Лабораторна робота №4

Програмна розробка методу кластеризації даних

Мета роботи: Здобути навички програмної розробки методу кластеризації даних.

Завдання: Розробити програму для ієрархічної кластеризації даних.

Вступ

У третій лабораторній роботі ви здобули навички використання алгоритмів кластеризації, які доступні у пакеті WEKA. У даній лабораторній роботі вам потрібно власноруч розробити програмну реалізацію одного з методів кластеризації даних — “Ієрархічна кластеризація”.

Метод ієрархічної кластеризації

Нехай вирішується задача кластеризації листів з електронної скриньки на наявність спаму. Усі електронні листи ще не позначені як “Спам” чи “Не спам”. Необхідно розробити програму, яка проаналізує назви листів і спробує визначити приналежність екземплярів слів до певних кластерів.

З першої лабораторної роботи вам відомо, що для класифікації екземплярів необхідно спершу позначити приналежність навчальних даних до класів, далі побудувати модель класифікації та перевірити модель на тестових даних. Якщо приналежність навчальних даних до класів є невідомою, тоді застосовують метод кластеризації даних.

У якості навчальних даних розглянемо назви електронних листів (див. рис. 1).

Назва 1: New meeting tomorrow (file)
Назва 2: Corporate party tomorrow
Назва 3: Free sales party
Назва 4: Free file for you
Назва 5: New greeting text
Назва 6: Free file upload

Рисунок 1. Навчальні дані

Виконувати кластерний аналіз можна за різними метриками слів, у тому числі: за кількістю букв у слові, відношенням слова до певної частини мови, за змістом. Складемо словник з переліку слів та кількості їх повторень та залишимо слова з кількістю повторень більше 1.

Таблиця 1. Словник слів та кількість їх повторень

Слово	Кількість
tomorrow	2
free	3
file	3
new	2
party	2
Інші	1

Нехай кожне слово з навчального набору речень можна описати двома атрибутами: загальною кількістю слів у реченні (атрибут_X) та загальною кількістю символів (не рахуючи пробіли та “(“ чи “)”) у реченні (атрибут_Y).

Таблиця 2. Перелік слів та їх атрибутів

Слово	атрибут_X	атрибут_Y
tomorrow	4	22
tomorrow	3	22
free	3	14
free	4	14
free	3	14
file	4	22
file	4	14
file	3	14
new	4	22
new	3	15
party	3	14
party	3	14

Тепер розрахуємо середні значення атрибутів для кожного слова.

Таблиця 3. Перелік слів та їх атрибутів

Слово	атрибут_X	атрибут_Y
tomorrow	3.5	22
free	3.33	14
file	3.67	16.67

new	3.5	18.5
party	3	14

При кластеризації схожість між екземплярами визначається метриками їх відносного розташування у просторі. Екземпляри, які описуються двома атрибутами можна зобразити у двомірному просторі Евкліда. Для розрахунку відстані між даними у просторі Евкліда прийнято застосовувати метрику відстані Евкліда, яку можна визначити наступним виразом:

$$D(A,B) = \sqrt{(B_x \quad A_x)^2 + \left(B_y \quad A_y\right)^2}, \tag{1}$$

де D — відстань між двома екземплярами у двомірному просторі, A та B — екземпляри даних, x та y — атрибути екземплярів.

Наступним кроком після розташування у просторі визначають два найближчі екземпляри, для цього будують таблицю відстаней (див. таблицю 4).

Таблиця 4. Таблиця відстаней слів, ітерація №1

Слово	tomorrow	free	file	new	party
tomorrow	0	sqrt(pow(3.5-3.33,2)+pow(22-14,2))	sqrt(pow(3.5-3.67,2)+pow(22-16.67,2))	sqrt(pow(3.5-3.5,2)+pow(22-18.5,2))	sqrt(pow(3.5-3,2)+pow(22-14,2))
free	sqrt(pow(3.5-3.33,2)+pow(22-14,2))	0	sqrt(pow(3.33-3.67,2)+pow(14-16.67,2))	sqrt(pow(3.33-3.5,2)+pow(14-18.5,2))	sqrt(pow(3.33-3,2)+pow(14-14,2))
file	sqrt(pow(3.5-3.67,2)+pow(22-16.67,2))	sqrt(pow(3.33-3.67,2)+pow(14-16.67,2))	0	sqrt(pow(3.67-3.5,2)+pow(16.67-18.5,2))	sqrt(pow(3.67-3,2)+pow(16.67-14,2))
new	sqrt(pow(3.5-3.5,2)+pow(22-18.5,2))	sqrt(pow(3.33-3.5,2)+pow(14-18.5,2))	sqrt(pow(3.67-3.5,2)+pow(16.67-18.5,2))	0	sqrt(pow(3.5-3,2)+pow(18.5-14,2))

party	sqrt(pow(3.5-3,2)+pow(22-14,2))	sqrt(pow(3.33-3,2)+pow(14-14,2))	sqrt(pow(3.67-3,2)+pow(16.67-14,2))	sqrt(pow(3.5-3,2)+pow(18.5-14,2))	0

Таблиця 4. Таблиця відстаней слів, ітерація №1 (продовження)

Слово	tomorrow	free	file	new	party
tomorrow	0	8.00	5.33	3.5	8.02
free	8.00	0	2.69	4.50	**0.33**
file	5.33	2.69	0	1.84	2.75
new	3.5	4.50	1.84	0	4.53
party	8.02	**0.33**	2.75	4.53	0

Знаходимо найменшу відстань. На першій ітерації слова free та party об’єднуємо у групу (див. таблицю 5).

Таблиця 5. Таблиця груп слів, ітерація №1

Слово	Група
tomorrow	0
free	0 → 1
file	0
new	0
party	0 → 1

Для об’єднаних слів розраховуємо центроїду (див. таблицю 6).

Таблиця 6. Перелік слів та їх атрибутів, ітерація №1

Слово	атрибут_X	атрибут_Y
tomorrow	3.5	22
free + party	(3.33 + 3)/2 = 3.17	(14 + 14)/2 = 14
file	3.67	16.67
new	3.5	18.5

Наступник кроком після об’єднання слів визначають два найближчі екземпляри, для цього будують таблицю відстаней (див. таблицю 7).

Таблиця 7. Таблиця відстаней слів, ітерація №2

Слово	tomorrow	free + party	file	new
tomorrow	0	sqrt(pow(3.5-3.17,2)+pow(22-14,2))	sqrt(pow(3.5-3.67,2)+pow(22-16.67,2))	sqrt(pow(3.5-3.5,2)+pow(22-18.5,2))
free + party	sqrt(pow(3.5-3.17,2)+pow(22-14,2))	0	sqrt(pow(3.17-3.67,2)+pow(14-16.67,2))	sqrt(pow(3.17-3.5,2)+pow(14-18.5,2))
file	sqrt(pow(3.5-3.67,2)+pow(22-16.67,2))	sqrt(pow(3.17-3.67,2)+pow(14-16.67,2))	0	sqrt(pow(3.67-3.5,2)+pow(16.67-18.5,2))
new	sqrt(pow(3.5-3.5,2)+pow(22-18.5,2))	sqrt(pow(3.17-3.5,2)+pow(14-18.5,2))	sqrt(pow(3.67-3.5,2)+pow(16.67-18.5,2))	0

Таблиця 7. Таблиця відстаней слів, ітерація №2 (продовження)

Слово	tomorrow	free + party	file	new
tomorrow	0	8.01	5.33	3.5
free + party	8.01	0	2.72	4.51
file	5.33	2.72	0	**1.84**
new	3.5	4.51	**1.84**	0

Знаходимо найменшу відстань. На першій ітерації слова free та party об’єднуємо у групу (див. таблицю 8).

Таблиця 8. Таблиця груп слів, ітерація №2

Слово	Група
tomorrow	0
free	0 → 1
file	0 → 2
new	0 → 2
party	0 → 1

Для об’єднаних слів розраховуємо центроїду (див. таблицю 9).

Таблиця 9. Перелік слів та їх атрибутів, ітерація №2

Слово	атрибут_X	атрибут_Y
tomorrow	3.5	22
free + party	3.17	14
file + new	(3.67 + 3.5)/2 = 3.59	(16.67 + 18.5)/2 = 17.59

Наступник кроком після об’єднання слів визначають два найближчі екземпляри, для цього будують таблицю відстаней (див. таблицю 10).

Таблиця 10. Таблиця відстаней слів, ітерація №3

Слово	tomorrow	free + party	file + new
tomorrow	0	sqrt(pow(3.5-3.17,2)+pow(22-14,2))	sqrt(pow(3.5-3.59,2)+pow(22-17.59,2))
free + party	sqrt(pow(3.5-3.17,2)+pow(22-14,2))	0	sqrt(pow(3.17-3.59,2)+pow(14-17.59,2))
file + new	sqrt(pow(3.5-3.59,2)+pow(22-17.59,2))	sqrt(pow(3.17-3.59,2)+pow(14-17.59,2))	0

Таблиця 10. Таблиця відстаней слів, ітерація №3 (продовження)

Слово	tomorrow	free + party	file + new
tomorrow	0	8.01	4.41
free + party	8.01	0	**3.61**
file + new	4.41	**3.61**	0

Знаходимо найменшу відстань. На першій ітерації слова free та party об’єднуємо у

групу (див. таблицю 11).

Таблиця 11. Таблиця груп слів, ітерація №3

Слово	Група
tomorrow	0
free	0 → 1 → 3
file	0 → 2 → 3
new	0 → 2 → 3
party	0 → 1 → 3

Для об’єднаних слів розраховуємо центроїду (див. таблицю 12).

Таблиця 12. Перелік слів та їх атрибутів, ітерація №3

Слово	атрибут_X	атрибут_Y
tomorrow	3.5	22
free + party + file + new	(3.17 + 3.59)/2 = 3.38	(14 + 17.59)/2 = 15.8

На рис. 2 зображено результат роботи методу ієрархічної кластеризації у вигляді дерева.

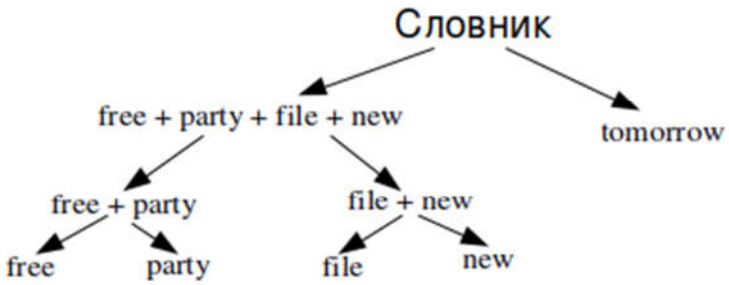

Рисунок 2. Результат ієрархічної кластеризації

Результат кластеризації можна розділити на необхідну кількість кластерів починаючи з кореня. Очевидно є два кластери, лівий позначимо як “SPAM”, а правий як “NON-SPAM”.

Як у першій лабораторній роботі спробуємо класифікувати тестове повідомлення “Free file tommorow”. Оскільки в реченні два слова відносяться до класу (кластеру) “SPAM” можна зробити висновок, що електронних лист також відноситься до класу “SPAM”.

Варіанти завдань

№	Тематики	№	Тематики	№	Тематики
1	Транспорт, здоров’я, шоу-бізнес	11	Музика, кіно, театр	21	Автоспорт, кіно, шоу-бізнес
2	Нерухомість, фінанси, енергетика	12	Енергетика, нерухомість, кіно	22	Інтернет, космос, наука
3	Автоспорт, хокей, баскетбол	13	Транспорт, здоров’я, шоу-бізнес	23	Нерухомість, фінанси, енергетика
4	Музика, кіно, театр	14	Нерухомість, фінанси, енергетика	24	Музика, наука, баскетбол
5	Інтернет, космос, наука	15	Хокей, фінанси, кіно	25	Хокей, фінанси, кіно
6	Автоспорт, кіно, шоу-бізнес	16	Автоспорт, хокей, баскетбол	26	Автоспорт, кіно, шоу-бізнес
7	Хокей, фінанси, кіно	17	Інтернет, космос, наука	27	Музика, наука, баскетбол
8	Енергетика, нерухомість, кіно	18	Музика, наука, баскетбол	28	Транспорт, здоров’я, шоу-бізнес

9	Музика, наука, баскетбол	19	Енергетика, нерухомість, кіно	29	Здоров’я, театр, космос
10	Здоров’я, театр, космос	20	Транспорт, здоров’я, шоу-бізнес	30	Транспорт, енергетика, баскетбол

Звіт по роботі

1. Розробити програму;
2. Вибрати за варіантом набір даних для побудови моделі кластеризації;
3. Вибрати набір даних для тестування;
4. Побудувати графік залежності точності моделі кластеризації від кількості екземплярів (слів, речень);
5. Побудувати графік залежності точності моделі кластеризації від кількості атрибутів екземплярів (запропонувати відмінні від наведених у прикладі);
6. Навести код програми;
7. Сформулювати висновок.