**МІНІСТЕРСТВО ОСВІТИ І НАУКИ,
МОЛОДІ ТА СПОРТУ УКРАЇНИ**

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

**Методичні рекомендації
до виконання лабораторних робіт
з навчальної дисципліни**

"МАТЕМАТИЧНІ МЕТОДИ І МОДЕЛІ ДОСЛІДЖЕННЯ ЕКОНОМІЧНИХ ПРОЦЕСІВ"

**для студентів спеціальностей 8.03050803 "Оподаткування",
8.03050701 "Маркетинг", 8.14010301 "Туризмознавство"
денної форми навчання**

Харків. Вид. ХНЕУ, 2012

Затверджено на засіданні кафедри економічної кібернетики.
Протокол № 2 від 08.09.2011 р.

Укладачі: Гур'янова Л. С.
Сергієнко О. А.
Нікіфорова О. В.
Трунова Т. М.
Чаговець Л. О.

М54 Методичні рекомендації до виконання лабораторних робіт з навчальної дисципліни "Математичні методи і моделі дослідження економічних процесів" для студентів спеціальностей 8.03050803 "Оподаткування", 8.03050701 "Маркетинг", 8.14010301 "Туризмознавство" денної форми навчання / укл. Гур'янова Л. С., Сергієнко О. А., Нікіфорова О. В. та ін. – Х. : Вид. ХНЕУ, 2012. – 64 с. (Укр. мов.)

Подано лабораторні роботи, метою яких є закріплення теоретичного і практичного матеріалу, набуття навичок роботи з пакетами прикладних програм, що підтримують реалізацію різних методів та моделей дослідження економічних процесів.

Рекомендовано для студентів економічних спеціальностей.

Вступ

В умовах трансформаційної економіки значно зросла роль факторів невизначеності, неповноти інформації при прийнятті управлінських рішень. Тому для підвищення ефективності розв'язання складних питань регулювання сучасних соціально-економічних систем за цих умов виникає необхідність ширшого застосування економіко-математичних методів та моделей для дослідження перебігу економічних процесів.

Метою лабораторних занять з дисципліни "Математичні методи та моделі дослідження економічних процесів" є вивчення можливостей практичного застосування методів моделювання економічних систем, що функціонують в умовах невизначеності та набуття навичок їх застосування.

Об'єктом вивчення дисципліни є соціально-економічні системи різного рівня ієрархії.

Предметом вивчення дисципліни є математичні методи і моделі, що дозволяють досліджувати соціально-економічні системи, які функціонують в умовах нестаціонарного зовнішнього середовища.

Наукову основу дисципліни складають теоретичні методи і моделі, математичний апарат, сучасні концепції, які визначають різні підходи до моделювання складних соціально-економічних систем.

Лабораторні роботи призначені для набуття навичок здійснення багатовимірного статистичного аналізу, зокрема різних варіантів процедур багатовимірної класифікації, використання методів повної та неповної редукції, моделювання економічних процесів та інтелектуального аналізу даних, зокрема за методом дерев класифікації.

Для виконання лабораторних робіт пропонується використовувати табличні редактори *OpenOffice 3.2 Calc* або *Microsoft Office Excel 2007* та спеціалізовану програму *See5.*

Лабораторна робота № 1. Ряди та їх статистичні характеристики

Мета – закріплення теоретичного й практичного матеріалу, придбання навичок роботи з варіаційними рядами в табличних редакторах *OpenOffice 3.2 Calc* та *Microsoft Office Excel 2007*.

Завдання: необхідно провести аналіз статистичних характеристик дискретного та інтервального ряду у *Calc* або *Excel*.

1. Розрахувати статистичні характеристики ряду (середнє, дисперсію, середнє квадратичне відхилення, моду, медіану, коефіцієнти асиметрії та ексцесу).

2. Побудувати гістограму та полігон розподілу випадкової величини, зробити висновки щодо характеру закону розподілу.

3. За допомогою критеріїв Пірсона та Колмогорова – Смірнова перевірити гіпотезу про нормальний закон розподілу.

Лабораторна робота виконується в рамках вивчення теми 1 "Математико-статистична обробка вибіркових даних".

Методичні рекомендації до виконання завдання

1. Запуск *OpenOffice або Microsoft Office* і підготовка даних.

Для початку роботи в пакеті після ініціювання ярлика *OpenOffice* необхідно обрати *"Электронная таблица" (або* після ініціювання ярлика Microsoft Office обрати Microsoft Office Excel*).*

Для введення даних (тексту чи числа) потрібно перейти до комірки, набрати дані і натиснути клавішу *Enter.* Кожен елемент даних займає одну комірку поля даних. Після заповнення всіх комірок одержимо таблицю, у якій зображено дискретний варіаційний ряд (рис. 1.1.).

Адреса кожної комірки складається з назви стовпця (буква алфавіту) та номера строки. Наприклад комірка A2, комірка F23.

Якщо треба вказати на блок комірок – слід надати адресу верхньої лівої та нижньої правої комірки блоку.

Наприклад, щоб указати на необхідність роботи з даними, наведеними на рис. 1.1., слід указати адресу, за якою вони розміщені – A2:B12.

	A	B	C
1	№ спостереженя	X	
2	1	313,10	
3	2	1387,50	
4	3	438,40	
5	4	425,50	
6	5	290,50	
7	6	124,60	
8	7	262,90	
9	8	330,20	
10	9	470,20	
11	10	14772,90	
12	11	11854,00	
13	12	10735,20	

Рис. 1.1. **Вихідні дані**

2. Розрахунок основних статистичних характеристик для дискретного ряду. Для розрахунку основних статистичних характеристик дискретного варіаційного ряду, необхідно скористатися Майстром функцій, діалогове вікно якого зображено на рис. 1.2 для пакета *OpenOfficeCalc* чи на рис. 1.3 для *Microsoft Office Excel*.

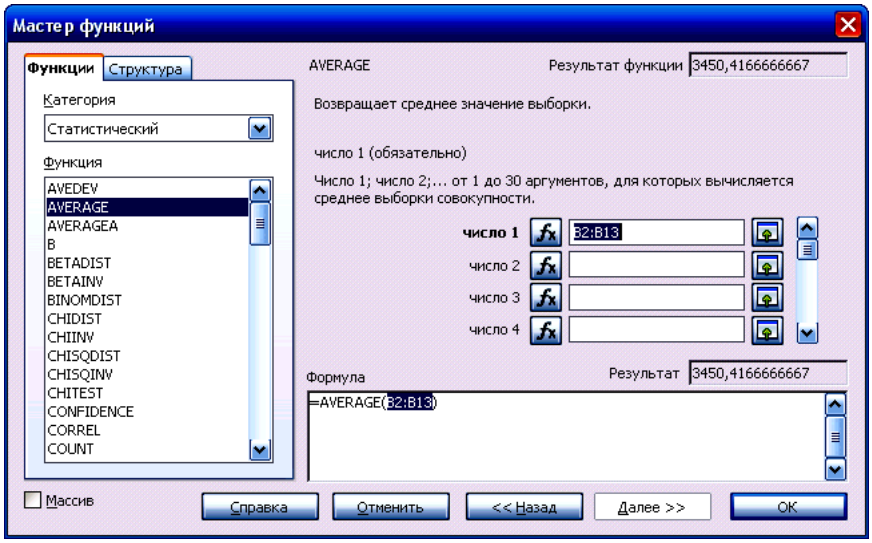


Рис. 1.2. **Діалогове вікно Майстра функцій OpenOfficeCalc**

Кожній із статистичних характеристик відповідає визначена функція, а саме: середнє значення (=AVERAGE(B2:B13)); дисперсія (=VAR(B2:B13)); стандартне відхилення (=STDEV(B2:B13)); мода (=MODE(B2:B13)); медіана (=MEDIAN(B2:B13)); коефіцієнт асиметрії (=SKEW(B2:B13)) та ексцесу (=KURT(B2:B13)).

У Microsoft Excel для розрахунку зазначених характеристик відповідають наступні функції: середнє значення – СРЗНАЧ; дисперсія – ДИСП; стандартне відхилення – СТАНДОТКЛОН; мода – МОДА; медіана – МЕДИАНА; коефіцієнт асиметрії – СКОС; коефіцієнт ексцесу – ЭКСЦЕСС.

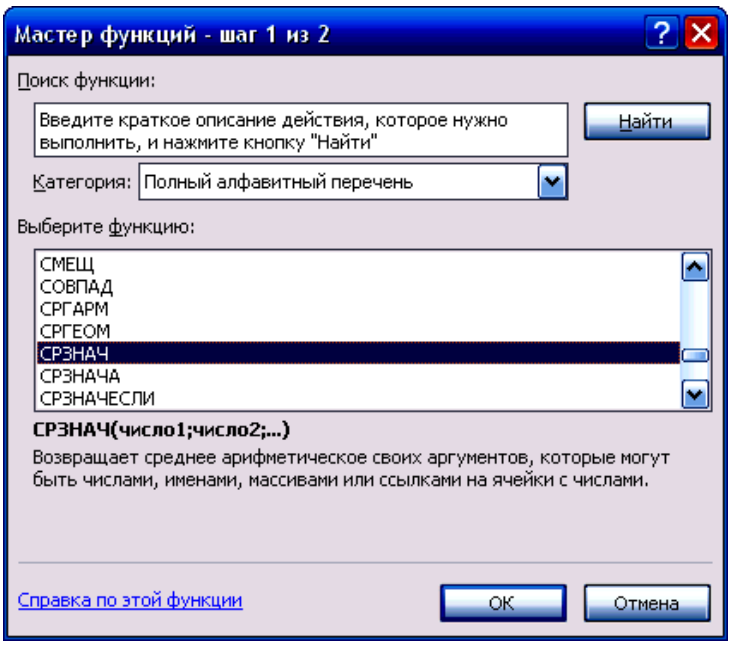


Рис.1.3. **Діалогове вікно Майстра функцій Microsoft Office Excel**

Результати розрахунку цих характеристик зображено на рис. 1.4.

	D	E
13		
14		Дискретний ряд
15	Середнє	3450,42
16	Дисперсія	30366499,27
17	Середнє квадратичне відхилення(СКВ)	5510,58
18	Мода	#ЗНАЧЕН!
19	Номер медіани	
20	Медіана	431,95
21	коефіцієнт асиметрії	1,42
22	коефіцієнт ексцесу	0,24
23		

Рис. 1.4. **Результати розрахунку основних статистичних характеристик для дискретного варіаційного ряду**

У даному випадку дискретний варіаційний ряд має множинне значення моди, тому що кожне зі значень цього ряду зустрічається однакову кількість разів.

3. Перетворення дискретного варіаційного ряду в інтервальний. Для побудови інтервального ряду необхідно розрахувати крок групування інтервалів за наступною формулою:

$$k = \frac{\max(X) - \min(X)}{n}, n = 1 + 3{,}22Lg(N), \quad (1)$$

де k – крок групування інтервалів;

X – даний дискретний варіаційний ряд, $X=(x_1,x_2,\ldots,x_N)$;

$\max(X) - \min(X)$ – розмах варіаційного ряду X;

n – кількість інтервалів;

N – кількість спостережень.

Для розрахунку елементів цієї формули необхідно скористатися функціями *Calc* MAX, MIN, ROUND (для округлення кількості інтервалів) (в *Microsoft Excel* відповідно – МАКС, МИН, ОКРУГЛ), що стають активними після введення у комірку знаку "=" у меню функцій, які зустрічаються більш часто (рис. 1.5.)

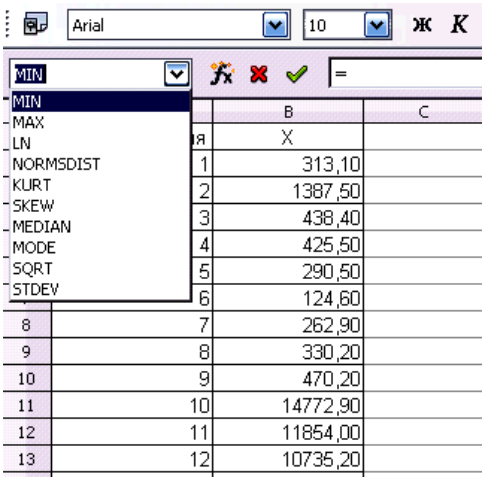


Рис. 1.5. **Меню функцій, що зустрічаються більш часто**

Після отримання проміжних результатів потрібно розрахувати крок групування, як зображено на рис. 1.6.

	A	B	C
14			
15	Максимальне значення	14772,9	
16	Мінімальне значення	124,6	
17	Розмах	14648,3	
18	n	=ROUND(1+3,22*LOG10(12) ;0)	
19	Крок групування	3662,075	
20			

Рис. 1.6. **Розрахунок кроку групування інтервалів ряду**

Формування границь інтервалів відбувається наступним чином: нижня границя першого інтервалу дорівнює мінімальному значенню дискретного варіаційного ряду (для усіх наступних інтервалів дорівнює верхній границі попереднього інтервалу); верхня границя усіх інтервалів дорівнює сумі нижньої границі відповідного інтервалу та кроку групування. Кількість інтервалів збільшується до тих пір, поки верхня границя не охопить максимальне значення ряду (рис. 1.7 – 1.8)

	A	B
15	Максимальне значення	14772,9
16	Мінімальне значення	124,6
17	Розмах	14648,3
18	n	4
19	Крок групування	3662,075
20		
21		
22		
23		
24		
25		
26	1	2
27	Нижня межа інтервалу	Верхня межа інтервалу
28	124,6	=A28+B$19

Рис. 1.7. **Розрахунок верхньої границі інтервалу**

	A	B
23		
24		
25		
26	1	2
27	Нижня межа інтервалу	Верхня межа інтервалу
28	124,6	3786,68
29	3786,68	7448,75
30	7448,75	11110,83
31	11110,83	14772,90
32	**Сума**	

Рис. 1.8. **Сформовані інтервали варіаційного ряду**

Після завершення формування границь інтервалів необхідно розрахувати частоти потрапляння до кожного інтервалу значень дискретного варіаційного ряду. Для цього можна скористатися функцією FREQUENCY (*OpenOffice*) або ЧАСТОТА (*Microsoft Excel*), обираючи у якості вихідного масиву заданий дискретний варіаційний ряд, а у якості масиву групування верхні межі побудованих інтервалів.

Остаточний вигляд інтервального ряду зображено на рис. 1.9.

	A	B	C
25			
26	1	2	3
27	Нижня межа інтервалу	Верхня межа інтервалу	Емпіричні частоти (f)
28	124,6	3786,68	9
29	3786,68	7448,75	0
30	7448,75	11110,83	1
31	11110,83	14772,90	2
32	**Сума**		**12**
33			

Рис. 1.9. **Сформований інтервальний варіаційний ряд**

4. Розрахунок основних статистичних характеристик для інтервального ряду. Для розрахунку основних статистичних характеристик для інтервального ряду необхідні наступні формули:

1. Середнє значення

$$\overline{X} = \frac{\sum x_i f_i}{\sum f_i}, \qquad (2)$$

де x_i – середина відповідного інтервалу;

f_i – частота відповідного інтервалу, $i = \overline{1,n}$;

n – кількість інтервалів.

2. Дисперсія та стандартне відхилення:

$$D(x) = \frac{\sum (x - \bar{x})^2 f_i}{\sum f_i}, \tag{3}$$

$$\sigma = \sqrt{D}. \tag{4}$$

3. Мода

$$Mo = \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} k_{mo} + x_{mo}, \tag{5}$$

де f_{mo} – частота модального інтервалу (модальний інтервал – це інтервал, якому відповідає найбільша частота ряду);

f_{mo-1} – частота інтервалу, що передує модальному;

f_{mo+1} – частота інтервалу, наступного за модальним;

k_{mo} – крок модального інтервалу;

x_{mo} – нижня границя модального інтервалу.

4. Медіана

$$Me = \frac{N_{me} - S_{me-1}}{f_{me}} k_{me} + x_{me}, \tag{6}$$

де f_{me} – частота медіанного інтервалу (медіанний інтервал – це інтервал, кумулятивна частота якого охоплює номер медіани, що розраховується за формулою $N_{me} = \frac{\sum f_i}{n}$);

S_{me-1} – кумулятивна частота інтервалу, що передує медіанному;

k_{me} – крок медіанного інтервалу;

x_{me} – нижня границя медіанного інтервалу

5. Коефіцієнт асиметрії

$$K_a = \frac{\bar{x} - Mo}{\sigma}. \tag{7}$$

6. Коефіцієнт ексцесу

$$K_э = \frac{1}{n}\frac{\sum(x-\bar{x})^4 f_i}{\sigma^4}. \qquad (8)$$

Таблиця проміжних результатів розрахунку зазначених статистичних характеристик зображена на рис. 1.10.

	A	B	C	D	E	F	G	H
25								
26	1	2	3	4	5	6	7	8
27	Нижня межа інтервалу	Верхня межа інтервалу	Емпіричні частоти (f)	Середина інтервалу (X)	X*f	(X-Xср)^2*f	Накопичені частоти S	(X-Xср)^4*f
28	124,6	3786,68	9	1955,64	17600,74	53643173,22	9	319732225931016,00
29	3786,68	7448,75	0	5617,71	0,00	0,00	9	0,00
30	7448,75	11110,83	1	9279,79	9279,79	23841410,32	10	568412846099584,00
31	11110,83	14772,90	2	12941,86	25883,73	146028638,22	12	10662181589727300,00
32	**Сума**		**12**		**52764,25**	**223513221,76**		**11550326661757900,00**
33								

Рис. 1.10. **Проміжні результати розрахунку статистичних характеристик інтервального варіаційного ряду**

Результати розрахунків статистичних характеристик для дискретного та інтервального рядів зображені на рис. 1.11

	A	B	C	D	E	F
14					Дискретний ряд	Інтервальний ряд
15	Максимальне значення	14772,9		Середнє	3450,42	4397,02
16	Мінімальне значення	124,6		Дисперсія	30366499,27	18626101,81
17	Розмах	14648,3		Середнє квадратичне відхилення(СКВ)	5510,58	4315,80
18	n	4		Мода	#ЗНАЧЕН!	1955,64
19	Крок групування	3662,075		Номер медіани		6
20				Медіана	431,95	2565,98
21				коефіцієнт асиметрії	1,42	0,57
22				коефіцієнт ексцесу	0,24	2,77
23						
24						

Рис. 1.11. **Результати розрахунку основних статистичних характеристик дискретного та інтервального рядів**

4.1. Графічне зображення дискретного та інтервального варіаційних рядів. Розглянемо графічне зображення рядів даних окремо в пакеті *OpenOfficeCalc* та *Microsoft Excel*.

4.1.1. Побудова діаграм в OpenOfficeCalc. Для більш наочного представлення варіаційних рядів доцільно використовувати графіки їх розподілу. Для цього необхідно в меню Вставка обрати "Диаграмма", після чого на екрані з'явиться меню Майстра діаграм, зображене на рис. 1.12.

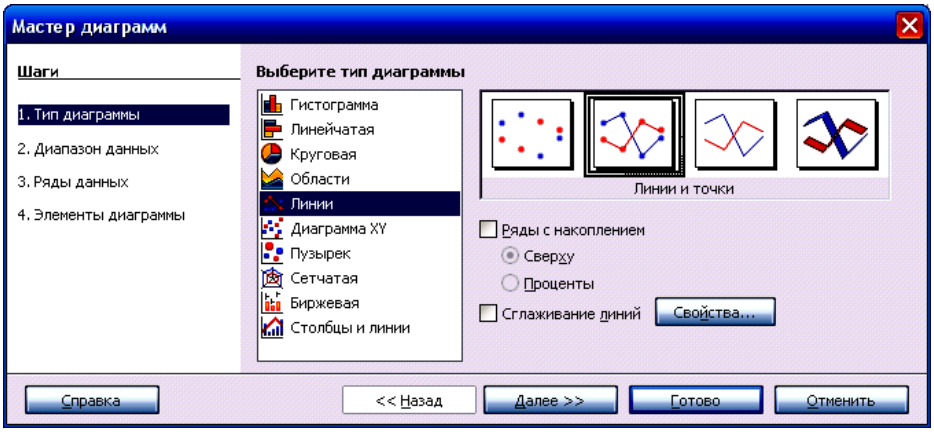


Рис. 1.12. **Майстер діаграм OpenOfficeCalc**

Для зображення дискретного ряду потрібно обрати тип діаграми Лінії (Лінії та точки) і натиснути Далее>>. Наступним кроком необхідно обрати діапазон даних, що необхідно зобразити, визначити розташування та підпис даних, як зображено на рис. 1.13. Після чого натиснути Готово. Графік зображення дискретного варіаційного ряду, у якого по осі X – номер об'єкта, а по осі Y – значення дослідженого показника представлено на рис. 1.14.

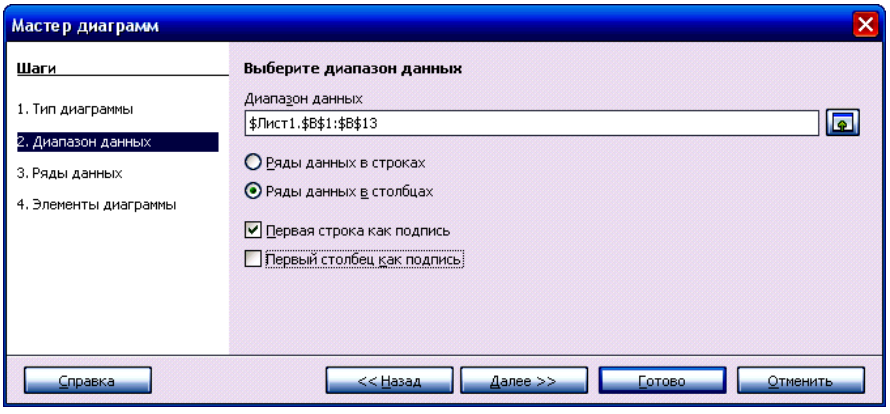


Рис. 1.13. **Вікно визначення діапазону даних OpenOfficeCalc**

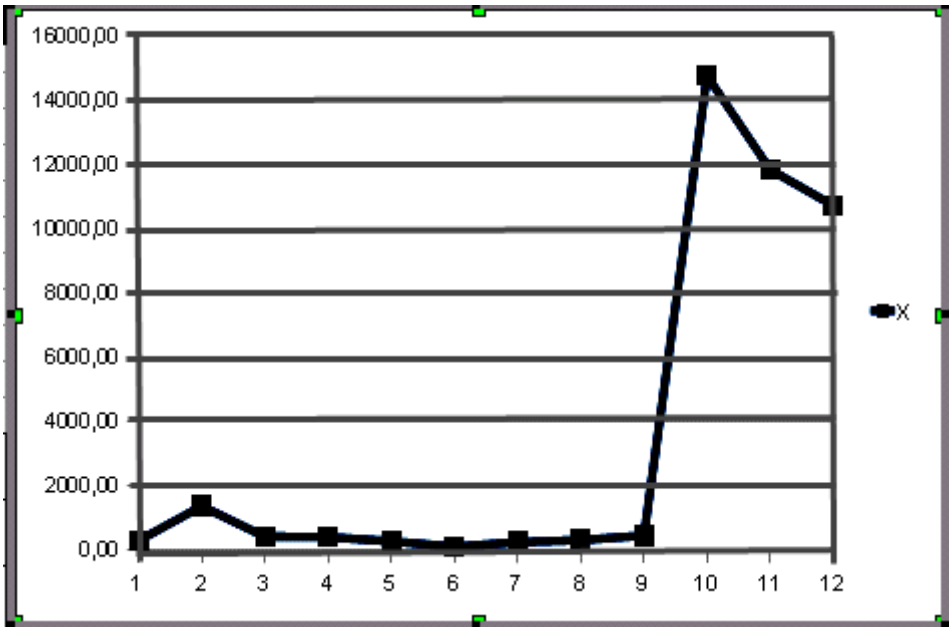


Рис. 1.14. **Графік зображення дискретного варіаційного ряду OpenOfficeCalc**

Для зображення інтервального ряду у Майстрі діаграм потрібно обрати тип діаграми Гістограма (Звичайна) і натиснути Далее>>. Після обрання діапазону даних необхідно визначити комірку з назвою досліджуваного показника (Частоти (f)) та самі значення, як зображено на рис. 1.15.

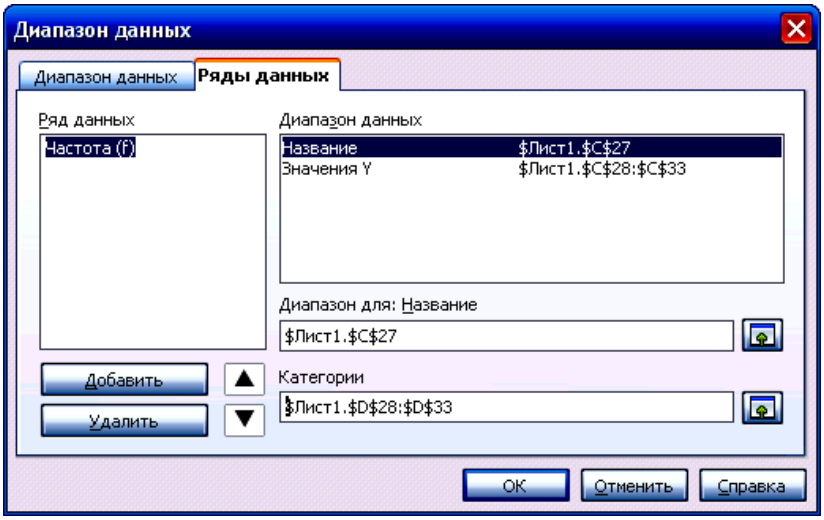


Рис. 1.15. **Визначення діапазону назви та значень досліджуваного показника за допомогою Майстра діаграм OpenOfficeCalc**

При ініціюванні кнопки "ОК" з’явиться гістограма розподілу частот (рис. 1.16)

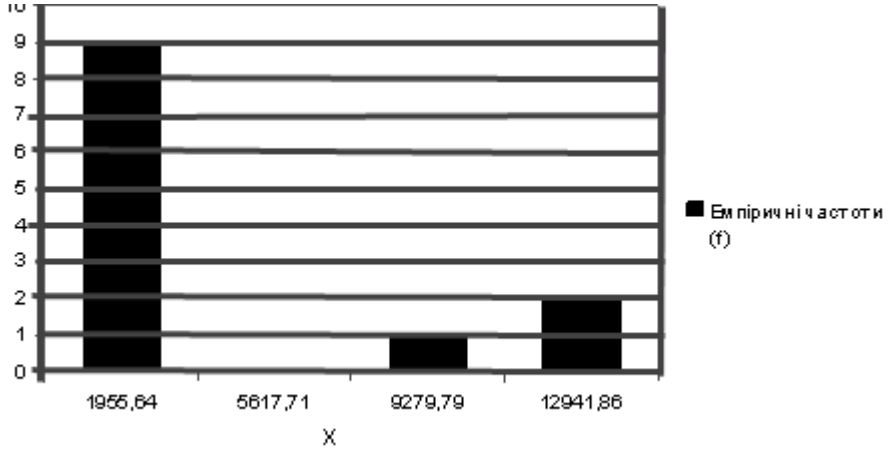


Рис. 1.16. **Гістограма розподілу частот інтервального ряду OpenOfficeCalc**

Наступним кроком доцільно графічно зобразити кумуляту частот інтервального ряду. Для цього у Майстрі діаграм необхідно обрати тип діаграми Лінії (Лінії та точки), а діапазоні даних визначити стовбцем, у якому знаходяться накопичені частоти інтервалів. У результаті з’явиться графік, зображений на рис. 1.17.

Рис. 1.17. **Кумулята частот інтервального ряду OpenOfficeCalc**

4.1.2. Побудова діаграм у Microsoft Office Excel. Для наочного представлення варіаційних рядів у середовищі *Microsoft Office Excel* необхідно в меню Вставка обрати "Диаграмма" після чого на екрані з'явиться меню Майстра діаграм, зображене на рис. 1.18.

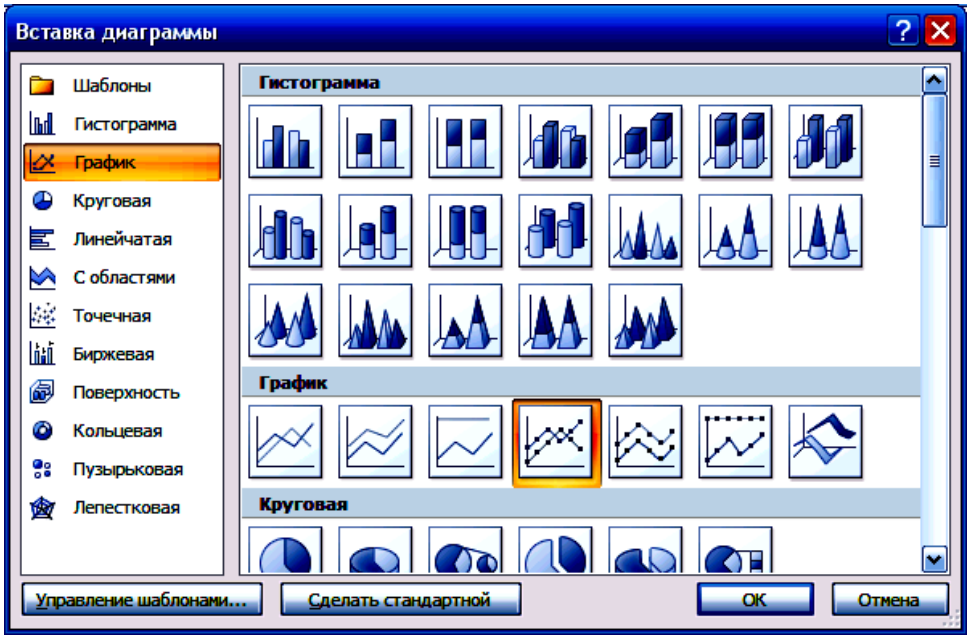


Рис. 1.18. **Обрання типу діаграми у Майстра діаграм Excel**

Для зображення дискретного ряду потрібно обрати тип діаграми "График (График с маркерами)" і натиснути ОК.

У контекстному меню порожнього рисунка, який щойно з'явився, потрібно обрати "Выбрать данные".

Наступним кроком необхідно обрати діапазон даних, що необхідно зобразити, визначити розташування та підпис, як зображено на рис. 1.19, 1.20. Після чого натиснути ОК.

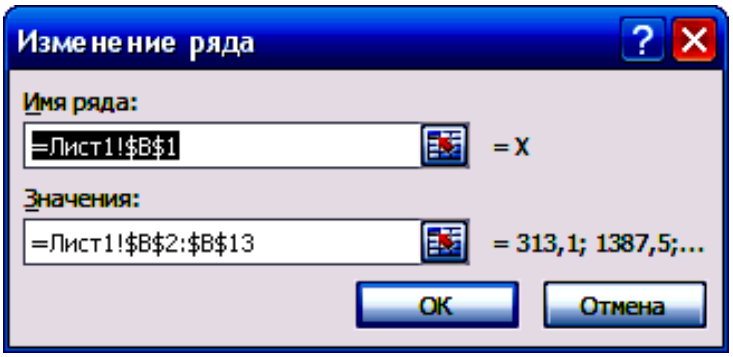


Рис. 1.19. **Визначення підпису та діапазону показника у Excel**

Графік зображення дискретного варіаційного ряду, у якого по осі X – номер об’єкта, а по осі Y – значення дослідженого показника представлено на рис. 1.21

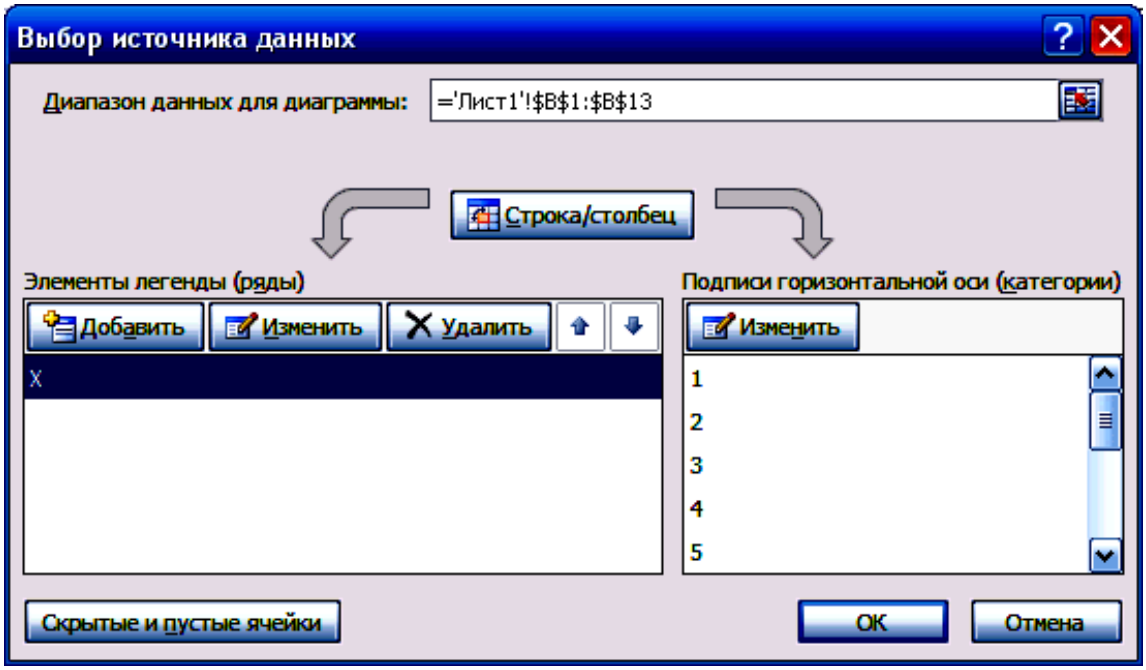


Рис. 1.20. **Вікно заповнених вихідних даних**

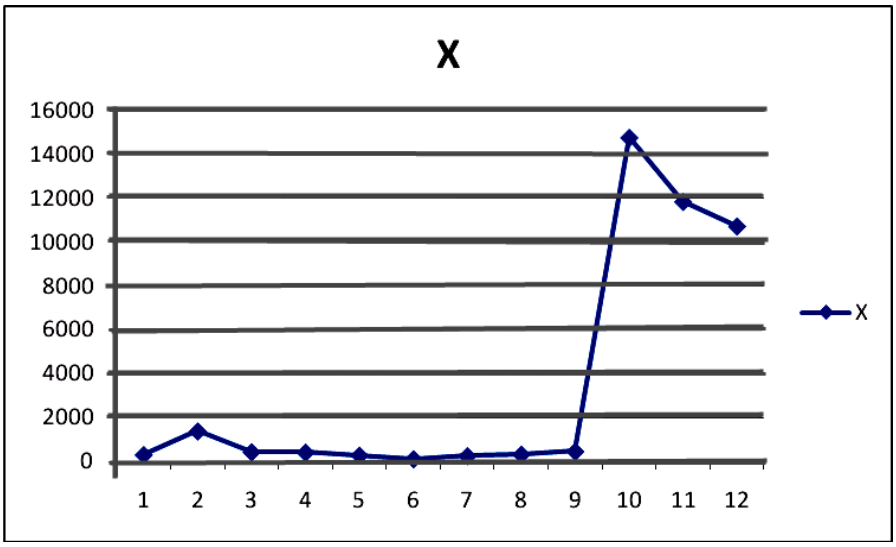


Рис. 1.21. **Графік зображення дискретного варіаційного ряду у Microsoft Office Excel**

Для зображення інтервального ряду у Майстра діаграм потрібно обрати тип діаграми "Гистограмма (Гистограмма с группировкой)" і

натиснути ОК. Наступним кроком здійснюється обрання діапазону даних і визначення комірки з назвою досліджуваного показника (Емпіричні частоти (f) та самі значення), а також визначення підписів горизонтальної осі, у якості якої виступають значення середини інтервалів, як зображено на рис. 1.22.

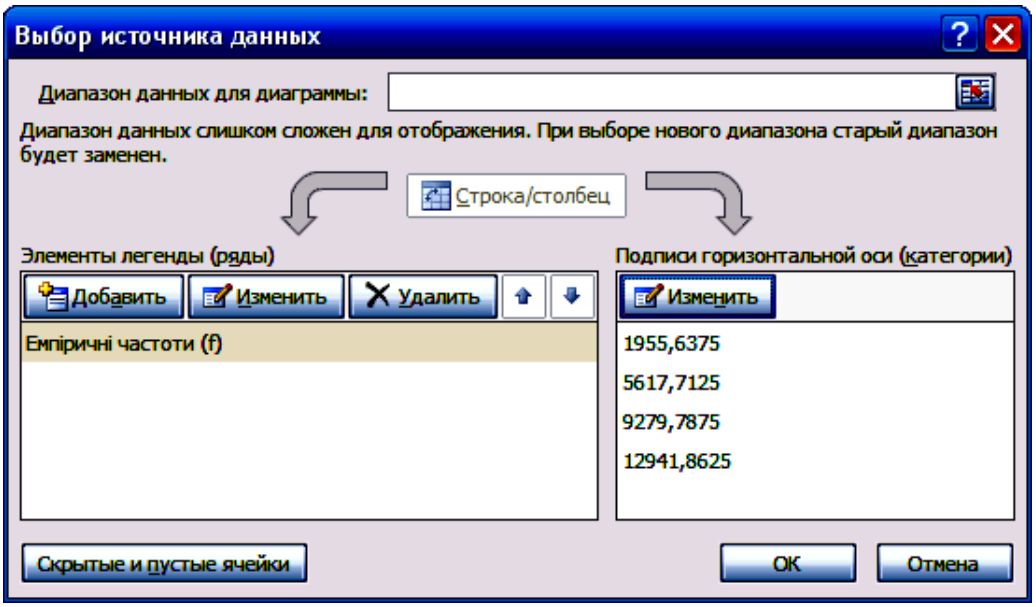


Рис. 1.22. **Вікно заповнених вихідних даних інтервального варіаційного ряду у Майстрі діаграм у Microsoft Office Excel**

При ініціюванні кнопки "ОК" з’явиться гістограма розподілу частот (рис. 1.23)

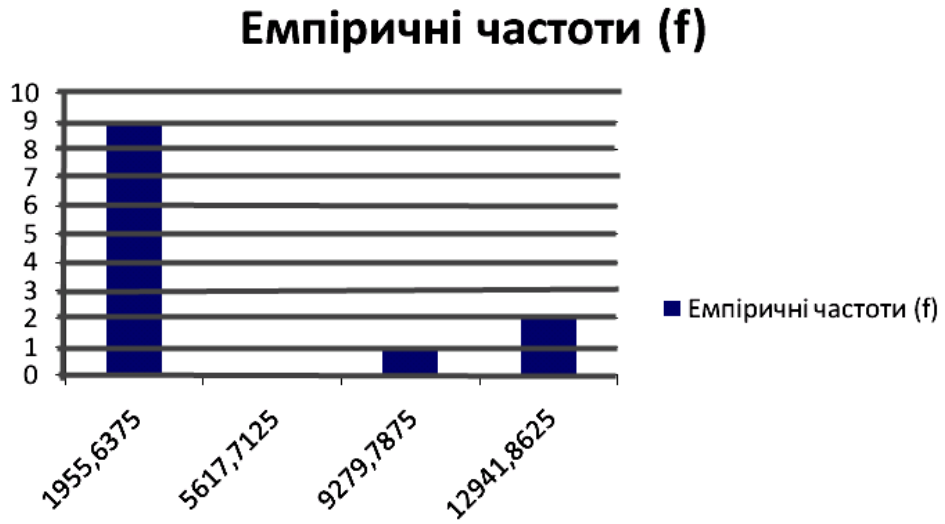


Рис. 1.23. **Гістограма розподілу частот інтервального ряду у Microsoft Office Excel**

Наступним кроком доцільно графічно зобразити кумуляту частот інтервального ряду. Для цього у Майстрі діаграм необхідно обрати тип діаграми "*График (График с маркерами)*", а діапазон даних визначити стовбцем, у якому знаходяться накопичені частоти інтервалів. У результаті з'явиться графік, зображений на рис. 1.24.

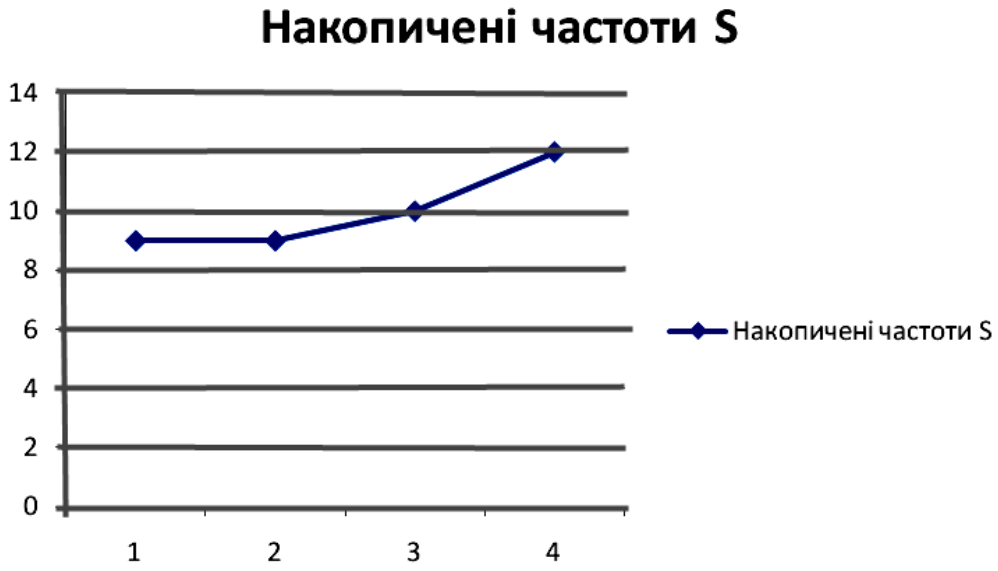


Рис. 1.24. **Кумулята частот інтервального ряду у Excel**

5. Перевірка гіпотези про нормальний закон розподілу ряду за допомогою критерію Пірсона та Колмагорова – Смірнова. Важливою умовою визначення характеру даного емпіричного ряду є побудова на базі емпіричних даних частот теоретичного нормального розподілу за наступною формулою:

$$f^* = \frac{nk}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}, \qquad (9)$$

де f^* – теоретичні частоти нормального розподілу;

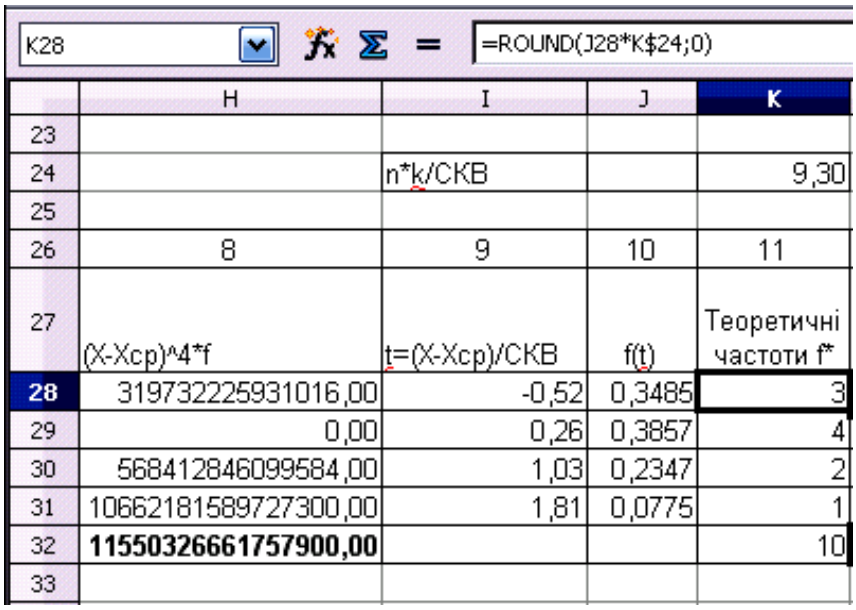
n – кількість спостережень;

k – крок групування інтервалів,

$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ – табличне значення функції нормального розподілу f(t), де

$$t = \frac{x - \bar{x}}{\sigma}. \qquad (10)$$

На рис. 1.25 зображена таблиця, в якій розраховані теоретичні частоти досліджуваного ряду.



K28 | =ROUND(J28*K$24;0)

	H	I	J	K
23				
24		n*k/СКВ		9,30
25				
26	8	9	10	11
27	(X-Xср)^4*f	t=(X-Xср)/СКВ	f(t)	Теоретичні частоти f*
28	319732225931016,00	-0,52	0,3485	3
29	0,00	0,26	0,3857	4
30	568412846099584,00	1,03	0,2347	2
31	106621815897273 00,00	1,81	0,0775	1
32	**11550326661757900,00**			10
33				

Рис. 1.25. **Розрахунок теоретичних частот нормального закону розподілу**

Стовбець 9 даної таблиці розраховано за формулою (10), у стовбці 10 знаходяться відповідні t табличні значення функції нормального розподілу, стовбець 11 розраховується множенням 10 на постійне значення $\frac{nk}{\sigma}$ відповідно до формули (9) та наданням отриманим частотам форми цілого числа.

Для перевірки гіпотези про нормальність розподілу досліджуваного ряду найчастіше використовуються критерії Колмагорова – Смірнова (11) та критерій згоди Пірсона (12):

$$\lambda = \frac{\max|S - S^*|}{\sqrt{n}}. \qquad (11)$$

$$\chi^2 = \sum \frac{(f - f^*)^2}{f^*}. \qquad (12)$$

На рис. 1.26 зображено порядок розрахунку зазначених критеріїв, де стовбець 12 розраховується шляхом накопичення значень

стовбця 11, 13 = 7 – 12, 14=(3 – 11)^2/11, λ | =M32/SQRT(12); X^2 | =SUM(N28:N33), λ, p(X^2) – табличні значення.

	I	J	K	L	M	N	O
23							
24	n*k/СКВ		9,30				
25							
26	9	10	11	12	13	14	
27	t=(X-Xср)/СКВ	f(t)	Теоретичні частоти f*	Накопичені теоретичні частоти S*	|S-S*|	(f-f*)^2/f*	
28	-0,52	0,3485	3	3	6,0000	12,00	
29	0,26	0,3857	4	7	4,0000	4,00	
30	1,03	0,2347	2	9	1,0000	0,50	
31	1,81	0,0775	1	10	1,0000	1,00	
32			10	**max**	**6,00**		
33							
34							
35				λ	1,73		
36				p(λ)	0,0032		
37				X^2	17,50		
38				p(X^2) (k=4-1)	0,0002		
39							

Рис. 1.26. **Розрахунок критеріїв Пірсона та Колмагорова – Смірнова**

Згідно отриманих результатів гіпотеза про нормальність досліджуваного ряду відхиляється. Близькі до нуля ймовірності свідчать про невипадкові розходження між емпіричними та теоретичними частотами, що можно візуально побачити на рис 1.27.

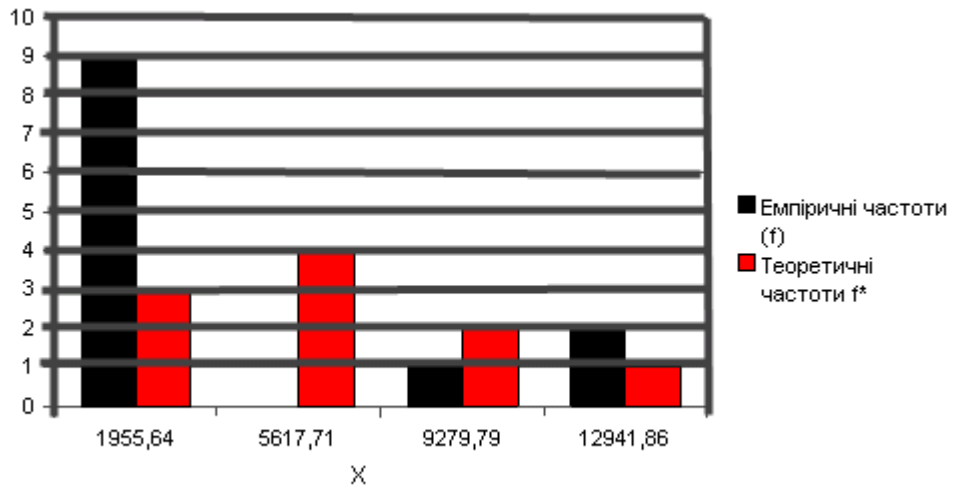


Рис. 1.27. **Гістограми розподілу емпіричних та теоретичних частот досліджуваного ряду**

Лабораторна робота № 2. Ієрархічні агломеративні процедури кластерного аналізу

Мета – закріплення теоретичного й практичного матеріалу, набуття навичок кластеризації за ієрархічними агломеративними процедурами в табличних редакторах *OpenOffice Calc 3.2* та *Microsoft Office Excel 2007*.

Завдання: за допомогою пакету *Calc* або *Excel* на підставі даних з табл. 2.1 провести кластеризацію об'єктів, описаних трьома показниками, за агломеративними процедурами. Кластеризацію провести методом найближчого сусіда, дальнього сусіда, середнього зв'язку, центрів тяжіння. При обчислені відстаней використовувати просту Евклідову відстань. Результати кластеризації представити у вигляді дендрограм. Порівняти результати кластеризації, отримані за різними методами. Зробити висновки щодо наявності природного розбиття сукупності об'єктів на кластери.

Лабораторна робота виконується в межах теми "Математико-статистична обробка вибіркових даних".

Таблиця 2.1

Вихідні дані для кластерного аналізу

№ об'єкта	Характеристики		
	x1	x2	x3
1	2	12	5
2	3	15	6
3	2	14	5
4	7	16	2
5	8	17	4
6	5	17	2

Методичні рекомендації до виконання завдання 1

Маємо шість об'єктів, описаних трьома показниками x1-x3. Геометрично кожному з об'єктів відповідає точка в трьовимірному просторі (де осями координат виступають шкали значень показників x1-x3). Графічно це може бути представлено графіком на рис. 1.1.

Кластеризація об'єктів за агломеративною процедурою припускає, що на першому етапі всі № об'єктів розглядаються як окремі кластери.

Розраховується відстань між ними та об’єднуються найближчі кластери. Матриця відстаней перераховується для отриманої меншої кількості кластерів (зазвичай *N*-1) і знову об’єднуються найближчі кластери. Потім знову перераховується матриця відстаней між кластерами і т. д. Процедура триває до тих пір, доки всі об’єкти не об’єднуються в один кластер.

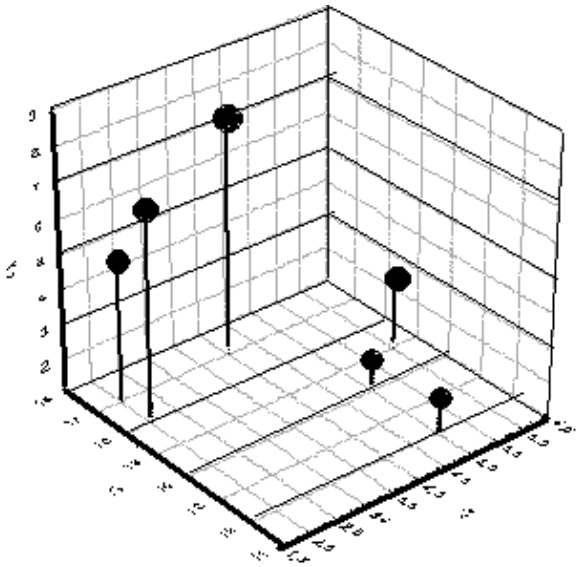


Рис. 2.1. **Геометричне представлення вихідних даних**

1. Розглянемо агломеративну процедуру за методом найближчого сусіда.

Для цього поперше розрахуємо матрицю відстаней між об’єктами. Геометрична відстань між точками на рис. 2.1 відповідає Евклідовій відстані:

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

де d_{ij} – відстань між і-тим та j-тим об’єктами;

x_{ik} – k-та координата і-того об’єкта (значення k-того показника для і-того об’єкта);

x_{jk} – k-та координата j-того об’єкта (значення k-того показника для і-того об’єкта);

m – кількість характеристик (показників), за якими описані об’єкти.

Щоб отримати матрицю відстаней у пакеті Calc або Excel набираємо таблицю з вихідними даними двічі: так як дані були надані в

завданні та транспоновану матрицю (рис. 2.2). Це дозволить ввести формулу для розрахунку Евклідової відстані в першу з комірок таблиці відстаней, зафіксувати потрібні комірки знаком $, та потім розтягнути такий тип на всю таблицю (рис. 2.2).

I8 =КОРЕНЬ(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2)

	B	C	D	E	H	I	J	K	L	M	N	P	Q
2		Характеристики				1	2	3	4	5	6		
3	№ об'єкта	x1	x2	x3	x1	2	3	2	7	8	5		
4	1	2	12	5	x2	12	15	14	16	17	17		
5	2	3	15	6	x3	5	6	5	2	4	2		
6	3	2	14	5									
7	4	7	16	2	1 Этап		1	2	3	4	5	6	Матриця евклідових відстаней
8	5	8	17	4		1	0	3,31662	2	7,07107	7,87401	6,55744	
9	6	5	17	2		2	3,31662	0	**1,73205**	5,74456	5,74456	4,89898	
10						3	2	**1,73205**	0	6,16441	6,78233	5,19615	мінімальна відстань =
11						4	7,07107	5,74456	6,16441	0	2,44949	2,23607	**1,732**
12	№ етапа	Об'єкти	Відстань			5	7,87401	5,74456	6,78233	2,44949	0	3,60555	
13	1	2+3	1,732			6	6,55744	4,89898	5,19615	2,23607	3,60555	0	

Рис. 2.2. **Розрахунок матриці відстаней між об’єктами**

У комірку I8 вводимо формулу:

для Calc - =SQRT(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2);

для Excel - =КОРЕНЬ(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2).

Завдяки тому, що в формулі розставлені знаки $, маємо можливість розтягнути введену формулу на весь діапазон I8-N13. Перевіряємо, якщо формула введена правильно – матриця відстаней буде симетричною матрицею з нулями на головній діагоналі.

Знаходимо мінімальну відстань між об’єктами. Як видно з рис. 2.2 – це відстань між 2 та 3 об’єктами. Об’єднуємо їх у новий кластер.

Для того, щоб формалізувати результати розрахунків сформуємо маленьку таблицю з трьох стовбців, що будуть містити № етапу, назви об’єктів, що на цьому етапі об’єднано, та відстані, на яких об’єднано об’єкти (рис. 2.2). Вносимо в цю таблицю дані щодо першого об’єднання (2 та 3 об’єкти).

Після цього будуємо нову таблицю для матриці відстаней між кластерами (рис. 2.3). На цьому етапі в нас кластерів вже 5, а не шість. Їх представляють 1, 2+3, 4, 5 та 6 об’єкти, відповідно.

Відстані між 1, 4, 5 та 6 кластерами вже розраховані в попередній таблиці. Перерахувати треба тільки відстані між новим кластером "2+3" та 1, "2+3" та 4, "2+3" та 5, "2+3" та 6 кластерами.

За методом найближчого сусіда за відстань між кластерами береться відстань між найближчими об’єктами двох кластерів. Тож, наприклад відстань між кластером "2+3" та 1 кластером, що містить

тільки один 1 об’єкт, розраховується, як мінімальна відстань з $d_{13} = 2$ та $d_{12} = 3{,}316$, і дорівнює, відповідно, $\min d = d_{2+3,1} = 2$.

Для розрахунків використовуємо вбудовану функцію пакета Calc MIN(_) або Excel – МИН(_).

J16 =MIN(J8:K8)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1																		
2			Характеристики						1	2	3	4	5	6				
3		№ об'єкта	x1	x2	x3			x1	2	3	2	7	8	5				
4		1	2	12	5			x2	12	15	14	16	17	17				
5		2	3	15	6			x3	5	6	5	2	4	2				
6		3	2	14	5													
7		4	7	16	2		1 Этап		1	2	3	4	5	6				
8		5	8	17	4			1	0	3,316625	2	7,071068	7,874008	6,557439				
9		6	5	17	2			2	3,316625	0	**1,732051**	5,744563	5,744563	4,898979				
10								3	2	**1,732051**	0	6,164414	6,78233	5,196152		минімальна		
11								**4**	7,071068	5,744563	6,164414	0	2,44949	2,236068		відстань =	1,732	
12		№ етапа	Об'єкти	Відстань				5	7,874008	5,744563	6,78233	2,44949	0	3,605551				
13		1	2+3	1,732				6	6,557439	4,898979	5,196152	2,236068	3,605551	0				
14		2	1+2+3	2,000														
15		3					2 Этап		1	2+3	**4**	5	6					
16		4						1	0	**2**	7,071068	7,874008	6,557439					
17		5	все					2+3	**2**	0	5,744563	5,744563	**4,898979**			минімальна		
18								**4**	7,071068	5,744563	0	2,44949	2,236068			відстань =	2,000	
19								5	7,874008	5,744563	2,44949	0	3,605551					
20								6	6,557439	4,898979	2,236068	3,605551	0					
21																		

Рис. 2.3. **Розрахунки другого етапу кластеризації за методом найближчого сусіда**

Після розрахунку матриці відстаней знову знаходимо найближчі кластери. Оскільки мінімальна відстань, рівна 2 знаходиться між 1 та "2+3" кластерами на наступному етапі об’єднуємо саме їх.

На наступному етапі будуємо матрицю відстаней вже 4*4, бо маємо лише 4 кластери: "1+2+3", 4, 5, 6 (рис. 2.4).

Перераховуємо відстані. Обираємо мінімальну відстань та об’єднуємо кластери 4 та 6 ($\min d = d_{46} = 2{,}236$).

I24 =MIN(I18:J18)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2			Характеристики						1	2	3	4	5	6			
3		№ об'єкта	x1	x2	x3			x1	2	3	2	7	8	5			
4		1	2	12	5			x2	12	15	14	16	17	17			
5		2	3	15	6			x3	5	6	5	2	4	2			
6		3	2	14	5												
7		4	7	16	2		1 Этап		1	2	3	4	5	6			
8		5	8	17	4			1	0	3,316625	2	7,071068	7,874008	6,557439			
9		6	5	17	2			2	3,316625	0	**1,732051**	5,744563	5,744563	4,898979			
10								3	2	**1,732051**	0	6,164414	6,78233	5,196152		минімальна	
11								**4**	7,071068	5,744563	6,164414	0	2,44949	2,236068		відстань =	1,732
12		№ етапа	Об'єкти	Відстань				5	7,874008	5,744563	6,78233	2,44949	0	3,605551			
13		1	2+3	1,732				6	6,557439	4,898979	5,196152	2,236068	3,605551	0			
14		2	1+2+3	2,000													
15		3	4+6	2,236			2 Этап		1	2+3	**4**	5	6				
16		4						1	0	**2**	7,071068	7,874008	6,557439				
17		5	все					2+3	**2**	0	5,744563	5,744563	**4,898979**			минімальна	
18								**4**	7,071068	5,744563	0	2,44949	2,236068			відстань =	2,000
19								5	7,874008	5,744563	2,44949	0	3,605551				
20								6	6,557439	4,898979	2,236068	3,605551	0				
21																	
22							3 Этап		1+2+3	**4**	5	6					
23								1+2+3	0	5,744563	5,744563	4,898979				минімальна	
24								**4**	5,744563	0	2,44949	**2,236068**				відстань =	2,236
25								5	5,744563	2,44949	0	3,605551					
26								6	4,898979	**2,236068**	3,605551	0					

Рис. 2.4. **Розрахунки третього етапу кластеризації за методом найближчого сусіда**

На наступному етапі маємо вже 3 кластери: "1+2+3", "4+6", "5". Будуємо матрицю відстаней розмірності 3*3 (рис. 2.5).

Знаходимо мінімальну відстань $\min d = d_{4+6,5} = 2{,}449$ та об'єднуємо кластери "4+6" та "5".

На останньому етапі маємо 2 кластери: "1+2+3", "4+5+6". Будуємо матрицю (рис. 2.5), знаходимо мінімальну відстань $d = 4{,}898$ та об'єднуємо всі об'єкти в один кластер.

T24

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
4		1	2	12	5			x2	12	15	14	16	17	17			
5		2	3	15	6			x3	5	6	5	2	4	2			
6		3	2	14	5												
7		4	7	16	2		1 етап		1	2	3	4	5	6			
8		5	8	17	4			1	0	3,316625	2	7,071068	7,874008	6,557439			
9		6	5	17	2			2	3,316625	0	**1,732051**	5,744563	5,744563	4,898979			
10								3	2	**1,732051**	0	6,164414	6,78233	5,196152		мінімальна	
11								4	7,071068	5,744563	6,164414	0	2,44949	2,236068		відстань =	**1,73205**
12		№ етапа	Об'єкти	Відстань				5	7,874008	5,744563	6,78233	2,44949	0	3,605551			
13		1	2+3	1,73205				6	6,557439	4,898979	5,196152	2,236068	3,605551	0			
14		2	1+2+3	2,00000													
15		3	4+6	2,23607			2 етап		1	2+3	4	5	6				
16		4	4+6+5	2,44949				1	0	**2**	7,071068	7,874008	6,557439				
17		5	все					2+3	**2**	0	5,744563	5,744563	4,898979			мінімальна	
18								4	7,071068	5,744563	0	2,44949	2,236068			відстань =	**2,00000**
19								5	7,874008	5,744563	2,44949	0	3,605551				
20								6	6,557439	4,898979	2,236068	3,605551	0				
21																	
22							3 етап		1+2+3	4	5	6					
23								1+2+3	0	5,744563	5,744563	4,898979				мінімальна	
24								4	5,744563	0	2,44949	**2,236068**				відстань =	**2,23607**
25								5	5,744563	2,44949	0	3,605551					
26								6	4,898979	**2,236068**	3,605551	0					
27																	
28							4 етап		1+2+3	4+6	5						
29								1+2+3	0	4,898979	5,744563					мінімальна	
30								4+6	4,898979	0	**2,44949**					відстань =	**2,44949**
31								5	5,744563	**2,44949**	0						
32																	
33							5 етап		1+2+3	4+5+6							
34								1+2+3	0	4,898979						мінімальна	
35								4+5+6	4,898979	0						відстань =	**4,89898**
36																	

Рис. 2.5. **Розрахунки четвертого етапу кластеризації за методом найближчого сусіда**

За результати розрахунків будуємо дендрограму. Будуємо вручну на аркуші паперу, бо пакети не дозволяють це зробити автоматично. На осі ординат відмічаємо об'єкти, а по осі абсцис – відстані, на яких об'єкти об'єднуються у кластери. Отримуємо наступний малюнок – рис. 2.6.

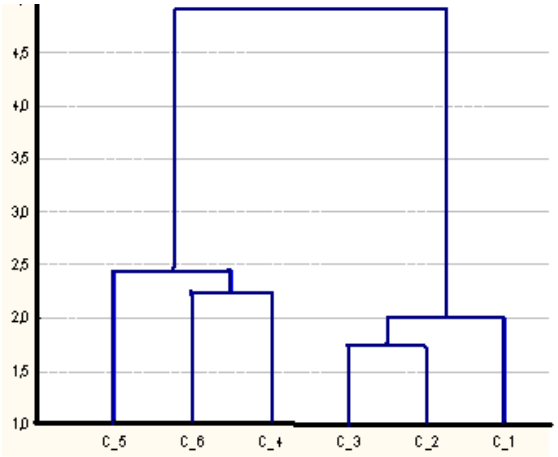


Рис. 2.6. **Дендрограма за методом найближчого сусіда**

Аналізуємо отриману дендрограму: проводимо лінію відсікання (рис. 2.7).

Бачимо, що існує природне розбиття сукупності об’єктів на кластери. А саме можна отримати два досить істотно відмінні кластери. У першому будуть міститися об’єкти 4, 5 та 6. В другому – 1, 2 та 3.

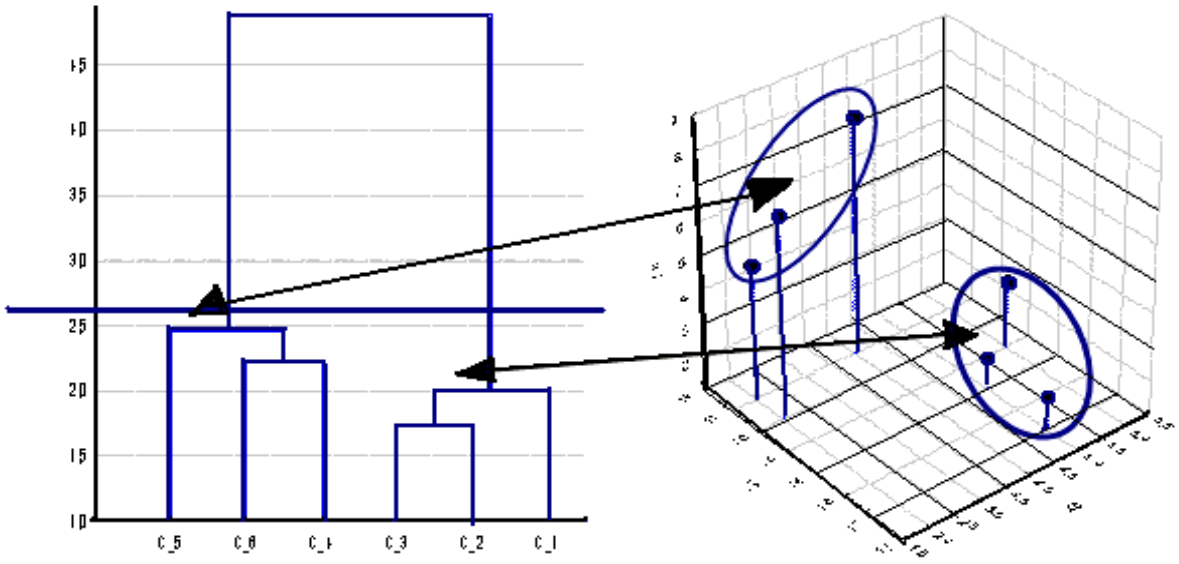


Рис. 2.7. **Відповідність між дендрограмою та розташуванням об’єктів у просторі ознак**

2. Розглянемо агломеративну процедуру за методом дальнього сусіда. Процедура кластеризації за методом дальнього сусіда здійснюється тим самим чином.

Вся процедура за етапами представлена на рис. 2.8.

Єдина відмінність: за правилом дальнього сусіда за відстань між кластерами береться відстань між найбільш далекими один від одного об’єктами у кластерах. Наприклад, на другому етапі кластеризації (рис. 8) у комірці J16 потрібно розрахувати відстань між кластерами "2+3" та "1". За таку відстань береться максимальне значення з відстаней між "1" та "2", та "1" та "3":

$$d_{2+3,1} = \max(d_{1,2}; d_{1,3}) = d_{1,2} = 3{,}31.$$

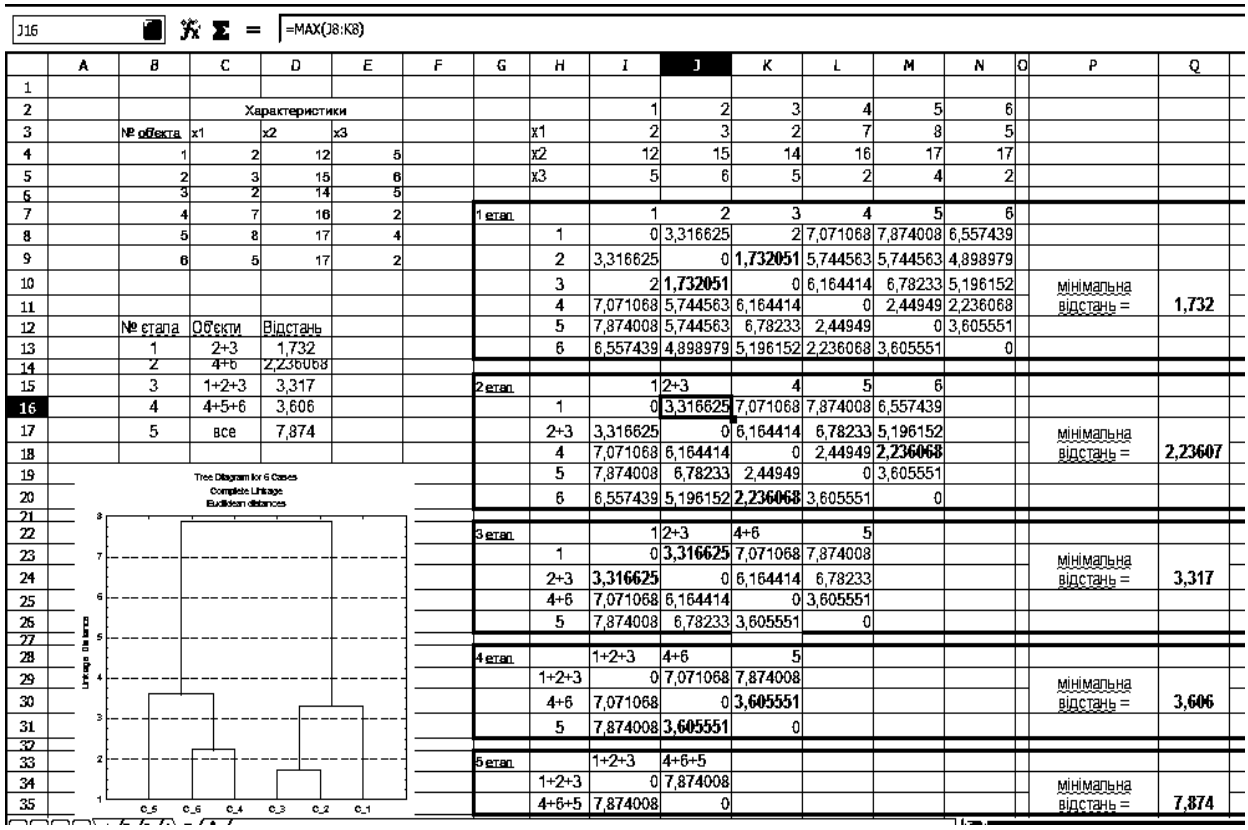


Рис. 2.8. **Процедура кластеризації за методом дальнього сусіда**

3. Розглянемо агломеративну процедуру за методом середнього зв'язку. Відмінність полягає у розрахунку відстаней між кластерами.

За відстань між двома кластерами береться середньоарифметичне всіх відстаней між об'єктами двох кластерів. Так на другому етапі відстань між кластерами "2+3" та "1", що міститься в комірці J16, треба розрахувати як:

$$d_{2+3,1} = \frac{d_{1,2}+d_{1,3}}{2} = \frac{3{,}31+2}{2} = 2{,}658.$$

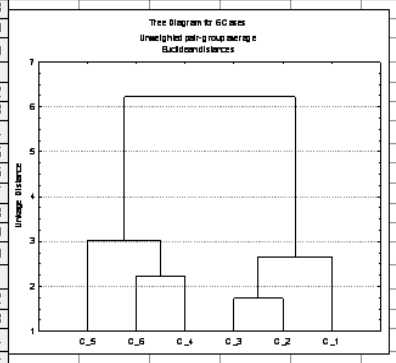
Вся процедура за етапами представлена на рис. 2.9.

4. Розглянемо агломеративну процедуру за методом центрів тяжіння. Ця процедура припускає додатковий етап: для кожного побудованого кластера розраховується координати його центру тяжіння (рис. 2.10).

J16 =МАКС(J8:K8)

№ об'єкта	Характеристики x1	x2	x3
1	2	12	5
2	3	15	6
3	2	14	5
4	7	16	2
5	8	17	4
6	5	17	2

№ етапа	Об'єкти	Відстань
1	2+3	1,732
2	4+6	2,23607
3	1+2+3	2,658
4	4+5+6	3,028
5	все	6,613



1 Этап

	1	2	3	4	5	6
x1	2	3	2	7	8	5
x2	12	15	14	16	17	17
x3	5	6	5	2	4	2

	1	2	3	4	5	6
1	0	3,31662	2	7,07107	7,87401	6,55744
2	3,31662	0	**1,73205**	5,74456	5,74456	4,89898
3	2	**1,73205**	0	6,16441	6,78233	5,19615
4	7,07107	5,74456	6,16441	0	2,44949	2,23607
5	7,87401	5,74456	6,78233	2,44949	0	3,60555
6	6,55744	4,89898	5,19615	2,23607	3,60555	0

мінімальна відстань = **1,732**

2 Этап

	1	2+3	4	5	6
1	0	3,31662	7,07107	7,87401	6,55744
2+3	3,31662	0	6,16441	6,78233	5,19615
4	7,07107	6,16441	0	2,44949	**2,23607**
5	7,87401	6,78233	2,44949	0	3,60555
6	6,55744	5,19615	**2,23607**	3,60555	0

мінімальна відстань = **2,23607**

3 Этап

	1	2+3	4+6	5
1	0	**2,65831**	6,81425	7,87401
2+3	**2,65831**	0	5,50103	6,26345
4+6	6,81425	5,50103	0	3,02752
5	7,87401	6,26345	3,02752	0

мінімальна відстань = **2,658**

4 Этап

	1+2+3	4+6	5
1+2+3	0	6,15764	7,06873
4+6	6,15764	0	**3,02752**
5	7,06873	**3,02752**	0

мінімальна відстань = **3,028**

5 Этап

	1+2+3	4+6+5
1+2+3	0	**6,61318**
4+6+5	**6,61318**	0

мінімальна відстань = **6,613**

Рис. 2.9. **Процедура кластеризації за методом середнього зв'язку**

Таким чином, перший етап – розрахунок матриці відстаней між кластерами, що містять кожний тільки по одному об'єкту, той же самий, що й в трьох попередніх методах. Але другий етап починається з розрахунку центра новоствореного об'єднаного кластеру: комірки T17-V17. Координати центру кластера розраховуються як середнє арифметичне відповідних координат елементів кластера. Наприклад, координата по ознаці x1 кластеру "2+3" визначається як T17=СРЗНАЧ(C5:C6) в *Excel* та T17=AVERAGE(C5:C6) в *Calc.*

Після розрахунку центра об'єднаного кластера відстань до інших кластерів від цього нового визначається за евклідовою відстанню між центрами кластерів. Наприклад, відстань між кластерами "2+3" та "1", що міститься в комірці J16 дорівнює:

в Excel - =КОРЕНЬ((C4-T$17)^2+(D4-U$17)^2+(E4-V$17)^2);
в Calc - =SQRT((C4-T$17)^2+(D4-U$17)^2+(E4-V$17)^2).

Вся процедура за етапами представлена на рис. 2.10.

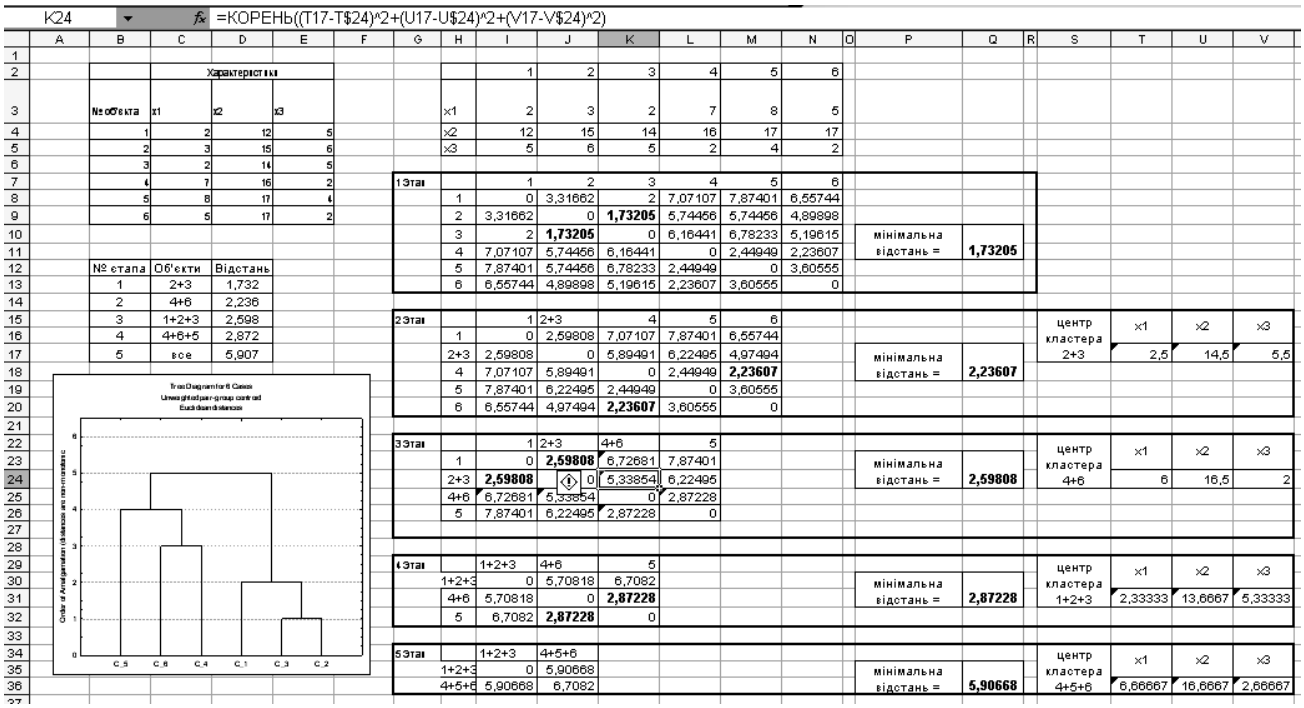


Рис. 2.10. **Процедура кластеризації за методом центрів тяжіння**

За результатами 4 процедур кластеризації будуємо дендрограми та у висновках аналізуємо отримані розбиття сукупності об’єктів на кластери, наявність природного розбиття та ін.

Лабораторна робота № 3. Ітеративні процедури кластерного аналізу на прикладі методу К-середніх

Мета – закріплення теоретичного й практичного матеріалу, набуття навичок кластеризації ітеративними процедурами кластерного аналізу в табличних редакторах *OpenOffice* Calc *3.2* та *Microsoft Office Excel 2007*.

Завдання: за допомогою пакету *OpenOffice Calc* або *Microsoft Office Excel* на підставі даних з табл. 2.1 провести кластеризацію об'єктів, описаних трьома показниками, за методом К-середніх.

Розбиття сукупності об’єктів на кластери провести для гіпотез, щодо наявності двох та трьох кластерів. За начальні центри кластерів обрати перші об’єкти з таблиці.

Лабораторна робота здійснюється за темою "Математично-статистична обробка вихідних даних".

Методичні рекомендації до виконання завдання

Маємо шість об’єктів, описаних трьома показниками x1-x3. Геометрично кожному з об’єктів відповідає точка в трьовимірному просторі (рис. 2.1). Значення показників, які характеризують об’єкти, виступають координатами точок відповідаючих об’єктам в n-вимірному евклідовому просторі, де n – кількість показників.

Кластеризація об’єктів за методом К-середніх є ітеративною процедурою. Ітерація (лат. слово iteration – "повтор") – це повторне використання математичної операції чи деякого набору операцій в серії для отримання деякого очікуваного результату.

Алгоритм методу К-середніх полягає у наступному:

0. Висуваємо гіпотезу, щодо кількості кластерів.

І ітерація.

1. Обираємо начальні значення центрів кластерів (наприклад за центри кластерів обираємо перші К об’єктів).

2. Розраховуємо відстані від всіх об’єктів до кожного з центрів кластерів.

3. Кожний об’єкт відносимо до того кластера, відстань до центра якого виявилася найменшою.

ІІ ітерація.

1. Перераховуємо центри кластерів (центр кластеру знаходиться як умовний об’єкт, характеристики якого розраховуються як середнє арифметичне характеристик об’єктів, що увійшли до кластера).

2. Розраховуємо відстані від всіх об’єктів сукупності до кожного з центрів кластерів.

3. Кожний об’єкт відносимо до того кластера, відстань до центра якого виявилася найменшою.

Друга ітерація повторюється до тих пір, поки:

- розбиття об’єктів на кластери не повториться два рази;
- або не виникне зациклення (розбиття повторюється через два – три рази – це може статися, коли деякий об’єкт є рівновіддаленим від центрів кластерів, чи розбиття на таку кількість кластерів не дуже характерне для цих вихідних даних);
- не досягнуто деякого значення критерію якості розбиття.

Щоб проілюструвати дію алгоритму, розглянемо умовний приклад. Існує 6 об'єктів, описаних двома показниками x1 та x2. Графічно ці об'єкти можуть бути представлені точками в двовимірному просторі ознак – рис. 3.1. Проведемо кластеризацію за гіпотезою, що кластерів 2. У якості центрів кластерів оберемо перші два об'єкти.

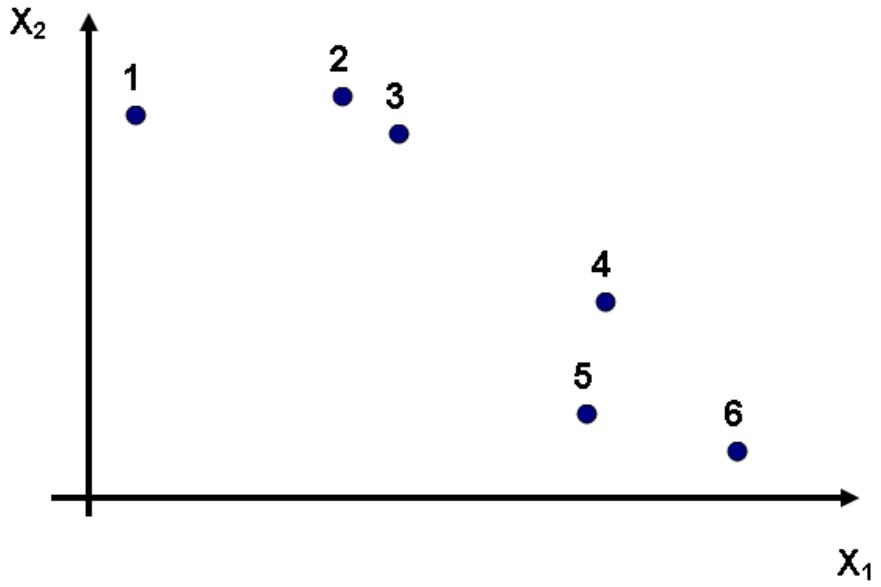


Рис. 3.1. **Представлення економічних об'єктів у двовимірному евклідовому просторі**

Оскільки до центру другого кластера (зараз це 2 об'єкт) від об'єктів 3; 4; 5 та 6 ближче ніж до центра першого кластера (об'єкт 1), то кластери мають вигляд як на рис. 3.2.

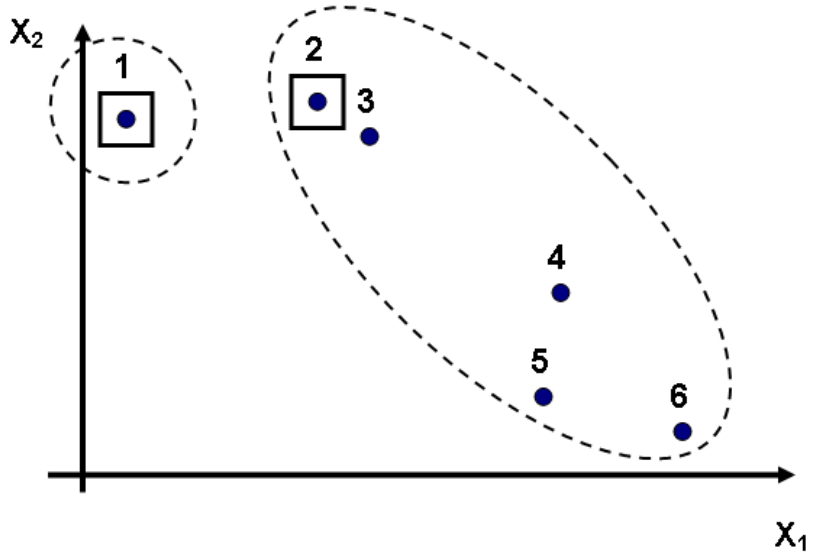


Рис. 3.2. **Результат I ітерації**

Після першої ітерації перераховуються центри кластерів. Центр першого кластера залишився незмінним, тоді коли центр другого перемістився нижче праворуч (рис. 3.3). Тепер 2 та 3 об’єктам ближче до центра першого кластера, тож кластери змінилися.

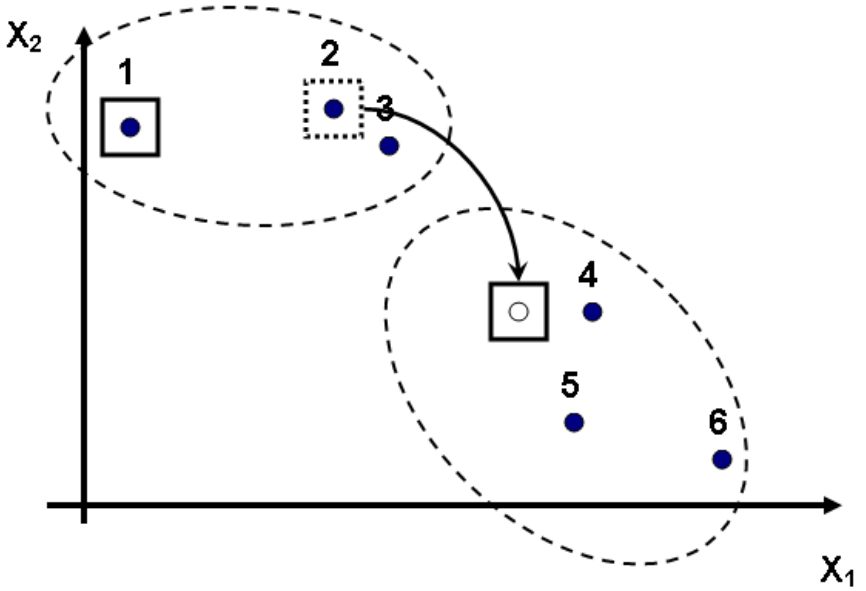


Рис. 3.3. **Результат II ітерації**

Після другої ітерації знову перераховуються центри кластерів. Центр і першого, і другого кластерів змістилися (рис. 3.4). Однак до переміщення об’єктів з кластера у кластер це не призвело. Тож ітеративний процес можна зупинити – рішення знайдено.

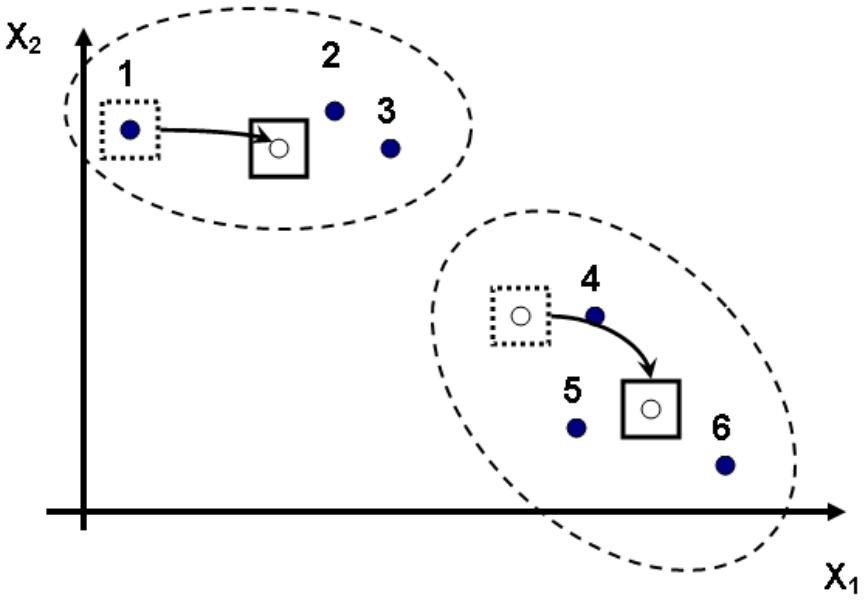


Рис. 3.4. **Результат III ітерації**

Використаємо наведений алгоритм на прикладі з табл. 2.1.

Проведемо розбиття на 2 кластери.

І ітерація.

1 крок. За начальні значення центрів двох кластерів оберемо перші два об’єкти сукупності (рис. 3.5).

2 крок. Відстані між об’єктами та центрами кластерів розрахуємо за простою Евклідовою відстанню:

$$d_{ij} = \sqrt{\sum_{k=1}^{3}(x_{ik} - x_{jk})^2},$$

де d_{ij} – відстань від і-того об’єкта до центра j-того кластера;

x_{ik} – k-та координата і-того об’єкта (значення k-того показника для і-того об’єкта);

x_{jk} – k-та координата центра j-того кластера (значення k-того показника для центра j-того кластера).

D21 =((D4-D$15)^2+(E4-E$15)^2+(F4-F$15)^2)^0,5

	A	B	C		E	F	G
1							
2			№ об'єкта	Показники			
3				x1	x2	x3	
4			1	2	12	5	
5			2	3	15	6	
6			3	2	14	5	
7			4	7	16	2	
8			5	8	17	4	
9			6	5	17	2	
10							
11							
12			**І ітерація**				
14				x1	x2	x3	
15			Центр 1 кластеру	2	12	5	
17				x1	x2	x3	
18			Центр 2 кластеру	3	15	6	
19							
20			№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру		
			1	**0,0000**	3,3166		
22			2	3,3166	**0,0000**		
23			3	2,0000	**1,7321**		
24			4	7,0711	**5,7446**		
25			5	7,8740	**5,7446**		
26			6	6,5574	**4,8990**		
27							
28						Об'єкти	
29			**1 кластер** за результатами 1 ітерації включає			1	
30			**2 кластер** за результатами 1 ітерації включає			2,3,4,5,6	
31							

Рис. 3.5. **І ітерація кластеризації за методом К-середніх**

3 крок. За результатами першої ітерації до 1 кластера слід віднести 1 об’єкт, бо для нього Евклідова відстань до центра першого кластера (d_{11}=0) є меншою ніж до центра другого (d_{12}=3.3166). Усі інші об’єкти слід віднести до кластера 2.

II ітерація.

1 крок. Перерахуємо центри кластерів. Нові центри кластерів будут мати такі значення показників x1-x3, які відповідають середнім арифметичним значенням цих показників для об’єктів, що увійшли до кластера. У пакеті Calc для їх розрахунку слід використати вбудовану функцію AVERAGE, в Excel – функцію СРЗНАЧ.

2 крок. Розраховуємо відстані до нових центрів кластерів тим самим чином, що і за I ітерації. Розрахунки структуруємо таким чином, як наведено на рис. 3.6.

3 крок. За результатами II ітерації в перший кластер слід віднести 1 та 3 об’єкти, в другий кластер – всі інші.

K18 =СРЗНАЧ(D5:D9)

	C	D	E	F	J	K	L	M
2	№ об'єкта	Показники						
3		x1	x2	x3				
4	1	2	12	5				
5	2	3	15	6				
6	3	2	14	5				
7	4	7	16	2				
8	5	8	17	4				
9	6	5	17	2				
12	I ітерація				II ітерація			
14		x1	x2	x3		x1	x2	x3
15	Центр 1 кластеру	2	12	5	Центр 1 кластеру	2	12	5
17		x1	x2	x3		x1	x2	x3
18	Центр 2 кластеру	3	15	6	Центр 2 кластеру	5	15,8	3,8
20	№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру		№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру	
21	1	0,0000	3,3166		1	0,0000	4,9880	
22	2	3,3166	0,0000		2	3,3166	3,0790	
23	3	2,0000	1,7321		3	2,0000	3,6986	
24	4	7,0711	5,7446		4	7,0711	2,6981	
25	5	7,8740	5,7446		5	7,8740	3,2373	
26	6	6,5574	4,8990		6	6,5574	2,1633	
28				Об'єкти				Об'єкти
29	1 кластер за результатами 1 ітерації включає			1	1 кластер за результатами 1 ітерації включає			1,3
30	2 кластер за результатами 1 ітерації включає			2,3,4,5,6	2 кластер за результатами 1 ітерації включає			2,4,5,6

Рис. 3.6. **I та II ітерації кластеризації за методом К-середніх**

III ітерація.

1 крок. Перераховуємо центри кластерів, зважуючи на їх новий склад.

2 крок. Розраховуємо відстані до нових центрів кластерів та заносимо ці розрахунки до таблиці (рис. 3.7).

3 крок. За результатами III ітерації до першого кластера слід віднести 1, 2 та 3 об’єкти, всі інши – до другого кластера.

D37 =СРЗНАЧ(D4;D6)

№ об'єкта	Показники		
	x1	x2	x3
1	2	12	5
2	3	15	6
3	2	14	5
4	7	16	2
5	8	17	4
6	5	17	2

I ітерація

	x1	x2	x3
Центр 1 кластеру	2	12	5

	x1	x2	x3
Центр 2 кластеру	3	15	6

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	0,0000	3,3166
2	3,3166	0,0000
3	2,0000	1,7321
4	7,0711	5,7446
5	7,8740	5,7446
6	6,5574	4,8990

	Об'єкти
1 кластер за результатами 1 ітерації включає	1
2 кластер за результатами 1 ітерації включає	2,3,4,5,6

II ітерація

	x1	x2	x3
Центр 1 кластеру	2	12	5

	x1	x2	x3
Центр 2 кластеру	5	15,8	3,8

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	0,0000	4,9880
2	3,3166	3,0790
3	2,0000	3,6986
4	7,0711	2,6981
5	7,8740	3,2373
6	6,5574	2,1633

	Об'єкти
1 кластер за результатами 1 ітерації включає	1,3
2 кластер за результатами 1 ітерації включає	2,4,5,6

III ітерація

	x1	x2	x3
Центр 1 кластеру	2	13	5

	x1	x2	x3
Центр 2 кластеру	5,75	16,25	3,5

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	1,0000	5,8630
2	2,4495	3,9211
3	1,0000	4,6233
4	6,5574	1,9685
5	7,2801	2,4238
6	5,8310	1,8371

	Об'єкти
1 кластер за результатами 1 ітерації включає	1,2,3
2 кластер за результатами 1 ітерації включає	4,5,6

Рис. 3.7. **I-III ітерації кластеризації за методом К-середніх**

IV ітерація.

1 крок. Перераховуємо центри кластерів, зважуючи на їх новий склад.

2 крок. Розраховуємо відстані до нових центрів кластерів (рис. 3.8).

3 крок. За результатми IV ітерації до складу першого кластеру слід віднести 1, 2 та 3 об’єкти, а до другого – 4, 5, 6. Це розбиття повнятю повторює те, яке було отримано за результатами попередньої ітерації. Тобто ітеративний процес слід перервати – розбиття об’єктів на кластери за методом К-середніх знайдено.

K37 =СРЗНАЧ(D4:D6)

№ об'єкта	Показники		
	x1	x2	x3
1	2	12	5
2	3	15	6
3	2	14	5
4	7	16	2
5	8	17	4
6	5	17	2

I ітерація

	x1	x2	x3
Центр 1 кластеру	2	12	5

	x1	x2	x3
Центр 2 кластеру	3	15	6

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	0,0000	3,3166
2	3,3166	0,0000
3	2,0000	1,7321
4	7,0711	5,7446
5	7,8740	5,7446
6	6,5574	4,8990

	Об'єкти
1 кластер за результатами 1 ітерації включає	1
2 кластер за результатами 1 ітерації включає	2,3,4,5,6

II ітерація

	x1	x2	x3
Центр 1 кластеру	2	12	5

	x1	x2	x3
Центр 2 кластеру	5	15,8	3,8

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	0,0000	4,9880
2	3,3166	3,0790
3	2,0000	3,6986
4	7,0711	2,6981
5	7,8740	3,2373
6	6,5574	2,1633

	Об'єкти
1 кластер за результатами 1 ітерації включає	1,3
2 кластер за результатами 1 ітерації включає	2,4,5,6

III ітерація

	x1	x2	x3
Центр 1 кластеру	2	13	5

	x1	x2	x3
Центр 2 кластеру	5,75	16,25	3,5

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	1,0000	5,8630
2	2,4495	3,9211
3	1,0000	4,6233
4	6,5574	1,9685
5	7,2801	2,4238
6	5,8310	1,8371

	Об'єкти
1 кластер за результатами 1 ітерації включає	1,2,3
2 кластер за результатами 1 ітерації включає	4,5,6

IV ітерація

	x1	x2	x3
Центр 1 клас	2,333	13,667	5,333

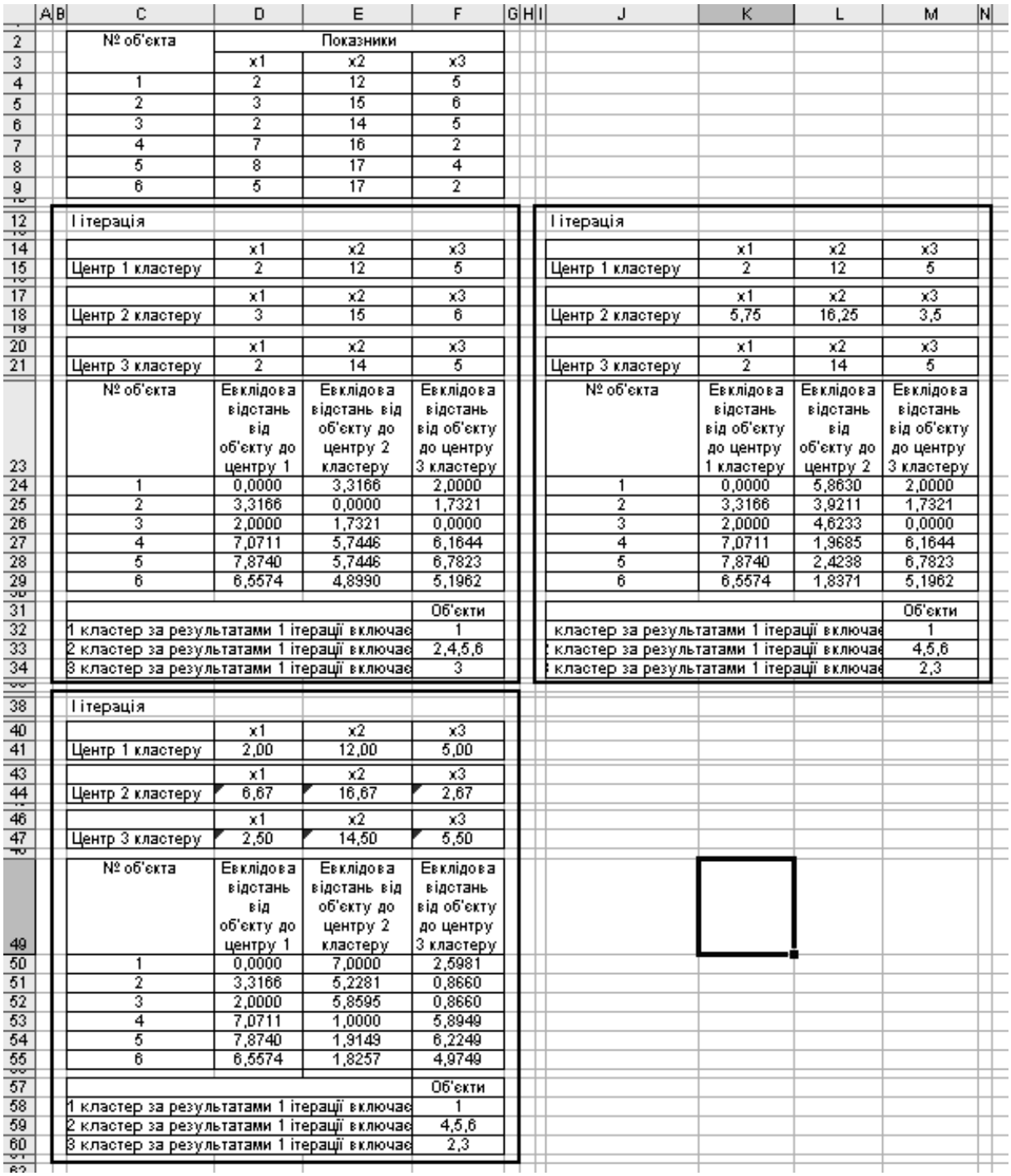
	x1	x2	x3
Центр 2 кластеру	6,667	16,667	2,667

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 1 кластеру
1	1,7321	7,0000
2	1,6330	5,2281
3	0,5774	5,8595
4	6,1914	1,0000
5	6,7082	1,9149
6	5,4160	1,8257

	Об'єкти
1 кластер за результатами 1 ітерації включає	1,2,3
2 кластер за результатами 1 ітерації включає	4,5,6

Рис. 3.8. **I-IV ітерації кластеризації за методом К-середніх**

За тим самим алгоритмом будуємо розбиття об’єктів на 3 кластери. На рис. 3.9 наведені результати розрахунків.



№ об'єкта	Показники		
	x1	x2	x3
1	2	12	5
2	3	15	6
3	2	14	5
4	7	16	2
5	8	17	4
6	5	17	2

I ітерація			
	x1	x2	x3
Центр 1 кластеру	2	12	5
	x1	x2	x3
Центр 2 кластеру	3	15	6
	x1	x2	x3
Центр 3 кластеру	2	14	5

№ об'єкта	Евклідова відстань від об'єкту до центру 1	Евклідова відстань від об'єкту до центру 2 кластеру	Евклідова відстань від об'єкту до центру 3 кластеру
1	0,0000	3,3166	2,0000
2	3,3166	0,0000	1,7321
3	2,0000	1,7321	0,0000
4	7,0711	5,7446	6,1644
5	7,8740	5,7446	6,7823
6	6,5574	4,8990	5,1962

	Об'єкти
1 кластер за результатами 1 ітерації включає	1
2 кластер за результатами 1 ітерації включає	2,4,5,6
3 кластер за результатами 1 ітерації включає	3

I ітерація			
	x1	x2	x3
Центр 1 кластеру	2,00	12,00	5,00
	x1	x2	x3
Центр 2 кластеру	6,67	16,67	2,67
	x1	x2	x3
Центр 3 кластеру	2,50	14,50	5,50

№ об'єкта	Евклідова відстань від об'єкту до центру 1	Евклідова відстань від об'єкту до центру 2 кластеру	Евклідова відстань від об'єкту до центру 3 кластеру
1	0,0000	7,0000	2,5981
2	3,3166	5,2281	0,8660
3	2,0000	5,8595	0,8660
4	7,0711	1,0000	5,8949
5	7,8740	1,9149	6,2249
6	6,5574	1,8257	4,9749

	Об'єкти
1 кластер за результатами 1 ітерації включає	1
2 кластер за результатами 1 ітерації включає	4,5,6
3 кластер за результатами 1 ітерації включає	2,3

I ітерація			
	x1	x2	x3
Центр 1 кластеру	2	12	5
	x1	x2	x3
Центр 2 кластеру	5,75	16,25	3,5
	x1	x2	x3
Центр 3 кластеру	2	14	5

№ об'єкта	Евклідова відстань від об'єкту до центру 1 кластеру	Евклідова відстань від об'єкту до центру 2	Евклідова відстань від об'єкту до центру 3 кластеру
1	0,0000	5,8630	2,0000
2	3,3166	3,9211	1,7321
3	2,0000	4,6233	0,0000
4	7,0711	1,9685	6,1644
5	7,8740	2,4238	6,7823
6	6,5574	1,8371	5,1962

	Об'єкти
кластер за результатами 1 ітерації включає	1
кластер за результатами 1 ітерації включає	4,5,6
кластер за результатами 1 ітерації включає	2,3

Рис. 3.9. **Результати I-III ітерації кластеризації за методом К-середніх для трьох кластерів**

За результатами трьох ітерацій отримаємо наступні кластери: 1 кластер – 1 об’єкт, 2 кластер – 2 та 3 об’єкти, 3 кластер – 4, 5 та 6 об’єкти.

Лабораторна робота № 4. Методи редукції

Мета – закріплення теоретичного й практичного матеріалу, набуття навичок розрахунку інтегральних показників рівня розвитку та знаходження репрезентантів в табличних редакторах *OpenOfficeCalc 3.2* та *Microsoft Office Excel 2007****.***

Завдання 1: побудова рейтингу компаній на основі інтегральних показників рівня розвитку. Необхідно за допомогою пакета *Calc* або *Excel* на підставі даних з табл. 1 побудувати рейтинг інвестиційної привабливості промислових підприємств. Впорядковування підприємств здійснити за методом рівня розвитку.

Таблиця 4.1

Дані щодо фундаментальних показників діяльності компаній

№	Назва компанії	№ групи	Співвідношення між ринковою і балансовою вартістю, %	Рентабельність власного капіталу, %	Співвідношення між прибутком і капіталізацією, %	Співвідношення між статутним фондом і власним капіталом, %	Коефіцієнт автономії %	Частка довгострокових зобов'язань, %
			X_1	X_2	X_3	X_4	X_5	X_6
1	Укрнафта	1	99	20,56	20,72	0,2	80	32,63
2	Маріупольський ММК ім. Ільїча	1	162,83	31,39	19,28	13,11	86,67	1,35
3	Запоріжсталь	1	124	21,97	17,7	5,7	82	21
4	Північний ДОК	1	165,39	54,6	33,01	24,26	46,69	0,05
5	Стирол	1	154,4	32,05	20,76	25,4	79,39	8,4
6	Нікопольський завод феросплавов	2	98,17	11,33	11,54	12,19	66,94	3,98
7	Західенерго	2	98	4,3	4,4	9,4	55	52
8	Харцизький трубний	2	184,37	0,63	0,34	8,34	64,36	4,21
9	Дніпрошина	2	42	12,31	29,6	10,9	61	41
10	Пивзавод "Славутич"	2	195,88	14,67	7,49	59,77	47,22	52,19

Завдання 2: пошук репрезентантів груп. Для наведених груп промислових підприємств (табл. 1), за методом центра тяжіння, визначити репрезентантів. Провести інтерпретацію отриманих результатів.

Лабораторна робота здійснюється в межах теми "Рішення задач зниження розмірності при знакового простору".

Методичні рекомендації до виконання завдання 1

1 крок. Вихідні дані для оцінки інвестиційної привабливості підприємств містять оцінки 10 великих компаній України за шістьома показниками x_1-x_6.

Оскільки показники неоднорідні, необхідно провести стандартизацію їх значень за формулою:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}, \; j=1..6,$$

де $\bar{x}_j$ – середнє арифметичне значення j-того показника;

S_j – середньоквадратичне відхилення j-того показника;

z_{ij} – стандартизоване значення j-того показника для і-того об'єкта.

Для розрахунку в пакеті *Calc* середнього арифметичного значення за показниками використаємо вбудовану функцію пакета AVERAGE(_). Для розрахунку середньоквадратичного відхилення – функцію STDEV(_) (рис. 1). В *Excel* цим функціям відповідають функції СРЗНАЧ(_) та СТАНДОТКЛОН(_). Розрахунок середніх значень показників та їх середньоквадратичних відхилень наведено на рис. 4.1.

Після розрахунку середніх та середньоквадратичних відхилень формуємо таблицю стандартизованих значень показників за об’єктами. Формулу вводимо в першу комірку таблиці, ставимо знаки $ для закріплення відповідних комірок, та розповсюджуємо введену формулу на всю таблицю.

2 крок. Наступним етапом розрахунку є виділення показників-стимуляторів та дестимуляторів. Показниками стимуляторами виступають ті показники, зростання значень яких відповідає покращенню рівня розвитку об’єктів, дестимуляторами – показники, що надають негативний вплив на рівень розвитку об’єктів.

STDEV | =(C3-C$13)/C$14

	A	B	C	D	E	F	G	H
1	№	Назва компанії	Співвідношення між ринковою і балансовою вартістю %	Рентабельність власного капіталу %	Співвідношення між прибутком і капіталізацією %	Співвідношення між статутним фондом і власним капіталом %	Коефіцієнт автономії %	Частка довгострокових зобов'язань %
2			x1	x2	x3	x4	x5	x6
3	1	Укрнафта	99,00	20,56	20,72	0,20	80,00	32,63
4	2	Маріупольський ММК ім. Ільїча	162,83	31,39	19,28	13,11	86,67	1,35
5	3	Запорожсталь	124,00	21,97	17,70	5,70	82,00	21,00
6	4	Північний ГОК	165,39	54,60	33,01	24,26	46,69	0,05
7	5	Стірол	154,40	32,05	20,76	25,40	79,39	8,40
8	6	Нікопольський завод феросплавів	98,17	11,33	11,54	12,19	66,94	3,98
9	7	Западенерго	98,00	4,30	4,40	9,40	55,00	52,00
10	8	Харцизський трубний	184,37	0,63	0,34	8,34	64,36	4,21
11	9	Днепрошина	42,00	12,31	29,60	10,90	61,00	41,00
12	10	Півзавод Славутіч	195,88	14,67	7,49	59,77	47,22	52,19
13	Середнє значення показника за вибіркою		132,40	20,38	16,48	16,93	66,93	21,68
14	Середньоквадратичне значення показника		45,76	15,03	10,01	16,03	13,87	20,05
15								
16	I. Таблиця стандартизованих значень показників							
17	№	Назва компанії	x1	x2	x3	x4	x5	x6
18	1	Укрнафта	=(C3-C$13)/C$14		0,42	-1,04	0,94	0,5
19	2	Маріупольський ММК ім. Ільїча	0,66	0,73	0,28	-0,24	1,42	-1,0
20	3	Запорожсталь	-0,18	0,11	0,12	-0,70	1,09	-0,0
21	4	Північний ГОК	0,72	2,28	1,65	0,46	-1,46	-1,0
22	5	Стірол	0,48	0,78	0,43	0,53	0,90	-0,6
23	6	Нікопольський завод феросплавів	-0,75	-0,60	-0,49	-0,30	0,00	-0,8
24	7	Западенерго	-0,75	-1,07	-1,21	-0,47	-0,86	1,5
25	8	Харцизський трубний	1,14	-1,31	-1,61	-0,54	-0,19	-0,8
26	9	Днепрошина	-1,98	-0,54	1,31	-0,38	-0,43	0,9
27	10	Півзавод Славутіч	1,39	-0,38	-0,90	2,67	-1,42	1,5

Рис. 4.1. **Розрахунок стандартизованих значень показників**

3 крок. Після визначення показників-стимуляторів та дестимуляторів слід побудувати еталонну точку $p_0(x_{01}, x_{02}, \ldots x_{0j}, \ldots x_{0m})$, $j=1..m$.

Еталонна точка в цьому завданні – це таке штучне підприємство, яке характеризується найкращими значеннями за кожним з показників x_1-x_6 серед всіх інших досліджуємих підприємств.

Тобто, якщо показник x_j є стимулятором, то значення цього показника для точки еталона розраховується як $x_{0j} = \max_i x_{ij}$, а якщо x_j є дестимулятором, то розраховується як $x_{0j} = \min_i x_{ij}$ (рис. 4.2).

MIN | =MAX(C18:C27)

	A	B	C	D	E	F	G	H
28								
29	II. Побудова точки-еталона							
30			x1	x2	x3	x4	x5	x6
31	Точка еталон Ро		=MAX(C18:C27)	2,28	1,65	2,67	1,42	-1,08

Рис. 4.2. **Розрахунок значень показників x_1-x_6 для точки еталона**

4 крок. Після побудови точки еталона слід розрахувати відстані між окремими точками, що характеризують об'єкти аналізу, і точкою еталоном. Чим ближче підприємство до точки еталона – тим більш розвинутим воно є.

Відстань між об'єктом та еталоном розраховується за формулою Евклідової відстані:

$$d_{0i} = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{0j})^2} \ .$$

Для того, шоб пришвидшити процес розрахунку відстаней, достатньо спочатку розрахувати відхилення об'єкта від еталона за кожним показником, а потім їх сумувати та взяти корінь із цієї суми (рис. 4.3).

Для цього в комірках C35:H44 розраховуємо відхілення за показниками, в комірках I35:I44 отримуємо Евклідову відстань, використовуючи вбудовані функції пакету Calc SQRT(_) та SUM(_) (в Excel їм відповідають функції КОРЕНЬ(_) та СУММ(_). Функція SQRT /КОРЕНЬ надає значення кореня з введеного числа, функція SUM /СУММ розраховує суму значень виділеного діапазону.

5 крок. Визначити значення таксономічних показників рівня розвитку можна за наступною формулою:

$$K_i = 1 - \frac{d_{0i}}{d}, \ (d = \bar{d}_0 + 2\sigma_0),$$

де $d_0 = \frac{\sum_{i=1}^{n} d_{oi}}{n}$ – середнє арифметичне значення евклідових відстаней між об'єктами та еталоном;

$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (d_{0i} - \bar{d}_0)^2}{n}}$ – середнє квадратичне відхилення значень евклідових відстаней між об'єктами та еталоном.

Розраховані таким чином значення інтегрального показника рівня розвитку містяться в останньому стовбці таблиці на рис. 4.3.

Інтерпретація значень показника рівня розвитку наступна: чим ближче значення показника до 1, тим на більш високому рівні розвитку знаходиться об'єкт, чим ближче до 0 – тим на більш низькому ($K_i \in [0,1]$).

STDEV | =(C18-C$31)^2

	A	B	C	D	E	F	G	H	I	J
15										
16	I. Таблиця стандартизованих значень показників									
17	№	Назва компанії	x1	x2	x3	x4	x5	x6		
18	1	Укрнафта	-0,73	0,01	0,42	-1,04	0,94	0,55		
19	2	Маріупольський ММК ім. Ільїча	0,66	0,73	0,28	-0,24	1,42	-1,01		
20	3	Запорожсталь	-0,18	0,11	0,12	-0,70	1,09	-0,03		
21	4	Північний ГОК	0,72	2,28	1,65	0,46	-1,46	-1,08		
22	5	Стірол	0,48	0,78	0,43	0,53	0,90	-0,66		
23	6	Нікопольський завод феросплавів	-0,75	-0,60	-0,49	-0,30	0,00	-0,88		
24	7	Западенерго	-0,75	-1,07	-1,21	-0,47	-0,86	1,51		
25	8	Харцизський трубний	1,14	-1,31	-1,61	-0,54	-0,19	-0,87		
26	9	Днепрошина	-1,98	-0,54	1,31	-0,38	-0,43	0,96		
27	10	Півзавод Славутіч	1,39	-0,38	-0,90	2,67	-1,42	1,52		
28										
29	II. Побудова точки-еталона									
30			x1	x2	x3	x4	x5	x6		
31		Точка еталон Ро	1,39	2,28	1,65	2,67	1,42	-1,08		
32										
33	III. Розрахунок відстаней до точки-еталона								Евклідова відстань до точки-еталона	Значення інтегрального показника рівня розвитку
34	№	Назва компанії	x1	x2	x3	x4	x5	x6		
35	1	Укрнафта	=(C18-C$31)^2	5,13	1,51	13,80	0,23	2,64	5,272	0,285
36	2	Маріупольський ММК ім. Ільїча	0,52	2,38	1,88	8,47	0,00	0,00	3,641	0,506
37	3	Запорожсталь	2,47	4,71	2,34	11,37	0,11	1,09	4,701	0,362
38	4	Північний ГОК	0,44	0,00	0,00	4,91	8,30	0,00	3,695	0,499
39	5	Стірол	0,82	2,25	1,50	4,60	0,28	0,17	3,100	0,579
40	6	Нікопольський завод феросплавів	4,56	8,29	4,60	8,81	2,02	0,04	5,321	0,278
41	7	Западенерго	4,57	11,20	8,16	9,87	5,21	6,71	6,762	0,082
42	8	Харцизський трубний	0,06	12,89	10,65	10,29	2,59	0,04	6,043	0,180
43	9	Днепрошина	11,31	7,92	0,12	9,29	3,42	4,17	6,019	0,183
44	10	Півзавод Славутіч	0,00	7,06	6,50	0,00	8,08	6,76	5,329	0,277
45										
46								$\bar{d}_0$ =	4,933	
47								σ =	1,191	
48								d =	7,370	

Рис. 4.3. **Розрахунок Евклідових відстаней від об'єктів до точки еталона та інтегрального показника K_i**

За значеннями інтегрального показника рівня розвитку можна провести ранжування підприємств. Для цього формуємо таблицю, яка б містила тільки назви підприємств та значення інтегрального показника. Виділяємо всі дані таблиці та обираємо в пункті головного меню "Дані" пункт "Сортування". У вікні Сортування обираємо сортування за стовбцем, у якому містяться значення інтегрального показника. Сортуємо підприємства за убуванням (рис. 4.4).

50	IV. Побудова рейтінгу компаній за рівнем розвитку			
51	№	Назва компанії	Значення інтегрального показника рівня розвитку	Рейтінг
52	5	Стірол	0,579	1
53	2	Маріупольський ММК ім. Ільїча	0,506	2
54	4	Північний ГОК	0,499	3
55	3	Запорожсталь	0,362	4
56	1	Укрнафта	0,285	5
57	6	Нікопольський завод феросплавів	0,278	6
58	10	Півзавод Славутіч	0,277	7
59	9	Днепрошина	0,183	8
60	8	Харцизський трубний	0,180	9
61	7	Западенерго	0,082	10

Рис. 4.4. **Побудова рейтингу підприємств**

Таким чином, найбільш інвестиційно привабливим за цим набором показників є підприємство Стирол, що набуло першого місця в рейтингу за значенням інтегрального показника К=0,579.

Методичні рекомендації до виконання завдання 2

1 крок. Знаходимо матриці відстаней між об'єктами в групі, скориставшись формулою Евклідової відстані:

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

де d_{ij} – відстань між i-тим та k-тим об'єктами,

x_{ik} – значення k-того показника для і-того об’єкта,

x_{jk} – значення k-того показника для j-того об’єкта.

Для першої групи матриця відстаней представлена діапазоном C14:G18, для другої групи підприємств – діапазоном C21:G25 (рис. 4.5).

H15 =СУММ(C15:G15)

	A	B	C	D	E	F	G	H	I
1	№	Назва компанії	№ групи	x1	x2	x3	x4	x5	x6
2	1	Укрнафта	1	99	20,56	20,72	0,2	80	32,63
3	2	Маріупольський ММК ім. Ільїча	1	162,83	31,39	19,28	13,11	86,67	1,35
4	3	Запорожсталь	1	124	21,97	17,7	5,7	82	21
5	4	Північний ГОК	1	165,39	54,6	33,01	24,26	46,69	0,05
6	5	Стірол	1	154,4	32,05	20,76	25,4	79,39	8,4
7	6	Нікопольський завод феросплавів	2	98,17	11,33	11,54	12,19	66,94	3,98
8	7	Западенерго	2	98	4,3	4,4	9,4	55	52
9	8	Харцизський трубний	2	184,37	0,63	0,34	8,34	64,36	4,21
10	9	Днепрошина	2	42	12,31	29,6	10,9	61	41
11	10	Півзавод Славутіч	2	195,88	14,67	7,49	59,77	47,22	52,19

13	Група 2	Нікопольський завод феросплавів	Западенерго	Харцизський трубний	Днепрошина	Півзавод Славутіч	Сума відстаней від об'єкта до всіх інших об'єктів
14	Нікопольський завод феросплавів	0,000	50,564	87,703	69,926	120,631	328,824
15	**Западенерго**	50,564	0,000	99,309	63,202	110,884	**323,959**
16	Харцизський трубний	87,703	99,309	0,000	150,443	74,978	412,433
17	Днепрошина	69,926	63,202	150,443	0,000	163,942	447,512
18	Півзавод Славутіч	120,631	110,884	74,978	163,942	0,000	470,434

20	Група 1	Укрнафта	Маріупольський ММК ім. Ільїча	Запорожсталь	Північний ГОК	Стірол	Сума відстаней від об'єкта до всіх інших об'єктів
21	Укрнафта	0	73,37047635	28,38336485	92,01789989	66,5108014	260,2825425
22	Маріупольський ММК ім. Ільїча	73,37047635	0	45,40750158	49,5800716	18,09519	186,4532395
23	Запорожсталь	28,38336485	45,40750158	0	71,00998028	39,859655	184,6605018
24	Північний ГОК	92,01789989	49,5800716	71,00998028	0	43,8138928	256,4218445
25	**Стірол**	66,51080138	18,09518997	39,85965504	43,81389277	0	**168,2795392**

Рис. 4.5. **Знаходження репрезентантів груп підприємств**

2 крок. Знаходимо суми відстаней від кожного об’єкта до всіх інших об’єктів відповідної групи: $d_i = \sum_j d_{ij}$ (стовбець H рис. 2.5). У пакеті для цього використовуємо вбудовану функцію SUM(_) (для Calc) або СУММ(_) (для Excel).

3 крок. Репрезентантом групи є об'єкт, якому відповідає мінімальна сума відстаней до інших об’єктів, тобто $\min d_i$. Якщо в групі знаходиться 2 елементи з однаковою сумою відстаней, то визначаються суми відстаней від цих елементів до раніше виділених репрезентантів, та як репрезентант обирається елемент з max відстанню.

За результатами розрахунків (рис. 4.5) репрезентантом першої групи є підприємство Західенерго, якому відповідає найменша сума відстаней до всіх інших об’єктів цієї групи. Репрезентантом групи 2 є підприємство Стирол.

Таким чином, маємо найбільш типових представників двох груп підприємств.

Лабораторна робота № 5. Побудова та використання дерев класифікації

Мета – закріплення теоретичного та практичного матеріалу з виявлення логічних закономірностей у даних, придбання навичок роботи в системі обробки даних *See5*

Завдання: використовуючи систему обробки даних *See5*, на підставі вхідних даних про результати роботи відкритих акціонерних товариств України побудувати модель класифікації їх інвестиційної привабливості:

1. Побудувати модель класифікації, провести її верифікацію.
2. Навести конфігурацію дерева та всі його характеристики за найбільш адекватною моделлю.
3. Сформувати та навести логічні правила віднесення спостережень до кожного з класів інвестиційної привабливості.
4. Дати економічну інтерпретацію результатів моделювання.

Лабораторна робота здійснюється у рамках теми "Побудова та використання дерев класифікації".

Методичні рекомендації до виконання завдання

Одним з найвагоміших серед сукупності методів для вирішення завдань класифікації є метод дерев класифікації. Цей метод дозволяє прогнозувати приналежність спостережень або об'єктів до певного класу категоріальної залежної змінної в залежності від відповідних значень однієї або декількох незалежних (предикторних) змінних. Дерева класифікації є більш гнучким засобом, ніж традиційні методи аналізу, оскільки вони здатні послідовно вивчати ефект впливу окремих змінних та виконувати одновимірне розгалуження на підставі предикторних змінних різних типів. Дерева класифікації використовують рекурсивний підхід, який дозволяє уникнути обмеження лінійного дискримінантного аналізу щодо максимальної кількості лінійних дискримінантних функцій.

Розглянемо порядок побудови моделей класифікації в системі обробки даних See5 на прикладі моделі класифікації інвестиційної привабливості акціонерних товариств.

1. **Створення файла імен змінних**. Для коректної роботи в системі See5 необхідно створити два файли – імен змінних і даних. Для формування файла змінних створюють новий текстовий документ за допомогою програми Блокнот, яку завантажують за допомогою меню *Пуск/Программы/Стандартные/Блокнот*, або через контекстне меню *Создать/Текстовый документ*. Створений файл має бути збережено з розширенням **.names* та типом файлу – *Все файлы.*

У файлі імен змінних подають назви змінних і класів, які використовують. Серед типів змінних розрізняють такі класи:

- *discrete* – номінальні ознаки;
- *continuous* – кількісні ознаки і мітки;
- *data* – дата у форматі *рік/місяць*/день або *рік-місяць-день*;
- *time* – час у форматі *години: хвилини: секунди;*
- *timestamp* – дата і час у форматі *рік/місяць/день* або *рік-місяць-день години: хвилини: секунди;*
- *discrete* – список дискретних значень змінної, розділеної комою;
- *ignore* – для ознаки, що виключається з аналізу;
- *label* – мітка для ідентифікації окремого об'єкта.

Значення явно визначених ознак безпосередньо використовують з файла даних, а неявно визначені ознаки задають через формули. Після

імені кожної неявно заданої змінної записують двокрапку і далі записують формулу. У формулі використовують дужки, а дискретні ознаки обмежуються лапками. Для цього використовують оператори:

*+, -, *, /, % (mod), <, >, <=, >=, <>, !=; and, or;*

sin( ), cos( ), tan( ), log( ), exp( ), int( );

залежно від формули, яка використовується, кінцевий результат може бути як кількісним, так і давати логічні значення *true/false*.

Для прикладу цільова ознака *class* набуває три значення: 1 – високий клас інвестиційної привабливості; 2 – середній клас інвестиційної привабливості; 3 – низький клас інвестиційної привабливості. Ознаки К1 – капіталізація на кінець року (млн грн), К2 – коефіцієнт автономії (%) – є кількісними. Слід зазначити, що під час підготовки файла імен змінних, пробіли, порожні рядки і знаки табуляції ігноруються системою (крім випадків, коли вони застосовуються в іменах змінних). Вертикальна риска "|" призначена для запису нагадувань або коментарів.

Таким чином, для розглянутого прикладу файл імен змінних *Primer.names* у прикладі виглядає таким чином:

class | the target attribute

class: 1,2,3.

К1: continuous

К2: continuous

2. Створення файлів даних. Фрагмент вхідних даних наведено в таблиці 5.1:

Таблиця 5.1

Вхідні дані

№ п/п	Назва ВАТ	Клас	Капіталізація на кінець року	Коефіцієнт автономії
1	2	3	4	5
1	Укрнафта	1	2 219.6	80
2	Маріупольський ММК ім. Ілліча	1	1 805.3	86.67
3	Запоріжсталь	1	864.6	82
4	Північний ГЗК	1	1536.8	46.69
5	Стирол	1	760.6	79.39

Закінчення табл. 5.1

1	2	3	4	5
6	Концерн "Галнафтогаз"	1	179.4	58.35
7	Міттал Стіл Кривий Ріг	1	3 171.7	90.05
	...	…	…	…
50	Хмельницькобленерго	3	53.3	62.47
51	Черніговобленерго	3	47.6	65.23
52	Крименерго	3	34.6	7.51
53	Південний ГЗК	3	287.6	45.24
54	Луганськтепловоз	3	53.2	38.89
55	Харківський тракторний завод	3	40.6	31.12
56	Дніпренерго	3	271.2	11.5
57	Полтаваобленерго	3	142.2	65.4

Для обробки даних системою *See5*, необхідно здійснити їх конвертацію. Для цього необхідно виділити таблицю та конвертувати в текст командою меню: *Таблица/Преобразовать/Таблицу в текст* зі знаком табуляції ",":

1,2219.6,80
1,1805.3,86.67
1,864.6,82
1,1536.8,46.69
1,760.6,79.39
1,179.4,58.35
1,3171.7,90.05
...
3,53.2,38.89
3,40.6,31.12
3,271.2,11.5
3,142.2,65.4
3,137.4,17.15
3,68.6,28.88
3,219.4,44
3,176.6,26.88
3,42.2,23.4
3,234.3,-7.06

Отримані дані, необхідно скопіювати в новий файл *Блокнота* з ім'ям *Primer.data*. Зберегти файл необхідно з типом *Все файлы*.

Для перевірки якості побудованого дерева рішень і відповідної множини логічних правил у системі *See5* передбачена можливість роботи зі спеціальними файлами, у яких містяться додаткові тестові дані – контрольна вибірка. Даний файл *Primer.test* є необов'язковим і, якщо використовується, має формат файла *Primer.data*. Інший допоміжний файл *Primer.cases* також є додатковим. Він містить об'єкти з невідомою класифікацією.

Файл, що має розширення *.costs*, містить інформацію про вартість різних помилок класифікації. Заповнення цього файла є необов'язковим. Разом з тим, призначення штрафів за помилки може бути дуже корисним для врахування економічних втрат з ухваленням неправильного управлінського рішення.

3. Розрахунки в системі.

3.1. Побудова конфігурації дерева. Запустіть програму *See5*. За допомогою пункту меню *File/Locate Data* завантажте файл вхідних даних *Primer.data*. Якщо завантаження відбулось, повідомлення *class and attribute definitions, training cases to be analyzed* стають явними.

У головному вікні системи *See5* розташовано п'ять кнопок (рис. 5.1).

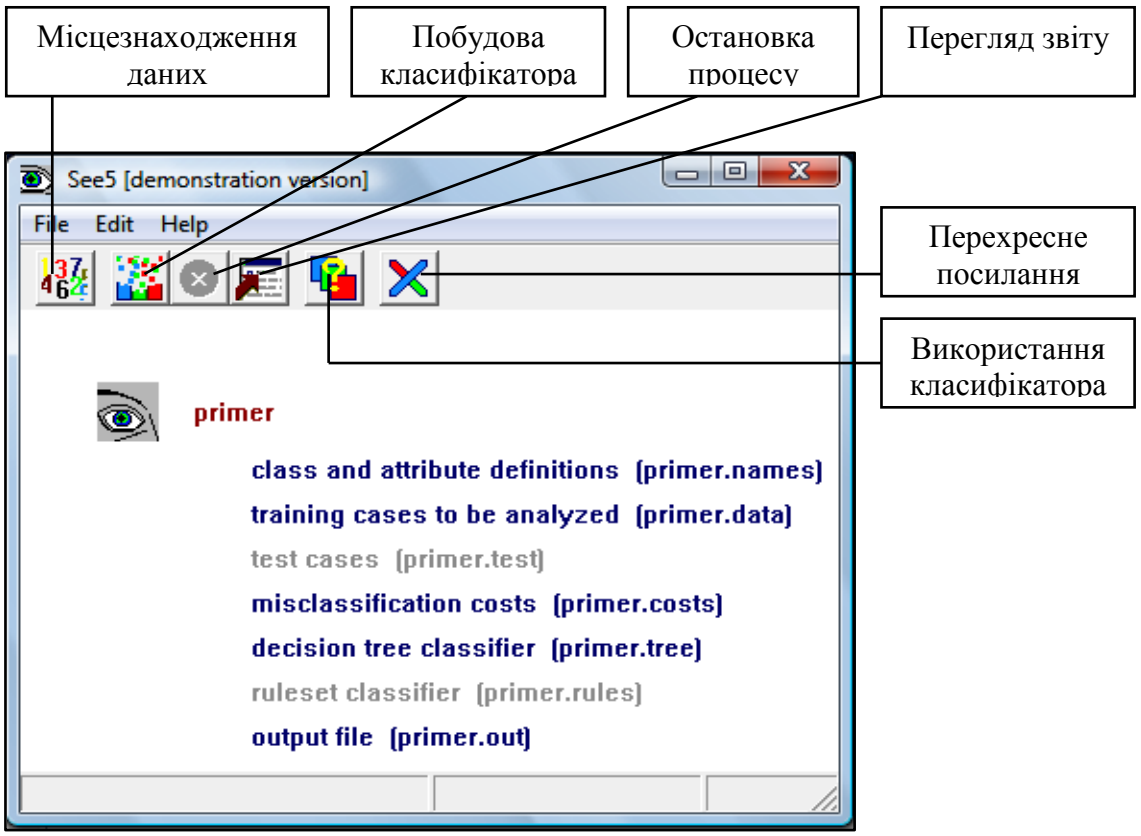


Рис. 5.1. **Головне вікно системи**

За допомогою кнопки *Locate Data* (місцезнаходження даних) викликається вікно для перегляду доступних файлів даних і їх завантаження в систему. Ініціалізацією кнопки *Construct Classifier* (побудова класифікатора) виконується звертання до вікна діалогу для вибору типу класифікатора й встановлення його параметрів. Кнопка *Stop* призначена для остановки процесу побудови дерева рішень. Кнопка *Use Classifier* (використання класифікатора) запускає процес інтерактивної класифікації одного або більше об'єктів.

За допомогою кнопки *Cross-Reference* (перехресне посилання) викликається вікно, у якому наочно розкриваються зв'язки між об'єктами навчальної вибірки і знайдених правил їх класифікації. Усі описані функції доступні також з меню *File*. У меню *Edit* надається можливість редагування файла імен даних і файла вартості помилок класифікації.

Для виклику вікна параметрів класифікації *Classifier Construction Options* (рис. 5.2) ініціалізуйте кнопку *Construct Classifier*  .

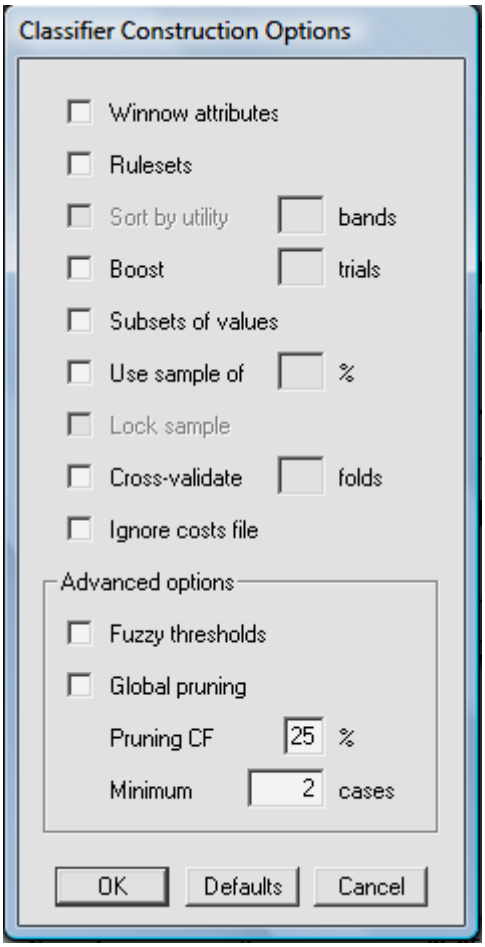


Рис. 5.2. **Вікно параметрів класифікації**

З натисканням кнопки ОК система генерує вікно результатів (рис. 5.3).

У першому рядку звіту про результати класифікації надається інформація про версію системи *See5*, яка використовується, поточний час. У наступних двох рядках розташовано інформацію, яка змінна є класифікуючою (змінна *class*) і прочитаний файл даних *Primer.data* містить 63 об'єкта, кожний з який описаний трьома ознаками (атрибутами). У наступних рядках звіту міститься побудоване дерево рішень. Кожна гілка дерева, яка має термінальну вершину, містить відповідний номер класу інвестиційної привабливості. За кожним номером класу у записі виду (*n*) або (*n**m*) міститься інформація про загальну кількість об’єктів, які належать цьому класу та кількість неправильно класифікованих об’єктів. Так, наприклад, перша гілка закінчується записом (12/2). Це означає, що цій гілці відповідає 12 об’єктів з визначеного (першого) класу, з яких 1 попадає помилково. Величини *n* або *m* можуть бути дробовими у випадку, коли до певної гілки належить деяка кількість об'єктів з невідомими значеннями ознак.

```
See5 [Release 2.08]     Wed Oct 12 22:12:24 2011
 Options:        Do not use global pruning
Class specified by attribute `class'
Read 63 cases (3 attributes) from primer.data
Decision tree:
K1 > 341.9: 1 (12/2)
K1 <= 341.9:
:...K2 <= 45.24: 3 (21/5)
    K2 > 45.24: 2 (30/15)
Evaluation on training data (63 cases):
          Decision Tree
        Size      Errors
          3   22(34.9%)   <<
         (a)   (b)   (c)    <-classified as
        ----  ----  ----
          10    10     1    (a): class 1
           2    15     4    (b): class 2
                5    16    (c): class 3
      Attribute usage:
         100%  K1
          81%  K2

Time: 0.0 secs
```

Рис. 5.3. **Результати класифікації**

У наступному розділі звіту наводяться характеристики побудованої моделі класифікації, які розраховані на навчальній вибірці. Так, з прикладу видно, що побудоване дерево рішень має 3 гілки (*size* = 3), а помилка класифікації спостерігається для 22 об'єктів, що складає 34,9 %.

У останній частині звіту подається таблиця класифікації. За даними цієї таблиці можна сказати, що з 1-го класу інвестиційної привабливості правильно класифіковано 10 об'єктів, 10 об'єктів помилково відносяться до класу 2, та 1 об’єкт до класу 3. Серед об'єктів 2-го класу інвестиційної привабливості 15 діагностовано правильно, 2 помилково належить першому класу, 4 об’єкти помилково належить третьому класу. Для 3-го класу класифіковано правильно 16 об’єктів і 5 помилково визнано об’єктами другого класу. На основі отриманих результатів класифікації конфігурація дерева за тренувальними даними матиме такий вигляд (рис. 5.4).

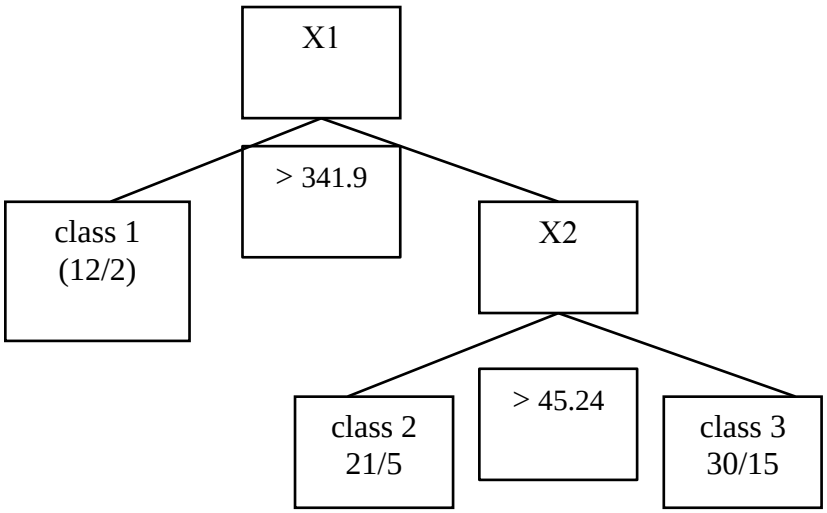


Рис. 5.4. **Конфігурація дерева класифікації**

3.2. Побудова правил класифікації. У випадку, якщо отримане дерево рішень є дуже складним для сприйняття, наприклад під час рішення задач високої розмірності для неоднорідних даних, у системі *See5* передбачена можливість конструювання набору правил *If ... then*. Для цього потрібно викликати вікно параметрів класифікації *Classifier*

Construction (див. рис. 5.2) і встановити прапорець у поле *Rulesets* (набір правил). Після проведення такої операції система додає у вікно звіту список правил, що відповідають побудованому дереву класифікації. За даними прикладу це буде така низка логічних правил (рис. 5.5).

```
Rules:
Rule 1: (12/2, lift 2.4)
        K1 > 341.9
        ->  class 1  [0.786]
Rule 2: (30/15, lift 1.5)
        K1 <= 341.9
        K2 > 45.24
        ->  class 2  [0.500]
Rule 3: (21/5, lift 2.2)
        K1 <= 341.9
        K2 <= 45.24
        ->  class 3  [0.739]
```

Рис. 5.5. **Правила класифікації**

Кожне правило містить такі елементи:

- номер правила (наприклад, Rule 2:);
- кількість об'єктів навчальної вибірки, які належать певному класу за правилом (наприклад, для другого правила – 30/15 – 30 об’єктів належать класу 2, 15 з який віднесено помилково);
- одна або кілька елементарних логічних подій, що входять до складу правила або складного логічного висловлення (для прикладу – у правилі № 2 логічна подія така: якщо K1 <= 341.9 ∩ K2 > 45.24, тоді об’єкт належить класу № 2;
- номер класу, якому відповідає дане правило;
- величина, що набуває значення від 0 до 1, що відбиває ступінь довіри до правила – характеристика точності правила (для правила № 2 – 0,5).

Побудовані множини правил використовується для ухвалення рішення про приналежність того або іншого об'єкта певному класу. Разом з цим існують ситуації, коли один і той самий об'єкт підпадає під дію відразу декількох правил, у тому числі правил, що описують різні класи. Такі внутрішні конфлікти може бути вирішено двома способами. За першим способом перевага надається тому правилу, яке має більш високий ступінь довіри (більш високу точність). Інший спосіб пов'язаний з узагальненням результатів різних правил для прийняття остаточного рішення. У системі *See5* існує інший спосіб — кожне правило, що спрацювало, подає запис для віднесення певного об'єкта до одного з класів. Записи додаються з вагами, рівними обчисленим ступеням довіри, і об’єкт вважається приналежним до класу, для якого набирається найбільша зважена сума записів.

3.3. Встановлення додаткових параметрів аналізу.

3.3.1. Рівень довіри та складність дерева. Для додаткового аналізу побудованого дерева класифікації та пошуку оптимальної конфігурації дерева у вікні параметрів класифікації системи передбачено параметр рівня довіри *pruning CF*, призначений для відсікання статистично не значимих гілок дерева класифікації. За замовчуванням система встановлює значення рівня довіри 25 %.

На точність класифікації може істотно впливати параметр *Minimum...cases*. У поле цього параметра встановлюється число, яке обмежує мінімальну кількість об'єктів на гілці дерева класифікації. Чим менше буде це число, тим більше за розміром стане дерево і тим точніше виконується підгонка дерева під необхідну класифікацію.

3.3.2. Посилення точності класифікації. Під час посилення рішення конструюється не одне, а декілька дерев класифікації. Головна вимога до таких дерев полягає в тому, що вони мають якнайменше дублювати один одного. Для цього на першому кроці конструюється початкове дерево рішень (див. рис. 5.3). Як випливає з поданих результатів, класифікація, виконана на основі початкового дерева, дає помилки на деяких об'єктах. Так, у нашому випадку спостерігається 22 помилки на 63 об'єктах навчальної вибірки (див. рис. 5.3). На другому кроці отримують істотну відмінність нового дерева від початкового.

Побудоване дерево також буде приводити до помилкових рішень, але вже на інших об’єктах. На кожному з наступних кроків роботи алгоритму нове дерево будується із врахуванням помилок усіх попередніх дерев рішень. Для запуску процесу посилення рішення потрібно встановити прапорець *Boost* у діалоговому вікні параметрів класифікації (див. рис. 5.2). У цьому ж вікні в полі *trials* необхідно задати загальну кількість дерев рішень. Зрозуміло, що побудова нової множини дерев рішень вимагає додаткового часу. Розроблювачі See5 стверджують, що з використанням 10 дерев рішень помилки класифікації знижуються в середньому на 25 %. За даними прикладу встановимо в полі *trials* 3 спроби побудови дерева. Ініціалізацією кнопки ОК отримуємо вікно звіту з інформацією про результати рішення (рис. 5.6).

Як видно зі звіту, перше дерево класифікує дані з помилкою 34,9 % друге й третє дерево – з помилкою 36,5 %. Але всі три дерева рішень разом класифікують дані з помилкою 33,3 (запис у рядку *boost*).

Незважаючи на високу швидкодію системи *See5*, конструювання дерев на повному наборі вхідних даних за їх великої кількості може займати досить багато часу. Це стає особливо помітним під час використання додаткових параметрів алгоритму, наприклад параметра для посилення рішення (*boosting*). Для вирішення цієї проблеми в системі *See5* існує можливість роботи не з повним набором даних, а з певною підвибіркою з вхідного набору. Для цього передбачений спеціальний параметр *Use sample of* (див. рис. 5.2). Під час використання цього параметра з вихідного набора випадковим чином формується x % об'єктів і за допомогою їх будується дерево класифікації. Подальше тестування побудованого дерева відбувається на іншій відмінній від вхідної вибірці обсягом x %, якщо $x < 50$ %, або на всіх об'єктах, що залишилися, якщо $x > 50$ %. На рис. 5.7 наведено результат побудови дерева рішень на вибірці з 60 % від обсягу вихідних даних.

```
See5 [Release 2.08]     Wed Oct 12 22:15:41 2011
Options:        3 boosting trials
Do not use global pruning
Class specified by attribute `class'
Read 63 cases (3 attributes) from primer.data
-----  Trial 0:  -----
Decision tree:
K1 > 341.9: 1 (12/2)
K1 <= 341.9:
:...K2 <= 45.24: 3 (21/5)
K2 > 45.24: 2 (30/15)
-----  Trial 1:  -----
Decision tree:
K1 > 341.9: 1 (11.3/2.4)
K1 <= 341.9:
:...K2 <= 45.24: 3 (20.2/6.1)
K2 > 45.24:
:...K1 <= 58.5: 2 (6/2.4)
K1 > 58.5: 1 (25.5/13.4)
-----  Trial 2:  -----
Decision tree:
K2 <= 46.88: 3 (17.4/3.4)
K2 > 46.88:
:...K1 <= 543.4: 2 (31.3/13)
K1 > 543.4: 1 (5.4)
Evaluation on training data (63 cases):
Trial       Decision Tree
-----     ----------------
Size      Errors
0         3   22(34.9%)
1         4   23(36.5%)
2         3   23(36.5%)
boost           21(33.3%)   <<
(a)   (b)   (c)    <-classified as
----  ----  ----
10     9     2    (a): class 1
2    15     4    (b): class 2
4    17    (c): class 3
Attribute usage:
100%  K1
100%  K2
Time: 0.0 secs
```

Рис. 5.6. **Результати класифікації з посиленням рішення**

На навчальній вибірці досягнуто ефект класифікації – помилка складає 31,6 %. Разом з тим, на контрольній вибірці, обсяг якої дорівнює половині обсягу вхідних даних, відсоток правильної класифікації різко підвищується до 48 %. Це свідчить про необхідність продовжити пошук більш стійкого варіанта рішення.

```
Options:        Use 60% of data for training        Do not use global pruning
   Class specified by attribute `Klass'
   Read 38 cases (3 attributes) from primer.data
   Decision tree:
   K2 <= 31.12: 3 (7/1)
   K2 > 31.12:
   :...K1 > 58.5: 1 (23/11)
      K1 <= 58.5:
      :...K1 <= 46.4: 2 (2)
         K1 > 46.4:
         :...K1 <= 53.3: 3 (4)
            K1 > 53.3: 2 (2)

   Evaluation on training data (38 cases):
            Decision Tree
          ----------------
          Size      Errors
            5   12(31.6%)   <<
           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            12              (a): class 1
             9     4     1    (b): class 2
             2           10    (c): class 3

         Attribute usage:
            100%  K2
             82%  K1

   Evaluation on test data (25 cases):
            Decision Tree
          ----------------
          Size      Errors
            5   12(48.0%)   <<

           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            8           1    (a): class 1
            5     1     1    (b): class 2
            5           4    (c): class 3
```

Рис. 5.7. **Результати класифікації за підвибіркою**

3.3.3. Врахування вартості помилок класифікації. За замовченням вважають, що всі види помилок класифікації є

еквівалентними. Загальну оцінку якості побудованого дерева класифікації обчислюють, підраховуючи загальне число помилок. Якщо вартість різних помилок вважають різною, тоді для обліку їх вартості створюють спеціальний файл, для нашого прикладу *Primer.costs*. Він містить рядки такого вигляду:

прогнозний клас, клас вхідних даних: вартість помилки,

де вартість помилки – додатне дійсне число.

Число рядків, що характеризують прогнозний клас – клас вхідних даних, у файлі може бути будь-яким. Якщо вартість помилки не визначено явно, то система призначає вартість, яка дорівнює 1.

Для прикладу, якщо вартість помилкового віднесення до класу низької інвестиційної привабливості буде дорівнювати 20, а вартість всіх інших видів помилок дорівнювати 4, тоді файл різної вартості помилок Primer.costs може мати такий вигляд:

```
                |costs file for Primer
1, 2: 4         | вартість помилкового віднесення класу 2 до класу 1
1, 3: 4         | вартість помилкового віднесення класу 3 до класу 1
2, 1: 4         | вартість помилкового віднесення класу 1 до класу 2
2, 3: 4         | вартість помилкового віднесення класу 3 до класу 2
3, 1: 20        | вартість помилкового віднесення класу 1 до класу 3
3, 2: 20        | вартість помилкового віднесення класу 2 до класу 3
```

3.3.4. Перехресна перевірка. Для обчислення надійних оцінок якості побудованої класифікації у системі *See5* використовується перехресна перевірка, яка здійснюється таким чином. Уся вибірка об'єктів розбивається на *m* множин однакової розмірності. Послідовно кожна множина використовується як контрольний набір об'єктів для тестування моделі, побудованої на основі тренувальних даних. Кількість множин вводиться в поле *Cross-validate* діалогового вікна для завдання опцій алгоритму конструювання класифікатора. Результат перехресної перевірки відбивається в нижній частині вікна звіту (рис. 5.8).

```
See5 [Release 2.08]     Wed Oct 12 22:16:27 2011
 Options:          Do not use global pruning          Cross-validate using 3 folds Class specified by
attribute `class' Read 63 cases (3 attributes) from primer.data
                                [ Fold 1 ] Decision tree:
                                K2 > 38.89:
                                :...K1 <= 543.4: 2 (25/12)
                                :   K1 > 543.4: 1 (6/1)
                                K2 <= 38.89:
                                :...K1 <= 169.1: 3 (7)
                                    K1 > 169.1:
                                    :...K2 <= 17.15: 3 (2)
                                        K2 > 17.15: 1 (2)
                                Evaluation on hold-out data (21 cases):
                                          Decision Tree
                                        ----------------
                                         Size      Errors
                                           5        10 (47.6 %)   <<
                                [ Fold 2 ] Decision tree:
                                K1 > 321: 1 (10/1)
                                K1 <= 321:
                                :...K2 <= 45.24: 3 (15/4)
                                    K2 > 45.24: 2 (17/8)
                                Evaluation on hold-out data (21 cases):
                                          Decision Tree
                                        ----------------
                                         Size      Errors
                                           3        10 (47.6 %)   <<
                                [ Fold 3 ] Decision tree:
                                K1 > 543.4: 1 (4)
                                K1 <= 543.4:
                                :...K2 <= 36.04: 3 (12/3)
                                    K2 > 36.04: 2 (26/14)
                                Evaluation on hold-out data (21 cases):
                                          Decision Tree
                                        ----------------
                                         Size      Errors
                                           3        7 (33.3 %)   <<
                                [ Summary ]
                                Fold       Decision Tree
                                ----     ----------------
                                         Size      Errors
                                  1        5       47.6%
                                  2        3       47.6%
                                  3        3       33.3%
                                 Mean    3.7     42.9%
                                 SE      0.7      4.8%
                                        (a)   (b)   (c)    <-classified as
                                       ----  ----  ----
                                          9    11     1    (a): class 1
                                          3    16     2    (b): class 2
                                          1     9    11    (c): class 3
```

Рис. 5.8. **Результат крос-перевірки**

3.3.5. Завдання м’яких порогів. У системі See5 передбачена ще одна можливість поліпшення якості класифікації. Ця можливість стосується не стільки точності результатів, скільки підвищення їх стійкості до можливих флуктуацій значень ознак. Вона пов'язана зі встановленням нечітких або м'яких порогів, на порівнянні з якими ґрунтується вибір тієї або іншої гілки дерева класифікації.

У діалоговому вікні для завдання параметрів алгоритму See5 (див. рис. 5.2) існує параметр для пом'якшення порогів – *fuzzy thresholds* (розмиті пороги). Під час звертання до нього замість одного порога задається три значення – нижня границя LB, верхня границя UB і центральне значення T. Якщо значення змінна набуває нижче LB або більше UB, то досліджуються відповідні гілки дерева. Якщо ж значення змінної належить інтервалу (LB; UB), то досліджуються одночасно дві гілки дерева і вибирається найбільш адекватний результат класифікації. Значення LB, UB і T система визначає автоматично.

За прикладом дерево рішень з розмитими порогами наведено на рис. 5.9. Тут кожен поріг подано у вигляді <= LB(T) або >= UB(T).

```
See5 [Release 2.08]     Wed Oct 12 22:20:45 2011
    Options:
        Probability thresholds
        Do not use global pruning
Class specified by attribute `class'
Read 63 cases (3 attributes) from primer.data
Decision tree:
K1 >= 864.6 (414.75): 1 (12/2)
K1 <= 321 (414.75):
:...K2 <= 45.01 (46.045): 3 (21/5)
    K2 >= 60.17 (46.045): 2 (30/15)
Evaluation on training data (63 cases):
          Decision Tree
        ----------------
        Size      Errors
          3   23(36.5%)   <<
         (a)   (b)   (c)    <-classified as
        ----  ----  ----
         10     8     3    (a): class 1
          2    13     6    (b): class 2
               4    17    (c): class 3
      Attribute usage:
        100%  K1
         87%  K2
```

Рис. 5.9. **Класифікація за мякими порогами**

3.4. Використання дерева класифікації. Після того як визначено конфігурацію дерева класифікації та перевірено її адекватність, стає можливим використання моделі в інтерактивному режимі для прогнозування класу на нових даних. Ініціювавши кнопку *Use Classifier,* активізуємо алгоритм класифікації даних, який відповідає останньому варіанту дерева класифікації. На екран виводиться вікно для введення прогнозних значень змінних.

Кількість змінних може бути різною, оскільки в залежності від результату реалізується та або інша гілка дерева класифікації. Після введення всіх значень на екран виводиться вікно, у якому вказуються прогнозний клас і рівень довіри до результату класифікації. За прикладом із завданням прогнозного значення змінної К1 = 2 000, отримуємо результат, за яким новий об'єкт належить до першого класу інвестиційної привабливості із рівнем довіри 0,83, до другого класу – із рівнем 0,17 (рис. 5.10).

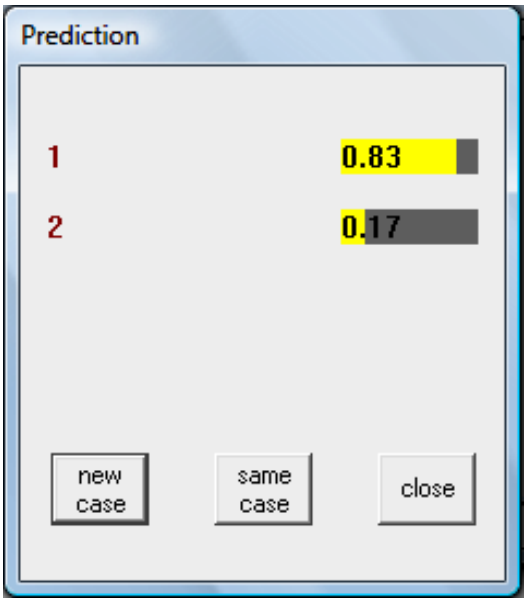


Рис. 5.10. **Вікно результатів прогнозування**

3.5. Поелементний перегляд крос-посилань. Для поелементного перегляду крос-посилань необхідно ініціювати кнопку Cross-Reference  у головному вікні програми. На екрані з'являється вікно вибору типу даних (рис. 5.11).

Натисканням кнопки ОК отримуємо вікно перехресних посилань, у лівій половині якого спочатку зображене дерево класифікації, а в правій подано список об'єктів, які класифіковано (рис. 5.12). Для поелементного

перегляду (рис. 5.13) для онлайнового об'єкта потрібно клацнути лівою кнопкою миші в правому полі вікна перехресних посилань на необхідному об'єкті – у лівому полі автоматично відбивається відповідна гілка. Так, для прикладу було вибрано об'єкт № 1 (рис.). Як бачимо, об'єктом 1 співвідноситься досить коротка гілка рішення (K2 > 36,04 ∩ K1 > 76). Аналогічним образом можна розібрати результати класифікації всіх інших доступних об'єктів вибірки.

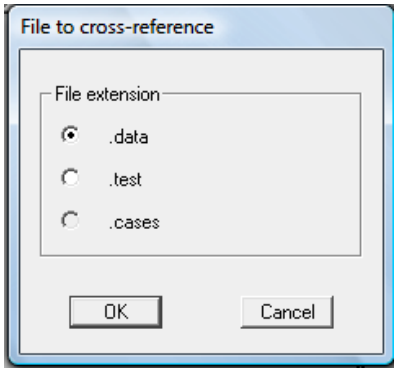


Рис. 5.11. **Вибір типу даних для крос-посилань**

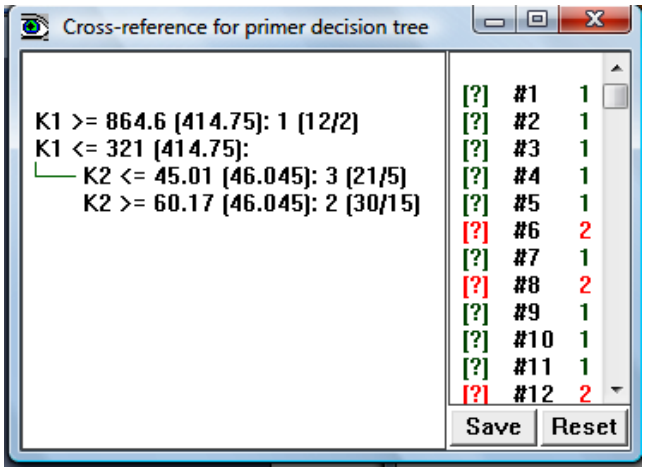


Рис. 5.12. **Вікно крос-посилань**

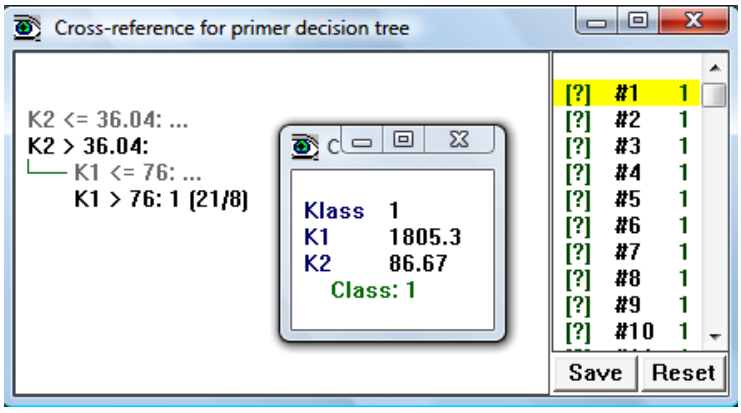


Рис. 5.13. **Вікно перегляду крос-посилань для вибраного об’єкта**

Рекомендована література

Основна

1. Андрейчиков А. В. Интеллектуальные информационные системы : учебник / А. В. Андрейчиков, О. Н. Андрейчикова. – М. : Финансы и статистика, 2004. – 424 с.

2. Барсегян А. А. Методы и модели анализа данных: OLAP и Data Mining / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург, 2004. – 336 с.

3. Дубров А. М. Многомерные статистические методы / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 1998. – 350 с.

4. Дюк В. Data Mining : учебный курс / Дюк В., Самойленко А. – СПб, 2001. – 368 с.

5. Лук'яненко І. Економетрика / І. Лук'яненко, Л. Красикова. – К. : Тов "Знання", КОО, 1998. – 494 с.

6. Магнус Я. Р. Эконометрика / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М. : Дело, 1997. – 248 с.

7. Плюта В. Сравнительный многомерный анализ в экономических исследованиях / В. Плюта. – М. : Статистика, 1980. – 151 с.

Додаткова

8. Бабешко Л. О. Основы эконометрического моделирования / Л. О. Бабешко. – М. : КомКнига, 2006. – 432 с.

9. Боровиков В. П. STATISTICA Статистический анализ и обработка данных в среде WINDOWS / В. П. Боровиков, И. П. Боровиков. – М. : Информационно-издательский дом "Филинъ", 1997. – 608 с.

10. Многомерный статистический анализ в экономике / Л. А. Сошникова, В. Н. Тамашевич, Г. Уебе, М. Шефер [под ред. проф. В. Н. Тамашевича. – М. : ЮНИТИ-ДАНА, 1999. – 598с.

НАВЧАЛЬНЕ ВИДАННЯ

Методичні рекомендації до виконання лабораторних робіт з навчальної дисципліни "МАТЕМАТИЧНІ МЕТОДИ І МОДЕЛІ ДОСЛІДЖЕННЯ ЕКОНОМІЧНИХ ПРОЦЕСІВ"

для студентів спеціальностей 8.03050803 "Оподаткування", 8.03050701 "Маркетинг", 8.14010301 "Туризмознавство" денної форми навчання

Укладачі: **Гур'янова** Лідія Семенівна
Сергієнко Олена Андріанівна
Нікіфорова Ольга Володимирівна та ін.

Відповідальний за випуск **Клебанова Т. С.**

Редактор **Бутенко В. О.**

Коректор **Бриль В. О.**

План 2012 р. Поз. № 304.
Підп. до друку Формат 60 х 90 1/16. Папір MultiCopy. Друк Riso.
Ум.-друк. арк. 4,0. Обл.-вид. арк. 5,0. Тираж прим. Зам. №

Видавець і виготівник — видавництво ХНЕУ, 61166, м. Харків, пр. Леніна, 9а

Свідоцтво про внесення до Державного реєстру суб'єктів видавничої справи ***Дк № 481 від 13.06.2001 р.***