень варіант.

Відібрані параметри та характеристики розподілу Вейбула є в певному сенсі ідентифікаційними показниками самого закону розподілу і дають важливу інформацію про індивідуальний характер випадкової величини. Наприклад, залежно від величини значення безрозмірного параметра *alfa* можна вказати на вид кривої щільності розподілу. Наприклад, якщо: $alfa \le 1$ – розподіл експоненційний, відсутня мода; коли $1 < alfa < 2$ – розподіл власне є одномодальним розподілом Вейбула, причому, якщо мінімальне значення вибірки $x_{min} \neq 0$, маємо зрізаний в області малих значень одномодальний розподіл Вейбула, крива функції щільності асиметрична; для $alfa = 2$ – випадок розподілу Релея; $alfa >> 2$ – крива щільності розподілу швидко наближається до симетричної і добре апроксимує криву нормального розподілу [7].

Параметр *beta* має розмірність випадкової величини, а його значення знаходяться в межах між першим і другим квартилями даної вибірки. Його часто називають параметром масштабу, оскільки його положення на числовій осі визначає границі кривої щільності.

Форматування аркушу таблиці "об'єкт–властивість". Для цього чинимо так.

1. Відкриваємо нову книгу. Вона містить $i+1$ аркуш, причому i аркуші (тут i – індекс оператора) призначені окремо для кожного з об'єктів, і окремий аркуші для побудови таблиці "об'єкт-властивість", матриці близькостей та проведення кластерного аналізу. Організовуємо інформацію на аркушах об'єктів так.

2. Вносимо на кожен аркуш індивідуальні дані. В перший рядок записуємо назви: комірка A1 – "номер", далі в стовпчик A3:AN – номери варіант, комірка B1 – "дані", в стовпчик B3:BN – значення даних, в комірки D1 і E1 – "описова статистика", результати описової статистики розміщуємо в D3:D18 і E3:E18, в комірки D20 і E20 – "функція розподілу", а в комірки D22:D26 і E22:E26 розміщуються значення класифікаційних ознак, визначених за функцією розподілу.

Далі надаємо на кожному аркуші такі назви коміркам: G1 – “сортування”, H1 – “зсув або приведення в початок координат”, I1 – “індекс або імовірність”, J1 – “модель1”, K1 – “модель2”, зміст яких наведено нижче.

3. На останньому аркуші визначаємо місце для таблиці “об’єкт–властивість”, наприклад, починаючи з стовпця B. В рядок B1:K1 після об’єднання комірок і центрування вносимо назву “Таблиця “об’єкт–властивість””. Далі вносяться назви класифікаційних ознак у рядок з комірками B2 – “Серед”, C2 – “Медіана”, D2 – “Стандартне відхилення”, E2 – “Ексцес”, F2 – “Асиметрія”, G2 – “Мінімум”, H2 – “alfa”, I2 – “beta”, J2 – “tau1”, K2 – “moda”, L – “tau2”. До стовпчика A3:A15 вносимо назви або номери об’єктів, а в стовпці в полі B3:L15 заповнюємо значеннями відібраних індивідуальних класифікаційних ознак за описовою статистикою та емпіричною функцією розподілу.

4. Виконавши в комірці D3 операції “Сервіс – Аналіз даних – Описова статистика” і задаючи відповідні індивідуальні дані, отримаємо в полі D3:E18 таблицю описової статистики, з якої вибираємо необхідні ознаки для таблиці “об’єкт–властивість” і вносимо їх у відповідні рядки $B_i – G_i$.

Визначення параметрів емпіричної функції розподілу. Для знаходження параметрів розподілу використано процедуру, зміст якої полягає в апроксимації обвідної варіаційного ряду перетвореної перестановкою координат на кумуляту. Для знаходження значень параметрів та характеристик функції розподілу чинимо так.

5. Копіюємо дані стовпчика B3:BN в стовпчик G3:GN і сортуємо їх за зростанням.

6. До стовпчика H3:HN вносимо зі стовпчика G3:GN дані, приведені в початок координат, тобто зменшені на величину значення найменшої у даній вибірці варіанти (віднімаємо від усіх варіант найменшу), в результаті значення в комірці H2 має дорівнювати нулю. Тут N відповідає обсягу індивідуальної вибірки.

7. До комірки I3 вносимо формулу “ =A3/N ” і, застосовуючи автозаповнення, заповнюємо значеннями комірки I3:IN, тобто отримуємо значення номерів зі стовпчика A3:AN, поділені на максимальне значення номера N.

8. За даними стовпчиків H3:HN і I3:IN будуємо графік кумуляти емпіричного розподілу, тобто графік функції

$$F(x) = 1 - \exp\left(-\left(\frac{x}{b} \right)^{a} \right). \qquad (1)$$

9. У комірках D25 і E25 задаємо ім’я і початкове значення параметра форми ***alfa-i*** = 1,5, а в комірках D26 і E26 задаємо ім’я і початкове значення параметра масштабу ***beta-i***, яке дорівнює значенню варіанти, що відповідає варіанті, яка лежить на межі першої і другої третин від початку варіаційного ряду або значення, що відповідає першому квартилю. Присвоюємо ці імена шляхом: **Вставка – Ім’я – Присвоїти**.

10. В комірку J3 вводимо формулу (1) в такому вигляді

“ = 1 - exp(– ((H3 / ***beta-i***)^ ***alfa-i***)) ”,

яку трактуємо як модель1 емпіричної кумуляти, і застосовуючи автозаповнення, обчислюємо значення J3:JN. Виділивши цей стовпчик, вносимо його дані на графік кумуляти. На графіку будуть зображені емпірична (зі стовпчика I3:IN) та теоретична (зі стовпчика J3:JN) кумуляти. Вони можуть навіть дуже різнитися між собою. Далі в комірці E28 визначаємо значення суми квадратів відхилень між ними, застосовуючи оператор **СУММКВРАЗН(I3:IN;J2:JN)**, тобто між даними цих стовпчиків.

11. Використовуючи надбудову “Пошук рішення”, для якої цільовою коміркою є комірка E28 зі значеннями суми квадратів різниць, і змінюючи значення комірок з початковими значеннями параметрів ***alfa-i*** і ***beta-i,*** оптимізуємо ці параметри. Отримані значення параметрів приймаємо як істинні для індивідуальних даних. У результаті значення параметрів зміняться, значення суми квадратів різниць суттєво зменшиться, а графіки емпіричної та теоретичної кумулят максимально зблизяться.

12. Для визначення класифікаційних ознак з функції розподілу, а саме моди та границь ширини функції щільності розподілу на половині її висоти треба ввести в комірку K3 формулу функції щільності розподілу Вейбула

$$f(x)=\frac{a}{b}\cdot\left(\frac{a}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{x}{b}\right)^{a}\right) \qquad (2)$$

у такому вигляді

“ = (*alfa-i*/*beta-i*)*((H3/*beta-i*)^(*alfa-i*-1))*exp(-((H3/*beta-i*)^*alfa-i*)) ”.

Перший множник, який є константою, можна опустити, якщо його величина суттєво більша або менша за одиницю, оскільки він впливає лише на масштаб графіка. Значення i відповідає номеру індивіда. Використовуючи автозаповнення, обчислюємо значення для стовпчика K3:KN. Максимальне значення в цьому стовпчику відповідає моді розподілу, а значення, вдвічі менші за максимальне, визначають границі півширини розподілу.

Значення моди, яке також можна визначити з рівняння (2), прирівнюючи до нуля його ліву частину і скоротивши перший множник $\frac{a}{b}$, тобто

$$0=\left(\frac{a}{b}\right)^{a-1}\cdot\exp\left(-\left(\frac{x_{\text{mod}}}{b}\right)^{a}\right),$$

на відміну значення, яке дає описова статистика, є істинним і точним значенням моди індивідуальної емпіричної функції розподілу.

Границі ширини кривої щільності розподілу на половині її висоти – *tau1* і *tau2* – визначають основну частку найімовірніших значень цієї випадкової величини, тобто частота значень варіант за цими межами є доволі низькою порівняно з частотою варіант у цих межах.

13. Визначені за описовою статистикою та з емпіричної функції закону розподілу індивідуальні класифікаційні ознаки: середнього, медіани, стандартного відхилення, ексцесу, асиметрії, мінімального значення та параметри *alfa-i, beta-i, tau1, моди, tau2* вводимо в таблицю “об’єкт-властивість”, яку зображено на рис. 1.

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		Серед	Медіа	СтВід	Эксце	Асиме	Мінім	alfa	beta	tau 1	moda	tau 2
3	Bur	1886	1840	322	3,837	1,257	1196	2,507	754	1471	1810	2191
4	Cub	574	523	156	1,418	1,361	348	1,72	231	380	487	662
5	Cup	1217	1180	218	2,826	1,069	719	2,731	538	940	1176	1418
6	Hav	754	727	169	1,714	1,208	483	1,869	291	537	676	879
7	Hod	709	674	171	1,742	1,261	462	1,603	263	487	604	817
8	Kol	430	415	105	6,887	2,169	241	2,721	195	321	407	496
9	Lob	682	643	164	1,038	0,968	363	2,25	347	467	629	831
10	Lot	608	593	178	2,029	1,163	347	1,595	289	373	502	737
11	Oli	655	619	166	3,005	1,503	396	2,066	270	464	591	763
12	Per	728	717	97	7,528	1,662	504	3,242	242	626	720	811
13	Pon	683	642	185	2,129	1,293	395	1,832	309	447	597	813
14	Bil	710	657	188	3,078	1,720	484	1,523	224	501	598	774
15	Syr	704	646	215	1,952	1,325	318	2,154	406	440	622	868
16												

Рис. 1. Таблиця “об’єкт–властивість”

Нормування значень класифікаційних ознак

Значення величин даних таблиці “об'єкт–властивість” переважно є дуже неоднорідними внаслідок значного розкиду величин значень одиниць вимірювання, що робить неможливим обґрунтоване подання використовуваних класифікаційних показників-ознак в одному масштабі.

Інакше кажучи, відстань між точками, які характеризують положення об'єкта в просторі його ознак, залежить від вибраного масштабу. А це означає, що результати кластерного аналізу, проведеного на підставі матриці відстаней (близькостей), створеної за такими значеннями ознак таблиці "об'єкт–властивість", не відповідатимуть істинній класифікації. Це пов'язано з тим, що ієрархічний агломеративний кластерний аналіз проводиться на підставі матриці близькостей і є доволі чутливим до розкиду величин значень відстаней.

Тому для уникнення такої неоднорідності значення в таблиці "об'єкт–властивість", як правило, попередньо нормують, тобто нормуванню підлягають усі значення кожного виду ознак за всіма об'єктами – нормують значення в стовпчиках кожної властивості.

Отже, в результаті виконаних процедур на $i+1$ аркуші побудовано таблицю "об'єкт–властивість" групи операторів з відповідними індивідуальними класифікаційними ознаками.

Способи нормування різні, серед них найпоширеніші такі: стандартизації, редукування [8] та приведення даних до інтервалу [0, 1]. Перший з них грунтується на перетворенні (z - перетворенні) за такою формулою $z_i = \frac{x_i - \overline{x}}{s}$, де z_i – перетворене значення; x_i – поточне значення, s – середньоквадратичне відхилення. В результаті цього перетворення середнє значення $\overline{z} = 0$, а стандартне відхилення $s_z = 1$. За другим способом [8] використовують властивість: якщо величина X має розподіл Вейбула з параметрами $alfa, beta$ і x_0, то величина $Y = \frac{X - x_0}{beta}$ також має розподіл Вейбула з параметрами $alfa$, $\overline{x} = 0$ і $s = 1$. У роботі нормування відповідає приведенню даних до одиничного відрізка і виконане так: $x_{i\,норм} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$, де $x_{\min}$ і $x_{\max}$ – найбільше і найменше значення вибірки. В результаті нормування значення величин ознак об'єктів у таблиці "об'єкт–властивість" знаходитиметься в межах $0 \le x_i \le 1$, що й відображено на рис. 2.

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		Серед	Медіа	СтВід	Эксце	Асиме	Мінім	alfa	beta	tau 1	moda	tau 2
3	Bur	1,000	1,000	1,000	0,431	0,240	1,000	0,572	1,000	1,000	1,000	1,000
4	Cub	0,099	0,076	0,264	0,059	0,327	0,112	0,115	0,064	0,051	0,057	0,098
5	Cup	0,541	0,537	0,537	0,275	0,084	0,501	0,703	0,614	0,538	0,548	0,544
6	Hav	0,222	0,219	0,321	0,104	0,199	0,253	0,201	0,172	0,188	0,192	0,226
7	Hod	0,192	0,182	0,330	0,108	0,243	0,231	0,047	0,122	0,144	0,140	0,189
8	Kol	0,000	0,000	0,039	0,901	1,000	0,000	0,697	0,000	0,000	0,000	0,000
9	Lob	0,173	0,160	0,299	0,000	0,000	0,128	0,423	0,272	0,127	0,158	0,198
10	Lot	0,122	0,125	0,362	0,153	0,162	0,111	0,042	0,168	0,045	0,068	0,142
11	Oli	0,155	0,143	0,308	0,303	0,445	0,162	0,316	0,134	0,124	0,131	0,158
12	Per	0,205	0,212	0,000	1,000	0,578	0,275	1,000	0,084	0,265	0,223	0,186
13	Pon	0,174	0,159	0,393	0,168	0,270	0,161	0,180	0,204	0,110	0,135	0,187
14	Bil	0,192	0,170	0,404	0,314	0,626	0,254	0,000	0,052	0,157	0,136	0,164
15	Syr	0,188	0,162	0,526	0,141	0,297	0,081	0,367	0,377	0,103	0,153	0,219
16												

Рис. 2. Нормовані значення класифікаційних ознак

Побудова матриці близькостей

Для побудови матриці близькостей або "відстаней" між m об'єктами, яка є квадратною матрицею розміром $m \times m$, за допомогою табличного процесора Ms Excel вибирають метрику для обчислення відстані (близькості, подібності), яка враховує значення їхніх класифікаційних ознак, використовуючи дані таблиці "об'єкт–властивість", яка має розмір $m \times n$, де m – кількість об'єктів, а n – кількість ознак.

При побудові матриці близькості неминуче виникає задача вибору міри близькості. Якщо розглядати об’єкти як елементи метричного простору, то як функції відстані можна використати метрику цього простору. Найчастіше відстань між об’єктами вимірюють в евклідовій метриці, яку ще називають евклідовою відстанню. Евклідова метрика є найбільш узгодженою з нашими інтуїтивними уявленнями про близькість цих об’єктів і подається такою формулою

$$D\left(x_i, x_j\right)_E = \sqrt{\sum_{k=1}^{q}\left(x_{ik} - x_{jk}\right)^2},$$

де $k = 1, 2, \mathbf{K}, q$ – номери ознак; q – їхня кількість; а i, j – номери об’єктів.

Оскільки матриця близькостей – це матриця значень попарно обчислених відстаней між об’єктами, то обчислення можна провести за допомогою двох однакових таблиць “об'єкт–властивість”, прийнявши першу за таблицю-оригінал, а другу – за таблицю-копію. Фіксуючи значення в таблиці-оригіналі двома символами $, тобто фіксуючи назву стовпчика і номер стрічки, а в таблиці копії одним символом $ лише назву стовпчика, за допомогою автозаповнення будуємо, починаючи із вказаної комірки, стовпчик матриці близькостей. Очевидно, що у вказаній комірці має бути формула для обчислення відстані. Нижче подано послідовність кроків побудови матриці близькостей.

Побудову матриці близькості здійснено так.

1. На останньому робочому аркуші, під розміщеною таблицею “об'єкт–властивість” вміщуємо через 2 – 3 рядки таблицю її нормованих значень, яку назвемо таблицею-оригіналом. Оскільки ця таблиця містить формули нормування, усунемо ці зв’язки так. Копіюємо всі значення і вставляємо їх на ті ж самі місця шляхом “Спеціальна вставка – значення і формати чисел – ОК”. У результаті будь-яке значення таблиці буде звичайним числом.

2. Праворуч від таблиці-оригіналу, на відстані одного стовпчика розміщують її таблицю-копію так, щоб номери рядків в обох були однакові. Наприклад, якщо перший елемент таблиці–оригіналу розміщено в комірці B3, то перший елемент таблиці-копії при $n = 11$ має бути розміщений у комірці N3.

3. Через кілька рядків під таблицями, враховуючи розміри матриці близькості $m \times m$, а в нашому випадку $m = 13$, визначають комірку для її першого елемента, наприклад, B20, в яку вводять формулу обчислення евклідової метрики, яка в Екселі для цієї задачі має вигляд:

```
=КОРЕНЬ(СУММ(($B$3-$N3)^2+($C$3-$O3)^2+($D$3-$P3)^2+($E$3-$Q3)^2+
                ($F$3-$R3)^2+($G$3-$S3)^2+($H$3-$T3)^2+($I$3-$U3)^2+
                    ($J$3-$V3)^2+($K$3-$W3)^2+($L$3-$X3)^2)).
```

Суть формули в тому, що близькість між об'єктами обчислюють за значеннями елементів у рядках, тому в цій комірці будуть нулі, оскільки це відстань між першим і ним самим об’єктом.

4. Реалізуємо формулу кліком ОК і, використовуючи автозаповнення, будуємо перший стовпчик матриці близькості, тобто заповнюємо значеннями відстаней всі m комірок, тобто обчислюємо значення в ст. C20:C32.

5. Копіюємо формулу з комірки C20 в комірку D20 так: “копіювати – спеціальна вставка – формули і формати чисел”. Двома кліками мишки активізуємо формулу в цій комірці, в результаті чого кольоровими рамками в обох таблицях буде виділено комірки перших рядків.

6. Перетягуємо рамки комірок першого рядка таблиці-оригіналу відповідно в другий рядок, так що виділеними рамками тепер будуть другий рядок таблиці оригіналу і перший рядок таблиці - копії. Реалізуємо цю формулу кліком ОК і автозаповненням формуємо другий стовпчик матриці близькостей.

7. Копіюємо формулу з комірки D20 в комірку E20: “копіювати – спеціальна вставка – формули і формати чисел”. Двома кліками мишки активізуємо формулу в цій комірці, в результаті чого кольоровими рамками в таблиці-оригінал будуть виділені комірки другого рядка, а в таблиці-копії – першого.

8. Перетягуємо рамки комірок другого рядка таблиці-оригіналу відповідно в наступний третій рядок, так що виділеними рамками тепер будуть третій рядок таблиці оригіналу і перший рядок таблиці-копії. Реалізуємо формулу і автозаповненням формуємо третій стовпчик матриці близькостей.

9. Аналогічно чинимо під час обчислення значень решти стовпчиків матриці близькостей. Після перетягнення рамок в останній рядок таблиці-оригіналу і реалізації формули отримуємо матрицю близькостей між цими об'єктами за цих властивостей.

10. Після обчислення останнього стовпчика O20:O32 у полі C20:O32 цього аркуша буде сформовано квадратну $m \times m$ матрицю попарних відстаней (близькостей) між даними об’єктами.

Ця матриця за своєю специфікою є симетричною відносно головної діагоналі, на якій розташовано нулі.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
17														
18		Bur	Cub	Cup	Hav	Hod	Kol	Lob	Lot	Oli	Per	Pon	Bil	Syr
19		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
20	*1*	**0,000**	2,613	1,315	2,253	2,375	2,956	2,357	2,512	2,395	2,458	2,350	2,408	2,254
21	*2*	2,613	**0,000**	1,436	0,386	0,282	1,265	0,537	0,258	0,382	1,393	0,297	0,497	0,530
22	*3*	1,315	1,436	**0,000**	1,073	1,239	1,890	1,092	1,349	1,215	1,415	1,176	1,378	1,060
23	*4*	2,253	0,386	1,073	**0,000**	0,193	1,384	0,371	0,336	0,382	1,306	0,202	0,547	0,407
24	*5*	2,375	0,282	1,239	0,193	**0,000**	1,387	0,498	0,227	0,401	1,397	0,199	0,450	0,486
25	*6*	2,956	1,265	1,890	1,384	1,387	**0,000**	1,477	1,376	1,014	0,778	1,285	1,143	1,308
26	*7*	2,357	0,537	1,092	0,371	0,498	1,477	**0,000**	0,481	0,570	1,357	0,419	0,865	0,421
27	*8*	2,512	0,258	1,349	0,336	0,227	1,376	0,481	**0,000**	0,443	1,436	0,224	0,551	0,467
28	*9*	2,395	0,382	1,215	0,382	0,401	1,014	0,570	0,443	**0,000**	1,057	0,286	0,400	0,413
29	*10*	2,458	1,393	1,415	1,306	1,397	0,778	1,357	1,436	1,057	**0,000**	1,295	1,288	1,286
30	*11*	2,350	0,297	1,176	0,202	0,199	1,285	0,419	0,224	0,286	1,295	**0,000**	0,464	0,304
31	*12*	2,408	0,497	1,378	0,547	0,450	1,143	0,865	0,551	0,400	1,288	0,464	**0,000**	0,656
32	*13*	2,254	0,530	1,060	0,407	0,486	1,308	0,421	0,467	0,413	1,286	0,304	0,656	**0,000**

Рис. 3. Матриця близькостей між об’єктами

Матриця близькостей дає кількісні значення “відстані” між об’єктами, за якими будують відповідну діаграму розбиття сукупності об’єктів на кластери.

Обчислення параметрів дендрограми

Основним призначенням кластерного аналізу є структурування даних, тобто виділення в них певної структури, яка полягає у розбитті множини об’єктів (елементів, даних) на певні групи, причому об’єкти в кожній групі є більш подібними між собою, ніж з рештою елементів з інших груп. Крім того, кластерний аналіз забезпечує обробку значних обсягів даних, стискаючи їх до невеликої кількості однорідних в тому чи іншому сенсі масивів. Важливе значення кластерний аналіз має стосовно груп часових послідовностей, оскільки забезпечує виділення однорідних періодів в одній послідовності та групування послідовності з подібною динамікою.

На відміну від алгебричного поняття ***розбиття множини*** на підмножини за відношенням еквівалентності, коли сукупність неперетинних підмножин еквівалентних елементів утворює фактор множину, в кластерному аналізі фігурує два відношення: відношення подібності, за яким будують матрицю подібності елементів (попередньо обґрунтувавши вибір відстані), та відношення стратегії об’єднання об’єкта з об’єктом, об’єкта з підгрупою та підгрупи з підгрупою. Формально задача кластерного аналізу подається так.

В інформаційній технології алгоритм ієрархічного агломеративного кластерного аналізу забезпечує вимоги “гіпотези компактності” і, на відміну від кластеризації методом k-середніх, дає

можливість візуального вибору кількості кластерів, не потребуючи нових обчислень. Такий вибір кластерів здійснюється на підставі дендрограми вибором відстані між кластерами. Визначення параметрів дендрограми в цього типу кластерному аналізі зводиться до об'єднання об'єктів, поданих матрицею близькостей, у групи, перерахунком відстаней між двома об'єктами, об'єктом і групою і двома групами. Кількість таких перерахунків є на одиницю менша від кількості об'єктів, причому самі перерахунки виконуються за одним і тим самим алгоритмом.

У математичному плані задача класифікації даних, тобто елементів формулюється як задача побудови розбиття елементів множини даних на деяке наперед задане чи відшукуване під час аналізу число не порожніх попарно неперетинних підмножин (класів) елементів.

Процедура, яка становить суть ієрархічної класифікації, полягає в тому, що відібрані для класифікації об'єкти даних разом з їх характеристичними ознаками зведено у таблицю "об'єкт–властивість" розміром $M \times N$, де M – кількість об'єктів, а N – кількість характеристичних ознак. Оскільки розмірність та величини значень ознак є різними, усі значення цієї таблиці нормують і використовують як основу для побудови матриці відстаней (подібності, схожості). Матриця відстаней $D = \|d_{ij}\|$ має розмірність M^2, а d_{ij}, $i, j = \overline{1, M}$, $i \neq j$ – елемент матриці, який відповідає кількісному значенню величини відстані між об'єктами i та j. На початку роботи процедури кожен об'єкт вважають окремим кластером. На першому кроці проглядається матриця D і відшукується її мінімальний елемент, тобто два об'єкти, наприклад, m_i і m_j, між якими відстань є найменшою, тобто

$$D\left(m_i, m_j\right) = \min_{i,j} d_{ij}. \tag{1}$$

Визначені умовою (1) об'єкти m_i і m_j об'єднуються в один $(M+1)$-й клас, стовпчик і рядок, що належить будь-якому з них з матриці D викреслюються, а замість другого вставляють стовпчик і рядок спеціально перерахованих значень, а саме значень відстаней від щойно створеного $(M+1)$-го класу до всіх решти ще необ'єднаних класів). Потім знову шукають найменшу відстань між двома об'єктами, і процедура повторюється. Якщо на k-му кроці всі елементи множини даних об'єднуються в один загальний клас, то на цьому процедура припиняється, якщо ні, то переходять до виконання наступного $(k+1)$-го кроку на основі модифікованої останнім об'єднанням матриці D.

Отже, спочатку кожен об'єкт розглядається як окремий кластер. Потім два найближчі кластери Q_i та Q_j об'єднують в один Q_{i+j}. Під час об'єднання виникає потреба обчислити відстань від нового кластеру Q_{i+j} до всіх інших. Для перерахунку значень об'єднуваних стовпчиків використовують формулу Ланса–Уільямса [9]:

$$D\left(Q_{i+j}, Q_m\right) = a_i D\left(Q_i, Q_m\right) + a_j D\left(Q_j, Q_m\right) + b\, D\left(Q_i, Q_j\right) + g \left| D\left(Q_i, Q_m\right) - D\left(Q_j, Q_m\right) \right|,$$

де $D(\cdot,\cdot)$ – функція відстані; Q_m – кластер, відстань до якого обчислюють, причому ($m \neq i, j$); a_i, a_j, b, g – числові параметри.

Значення параметрів a_i, a_j, b, g відповідають різним способам обчислення відстаней між кластерами і дають різні результати розбиття, тому їх часто називають стратегіями, до яких зараховують такі: найближчого сусіда, найдальшого сусіда, групового середнього, центроїдну, гнучку []. У цій роботі використано гнучку стратегію з такими параметрами: $a_i = a_j = 0.625$, $b = -0.25$, $g = 0$.

Процедура побудови дендрограми. Визначимо на аркуші з матрицею близькості область для перерахунку значень об'єднуваних елементів в одну групу і позначимо її для зручності так: B35 – "лівий", C35 – "правий", D35 – "новий", а також область характеристик дендрограми: J35 –

“номер”, тобто номер об’єднання, K35 – “елементи” – ті, які об’єднуються, L35 – “відстань”, за якою вони об’єднуються. Крім того, для зручності замінимо імена об’єктів цифрами від 1 до 13 у тій самій послідовності. Процедуру можна описати так.

1. Шукають в матриці близькостей два об’єкти, між якими є найменша відстань. Такими об’єктами на початку аналізу є 4-й і 5-й, тобто **Hav-1** і **Hod-1**. Значення відстані між ними, яке дорівнює 0,193, подані в комірках E24 і F23. Оскільки об’єднуються два елементи матриці, назвемо їх відповідно “лівим” та “правим”, а перерахований стовпчик – “новим”.

2. Виділяємо лівий стовпчик і перетягуємо його область перерахунку під назву “лівий” через один рядок, тобто його перший елемент розміщений в комірці B37. Аналогічно чинимо з правим, перший елемент якого розміщується в комірці C37.

3. Місце правого стовпчика в матриці близькостей ліквідуємо шляхом виділення усіх стовпчиків разом з їх номерами, що стоять праворуч від нього, і перетягуємо їх вліво. Аналогічно чинимо з рядком, що відповідає правому стовпчику – виділяємо рядки під цим рядком разом з їхніми номерами і перетягуємо їх на один рядок вище. Замінюємо номер лівого рядка на новий: 4 на 14.

4. Заповнюємо область характеристик діаграми. До комірки J37 вводимо номер об’єднання в групу 4 і 5 елементів, тобто 14; в комірці K37 вказуємо, які об’єкти чи групи об’єднуються, тобто 4+5, а комірку L37 записуємо відстань між цими об’єктами, тобто 0,193.

5. До комірки D37 вводимо формулу для перерахунку значень об’єднуваних елементів за даної стратегії, в нашому прикладі вибрано гнучку стратегію [1], яку можна використати для будь-якої міри відмінностей:

=0,625*(B37+C37)-0,25*C40,

де B37 і C37 перші елементи лівого і правого стовпчиків, а C40 фіксоване значення відстані між об’єднуваними елементами, взяте з правого стовпчика.

6. Використовуючи автозаповнення, обчислюємо дані “нового” стовпчика D37:D49. Відкидаємо рядок з нулем в правому стовпчику, тобто B41:D41, і найменше значення нового стовпчика, яке знаходиться навпроти нуля лівого стовпчика, тобто значення в комірці D40, яке дорівнює 0,074, замінюємо на нуль.

7. Виділяємо і копіюємо новий стовпчик і вставляємо його в матрицю близькостей, позбуваючись зв’язку з формулами завдяки використанню “Спеціальна вставка – значення і формати чисел – ОК” для стовпчика, у нашому випадку 14, і “Спеціальна вставка – значення і формати чисел – транспонувати – ОК” для 14-го рядка. В результаті розмірність матриці близькостей зменшилась на одиницю.

8. Повторюємо п.1 щодо пошуку в існуючій матриці близькостей двох об’єктів, відстань між якими є найменшою. Такими об’єктами є 14 і 11, відстань між якими дорівнює 0,203. стовпчик 14-й є лівим, а 11-й – правим. Як в п. 2 – лівий стовпчик перетягуємо в область перерахунку під назву “лівий” через один рядок, тобто його перший елемент розміщений в комірці B37. Аналогічно чинимо з правим, перший елемент якого розміщується в комірці C37. У стовпчику D37:D48 візуалізується помилка. Активізуємо комірку D37, в якій помилка має такий вигляд

=0,625*(#ССЫЛКА!+#ССЫЛКА!)-0,25*#ССЫЛКА!

Виділяємо першу “#ССЫЛКА!” і замінюємо її на комірку B37, другу “#ССЫЛКА!” на комірку C37, третю на комірку правого стовпчика з найменшим значенням, фіксуючи її кнопкою клавіатури **F4**. Використовуючи автозаповнення, обчислюємо дані “нового” стовпчика D37:D48. Відкидаємо в стовпчиках: B37:B48, C37:C48, D37:D48 рядок з нулем у правому стовпчику, а найменше значення нового стовпчика, яке знаходиться навпроти нуля лівого стовпчика, замінюємо на нуль.

9. Повторюємо п.1 щодо пошуку в існуючій матриці близькостей двох об’єктів, відстань між якими є найменшою. Такими об’єктами є 14 і 11, відстань між якими дорівнює 0,203, стовпчик 14-й є лівим, а 11-й – правим. Як в п. 2 – лівий стовпчик перетягуємо в область перерахунку під назву “лівий” через один рядок, тобто його перший елемент розміщений у комірці B37. Аналогічно

чинимо з правим, перший елемент якого розміщується в комірці C37. У стовпчику D37:D48 візуалізується помилка. Активізуємо комірку D37, в якій помилка має такий вигляд

=0,625*(#ССЫЛКА!+#ССЫЛКА!)-0,25*#ССЫЛКА!

Виділяємо першу “#ССЫЛКА!” і замінюємо її на комірку B37, другу “#ССЫЛКА!” на комірку C37, третю на комірку правого стовпчика з найменшим значенням, фіксуючи її кнопкою клавіатури **F4**. Використовуючи автозаповнення обчислюємо, дані “нового” стовпчика D37:D48. Відкидаємо в стовпчиках: B37:B48, C37:C48, D37:D48 рядок з нулем у правому стовпчику, а найменше значення нового стовпчика, яке знаходиться навпроти нуля лівого стовпчика замінюємо на нуль.

10. Аналогічно, повторюючи пункти 1 – 7, визначаємо відстані між іншими об’єднуваними елементами. Зменшення розмірності матриці відбувається до значення 2×2. Значення в цій матриці є єдине і означає відстань між двома найбільшими кластерами. Його номер визначається як $2m-1$.

D37 fx =0,625*(B37+C37)-0,25*C40

	A	B	C	D	E	F	G	H	I	J	K	L	M
16													
17				Таблиця подібності об'єктів									
18													
19		***1***	***2***	***3***	***14***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***
20	***1***	**0**	2,613	1,315		2,956	2,357	2,512	2,395	2,458	2,350	2,408	2,254
21	***2***	2,613	**0**	1,436		1,265	0,537	0,258	0,382	1,393	0,297	0,497	0,530
22	***3***	1,315	1,436	**0**		1,890	1,092	1,349	1,215	1,415	1,176	1,378	1,060
23	***14***												
24	***6***	2,956	1,265	1,890		**0**	1,477	1,376	1,014	0,778	1,285	1,143	1,308
25	***7***	2,357	0,537	1,092		1,477	**0**	0,481	0,570	1,357	0,419	0,865	0,421
26	***8***	2,512	0,258	1,349		1,376	0,481	**0**	0,443	1,436	0,224	0,551	0,467
27	***9***	2,395	0,382	1,215		1,014	0,570	0,443	**0**	1,057	0,286	0,400	0,413
28	***10***	2,458	1,393	1,415		0,778	1,357	1,436	1,057	**0**	1,295	1,288	1,286
29	***11***	2,350	0,297	1,176		1,285	0,419	0,224	0,286	1,295	**0**	0,464	0,304
30	***12***	2,408	0,497	1,378		1,143	0,865	0,551	0,400	1,288	0,464	**0**	0,656
31	***13***	2,254	0,530	1,060		1,308	0,421	0,467	0,413	1,286	0,304	0,656	**0**
32													
33													
34													
35		**лівий**	**правий**	**новий**		**номер елемент відстань**							
36		**4**	**5**										
37		2,253	2,375	2,845		14	4+5	0,193					
38		0,386	0,282	0,370									
39		1,073	1,239	1,397									
40		0,000	**0,193**	0,072									
41		0,193	0,000	0,072									
42		1,384	1,387	1,684									
43		0,371	0,498	0,495									
44		0,336	0,227	0,304									
45		0,382	0,401	0,441									
46		1,306	1,397	1,641									
47		0,202	0,199	0,203									
48		0,547	0,450	0,575									
49		0,407	0,486	0,510									

Рис. 4. Організація даних на листі Excel для визначення параметрів дендрограми

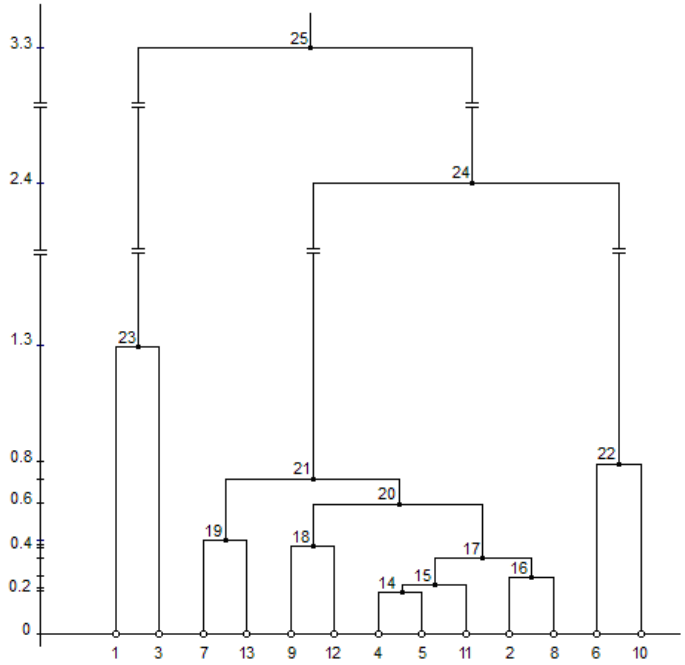
Побудова дендрограми

На жаль, графічне забезпечення Microsoft Excel не має можливості безпосередньо побудувати дендрограму, оскільки її побудова пов’язана лише не лише з масштабом відстаней, але і з місцем

об’єктів, які тепер належать конкретним кластерам. У [10] наведено загальні принципи основних методів побудови дендрограм, оскільки математичні алгоритми та їхні програмні реалізації є доволі складними, за невеликої кількості об’єктів та їх ознак (практично декілька десятків) користуються ручним методом. Хоча цей підхід є доволі рутинним, зате через декілька ітерацій отримують відповідну дендрограму. За ручним способом процес побудови дендрограми розпочинається з кореня дерева. Наприклад, номер 25 вказує на об’єднання двох останніх груп 23 і 24, тобто можемо побудувати в довільному масштабі дві прямокутні гілки. Своєю чергою, номер 23 є групою з двох початкових об’єктів і більше розгалужень він не має. Натомість вузол 24 об’єднує 21 і 22 групи, причому 22 група є об’єднанням 6-го і 10-го об’єктів. Аналогічно аналізуємо групу 21. Після побудови такого дерева стає зрозумілим потрібний порядок об’єктів, надалі лише корегують положення гілок для усунення перетинів. Останнім кроком є геометричне масштабування вертикальних відрізків для відтворення величин відстаней. Далі приводять вертикальні відрізки у відповідність зі шкалою відстаней. У результаті дендрограма набуває вигляду, як на рис. 5.

Параметри дендрограми		
номер	елемент	відстань
14	4+5	0,193
15	14+11	0,203
16	2+8	0,258
17	15+16	0,338
18	12+9	0,4
19	13+7	0,421
20	18+17	0,593
21	19+20	0,702
22	6+10	0,778
23	3+1	1,315
24	21+22	2,36
25	23+24	3,308

а



б

Рис. 5. Параметри (а) та графічне зображення (б) дендрограми

Маємо чітко виражені три кластери, які характеризується параметрами, наведеними в таблиці.

Характеристики кластерів

№ з/п	Кількісні та якісні характеристики	Кластери		
		Перший Об’єкти 1, 3	Другий Об’єкти 2, 4, 5, 7, 8, 9, 11, 12	Третій Об’єкти 6, 10
1	2	3	4	5
1	Середнє арифметичне	1510	676	579
2	Стандартне відхилення	270	177	101
3	Медіана	1510	636	566
4	Мінімальне значення	958	400	373
5	Асиметрія	1,163	1,311	1,916

Продовження таблиці

1	2	3	4	5
6	Ексцес	3,332	2,012	7,207
7	Tay1	1206	455	476
8	Мода	1493	590	564
9	Tay2	1805	794	654
10	*alfa*	2,619	1,846	2,982
11	*beta*	646	292	219

Ці три кластери отримано за умови, що відстань між ними перевищує значення 1,3 за шкалою ординат. Очевидно, вибираючи іншу відстань, можемо мати іншу кількість кластерів. У цьому випадку позначимо ці кластери як перший, другий і третій. У результаті розподіл об'єктів за кластерами є такий:

- перший кластер – об'єкти 1 і 3;
- другий кластер – об'єкти 7, 13, 9, 12, 4, 5, 11, 2, 8;
- третій кластер – об'єкти 6 і 10.

Їхні усереднені властивості, визначені за таблицею "об'єкт–властивість", наведено у таблиці. За цими результатами в сенсі оперативності розпізнавання об'єктів заданого класу на аерокосмічних зображеннях найкращими є оператори третього кластера, хоча вони доволі сильно відрізняються між собою. Середнє значення часу розпізнавання для них – 579 ms, стандартне відхилення – 100 ms, мода – 563 ms.

Оператори, що ввійшли до другого кластера, мають дещо гірші показники, проте, як показує дендрограма, оператори 4, 5, 11, 2, 8 можуть бути представлені і окремим підкластером.

Найгірші показники в операторів першого кластера.

Висновок

Застосування процедури ієрархічного агломеративного кластерного аналізу дає, на відміну від інших методів, весь спектр кластерів, а вже вибрати ту чи іншу кількість кластерів, тобто рівень дерева, має сам дослідник.

У результаті виконаної процедури вирішено такі завдання:

– побудовано таблицю "Об'єкт–властивість" та приведено її значення до [0, 1] інтервалу;

– приведений алгоритм побудови таблиці близькостей в середовищі табличного процесора Ms Microsoft Excel 2003 матриці близькостей шляхом автозаповнення її стовпців за рахунок обчислення двох однакових таблиць "об'єкт–властивість", з яких одна є оригінальною, а друга – її копією;

– приведений алгоритм визначення параметрів дендрограми ієрархічного агломеративного кластерного аналізу в середовищі Ms Microsoft.

Крім того, наведено поділ об'єктів на три кластери та показано усереднені характеристики для вибраних кластерів. Загалом приведена процедура ієрархічного агломеративного кластерного аналізу хоча і є достатньо простою, забезпечує повне або часткове проведення інших модифікацій кластерного аналізу, які використовують матрицю близькостей, навіть якщо дано подані одновимірними вибірками випадкових значень.

1. Мягченко О.П. Безпека життєдіяльності людини і суспільства: навч. посібник – К.: Центр учбової літератури, 2010. – 384 с. 2. Шафран Л.М., Псядло Є.М., Чумаєва Ю.В., Огуленко А.П., Стадник А.Л. Пути повышения надежности психофизиологического профессионального отбора операторов Запорожской АЭС // Український журнал з проблем медицини праці. – 2012. – № 4(33). –

С. 48 – 55. 3. Савченко Т. Н. Применение методов кластерного анализа для обработки данных психологических исследований // Экспериментальная психология, 2010, том 3, № 2, с. 67–85. 4. Aravind H., C. Rajgopal, K. P. Soman. A Simple Approach to Clustering in Excel // International Journal of Computer Applications (0975 – 8887), Volume 11 – No.7, December 2010. 5. Горчаков А.А. Математический апарат для инвестора // Аудит и финансовій анализ. – ІІІ кв. 1997. – С. 1 – 57. 6. Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни / Пер. с англ. О.В. Селезнева. – М.: инансы и статистика, 1988. – 191 с. 7. Вадзинский Р.Н. Справочник по вероятностным распределениям. – Спб.: Наука, 2001. – 295 с. 8. Таблицы по математической статистике / П. Мюллер, П. Нойман, Р. Шторм; пер. с нем. и предисл. В.М. Ивановой. – М.: Финансы и статистика, 1982. – 278 с. 9. Уиллиамс У.Т., Ланс Дж. Н. Методы иерархической классификации. - в кн. Статистические методы для ЭВМ / Под ред. Э. Рэлстона, Г.С. Уилфа: Пер с англ./ Под ред. М.Б. Малютова. – М.: Наука, 1986. – 464 с. 10. Бутвиловский А.В. Основные методы молекулярной эволюции: монография / А.В. Бутвиловский, Е.В. Барковский, В.Э. Бутвиловский, В.В. Давыдов, Е.А. Черноус, В.В. Хрусталев; под общ. ред. проф. Е.В. Барковского. – Мн.: 2009. – 210 с.

004.032.26; 004.852; 004.94

П.О. Кравець
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

МОДЕЛЬ СТОХАСТИЧНОЇ ГРИ НЕЙРОАГЕНТІВ

Розроблено нейроагентну ігрову модель колективного прийняття рішень в умовах невизначеності. Виконано формулювання стохастичної гри та для її розв'язування використано адаптивні методи навчання штучних нейронних мереж без учителя. Розроблено ігровий алгоритм та програмну модель нейроагентного прийняття рішень. Збіжність стохастичної гри нейроагентів підтверджено результатами комп'ютерного експерименту. Досліджено впливи параметрів ігрової моделі на швидкість навчання нейроагентів.

Ключові слова: колективне прийняття рішень, умови невизначеності, стохастична гра нейроагентів, адаптивні методи навчання.

The neuroagent game model of collective decision-making in the conditions of uncertainty is developed. The formulation of stochastic game is executed. Adaptive learning methods of artificial neural networks without the teacher are used for the game solving. The convergence of neuroagent stochastic game is confirmed by results of computer experiment. Influences of parameters of game model on the neuroagent learning rate are investigated.

Key words: collective decision making, uncertainty conditions, neuroagent stochastic game, adaptive learning methods.

Вступ до проблеми колективного прийняття рішень в умовах невизначеності

Для розв'язування задач розподіленого керування та прийняття рішень у технічних, економічних, інформаційних та соціальних системах існує необхідність у колективному виборі варіантів рішень, що задовольняють одну з умов багатокритеріальної оптимальності, наприклад, Неша, Парето тощо [1–7]. Ці умови тією чи іншою мірою визначають ступінь вигідності та справедливості колективно досягнутого рішення.