МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ
Завідувач кафедри інтелектуальних інформаційних систем, д.т.н., проф.,
_____Ю.П.Кондратенко
«____»_____2022 року

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

АНАЛІЗ МЕТОДІВ ЛУВЕНА ТА К-СЕРЕДНІХ ПРИ ВИРІШЕННІ ЗАДАЧ КЛАСТЕРИЗАЦИЇ ДЛЯ ВИЗНАЧЕННЯ СПІЛЬНОТ НА МНОЖИНІ ЕЛЕМЕНТІВ

Спеціальність 124 «Системний аналіз»

124 – МКР – 607.21610803

Студент _____ Голуб Р.Р.

«17» лютого 2022 р.

Консультант_____Дворецький М.Л.
канд. тех. наук, ст. викладач кафедри ІПЗ

«17» лютого 2022 р.

Миколаїв – 2022

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет ім. Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

Освітньо-кваліфікаційний рівень **магістр**
Галузь знань **12 «Інформаційні технології»**
(шифр і назва)
Спеціальність **124 «Системний аналіз»**
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних інформаційних систем, д-р техн. наук, проф.

_____ Ю. П. Кондратенко

«____» _____20__ р.

З А В Д А Н Н Я

на магістерську кваліфікаційну роботу
Голубу Ростиславу Руслановичу

1. Тема магістерської кваліфікаційної роботи «Аналіз методів лувена та к-середніх при вирішенні задач кластеризації для визначення спільнот на множині елементів».

Керівник роботи Дворецькій Михайло Леодінович, канд. тех. наук, ст. викладач кафедри ІПЗ

Затв. наказом Ректора ЧНУ ім. Петра Могили від «__» ____ 20__ р. № ____

2. Строк подання студентом роботи 20__ р.

3. Вхідні (початкові) дані до роботи: набори даних мереж з різною характеристикою.

Очікуваний результат роботи: порівняльна характеристика методів Лувена та К-середніх на різних наборів данних. Аналіз експериментів та отриманих результатів

4. Перелік питань, що підлягають розробці (зміст пояснювальної записки):

здійснення аналізу мережевих сервісів у сфері продажу автомобілів та дослідження теоретичних засад створення рекомендаційних систем автосалону на основі уподобань користувачів;

обґрунтуваня вибору наборів даних та інструментальних засобів для експериментів;

здійснення експериментів та аналіз отриманих результатів

5. Перелік графічного матеріалу: презентація, рисунки, таблиці.

6. Завдання до спеціальної частини: Охорона праці

7. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис
Спеціальна частина з охорони праці	д.б.н.,професор Л. І. Григор'єва	
Методична частина	канд. тех. наук, ст. викладач кафедри ІПЗ Дворецький М.Л	

Керівник роботи канд. тех. наук, ст. викладач кафедри ІПЗ Дворецький М.Л._____

(наук. ступінь, вчене звання, прізвище та ініціали)

_____

(підпис)

Завдання прийнято до виконання ___Голуб Р. Р._____

(прізвище та ініціали)

_____

(підпис)

Дата видачі завдання «_____» _____ 20__ р.

КАЛЕНДАРНИЙ ПЛАН

Виконання магістерської кваліфікаційної роботи

Тема: Аналіз методів лувена та к-середніх при вирішенні задач кластеризації для визначення спільнот на множині елементів

№	Найменування роботи	Початок	Закінчення	Примітки
1	Визначення керівника і теми МКР. Подання заяви на затвердження теми МКР	01.09.2021	10.10.2021	Виконано
2	Отримання завдання на виконання МКР	19.10.2021	22.10.2021	Виконано
3	Складання календарного плану на період виконання МКР	23.10.2021	26.10.2021	Виконано
4	Огляд літератури за темою дослідження	27.10.2021	10.11.2021	Виконано
5	Проходження переддипломної практики, збір та аналіз матеріалів до МКР	29.11.2021	18.12.2021	Виконано
6	Аналіз предметної області та розробка технічного завдання.	23.12.2021	02.01.2022	Виконано
7	Опис фахової частини МКР	03.01.2022	25.01.2022	Виконано
8	Розробка спеціальної частини з охорони праці та методичної частини	26.01.2022	30.01.2022	Виконано
9	Попередній захист МКР на засіданні комісії кафедри	31.01.2022	31.01.2022	Виконано
10	Корегування роботи за результатами попереднього захисту	01.02.2022	03.02.2022	Виконано
11	Остаточне оформлення пояснювальної записки та слайдів доповіді для захисту	04.02.2022	08.02.2022	Виконано
12	Подання МКР рецензенту	11.02.2022	12.02.2022	Виконано
13	Рецензування МКР	13.02.2022	14.02.2022	Виконано
14	Подання МКР, її електронної копії та інших документів (відгуку, рецензії) до захисту	16.02.2022	17.02.2022	Виконано
15	Захист МКР перед екзаменаційною комісією (ЕК)	24.02.2022	24.02.2022	Виконано

Розробив студент _____Голуб Р.Р._____ _____

(прізвище та ініціали) (*підпис*)

Керівник роботи канд. тех. наук.. ст. викл. кафедри ІПЗ__Дворецький М.Л._____

(*наук. ступінь, вчене звання, прізвище та ініціали)* *(підпис)*

«24» жовтня 2021 р.

Анотація

Виявлення спільноти на графах складних реальних мережевих систем є значною областю досліджень науки про дані. Спільнота або кластер - це спільнота, яка має багато ребер, що з'єднують вершини, включені до кластера, тоді як менше ребер, що з'єднуються з вершинами, не включеними. Критерії включення в спільноту засновані на даних вершин і ребер. Дані системи допомагають формувати спільноти системи. Ці спільноти допомагають аналітикам даних отримати високий рівень зору на виборців системи. Така точка зору має вирішальне значення для виконання поведінкового аналізу, прийняття управлінських рішень, стратегії маркетингових планів, рекомендацій тощо. Багато алгоритмів виявлення спільноти були запропоновані в дослідницькому товаристві. З них алгоритм Лувена та алгоритм К-середніх є двома популярними варіантами алгоритмів. Ця робота виконує порівняльний аналіз двох алгоритмів виявлення спільноти, використовуючи різні реальні набори даних різної складності та оцінюючи їх ефективність проти них. Порівняння висвітлять сильні сторони та обмеження кожного алгоритму та запропонують ідеальні сценарії для їх застосування.

Об’єкт дослідження. Об’єктом дослідження є процес вирішення задачі кластеризації для визначення спільнот на множині елементів

Предмет дослідження.

Предметом дослідження є аналіз ефективності методів Лувена та k-середніх для визначення спільнот на множинах елементів різного розміру

Ключові слова: кластеризація, алгоритм Лувена, алгоритм К-засобів, набори даних, системний аналіз.

Abstract

A significant area of data research is the detection of the community on charts of complex real network systems. A community or cluster has many edges connecting the tops included in the cluster, while fewer edges connecting to the tops, are not included. Criteria for inclusion in the community are based on data vertices and edges. These systems help to form systems' communities. These communities help data analysts get a high-level view of system voters. This point of view is crucial for behavioral analysis, making managerial decisions, marketing plans strategy, recommendations, etc. Numerous community detection algorithms were offered in the Research Component. The Louvain algorithm and K-means algorithm are the most popular among voters. This work performs a comparative analysis of two community detection algorithms using real data sets of different complexities and evaluating their performance against them. This comparison highlights the pros and cons of each algorithm and will offer ideal scenarios for their application. Thus, it can be easily found out lots of advantages on both algorithms from varying complexity and evaluating their effectiveness.

Object of research. The object of research is the process of solving the problem of clustering to identify communities on a set of elements

Subject of research.The subject of the study is the analysis of the effectiveness of Leuven's and k-means methods for determining communities on sets of elements of different sizes.

Keywords: Louvain algorithm, K-means algorithm, data sets, community detection, analysis.

Пояснювальна записка

до магістерської кваліфікаційної роботи

на тему:

«АНАЛІЗ МЕТОДІВ ЛУВЕНА ТА К-СЕРЕДНІХ ПРИ ВИРІШЕННІ ЗАДАЧ КЛАСТЕРИЗАЦИЇ ДЛЯ ВИЗНАЧЕННЯ СПІЛЬНОТ НА МНОЖИНІ ЕЛЕМЕНТІВ»

Спеціальність 124 «Системний аналіз»

124 – МКР – 607. 21610803

Студент _____ Голуб Р.Р.

«__» _____ 20___ р.

Консультант _____ Дворецький М.Л.

канд.тех.наук, ст. викладач кафедри ІПЗ

«__» _____ 20___ р.

м. Миколаїв – 2022

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

NMI – Нормалізована взаємна інформація

AMI – Відрегульована взаємна інформація

ARI – скоригований індекс Ранда

FMI – Індекс Фаулкса-Мэллоуса

F1 – F-міра

SSE – Сума квадратів помилок

-

ВСТУП

Актуальність теми дослідження. Важливі особливості системи можуть бути виявлені як набори даних, які можуть бути представлені у вигляді графіків. Наприклад, соціальна мережа, така як Facebook, надає спільну платформу для людей, що дозволяє їм з'єднуватися та обмінюватися даними. Представлення соціальної мережі через граф буде складатися з вузлів і ребер. Вузол буде представляти людину, а ребро буде вказувати на взаємодію людини (наприклад, дружба, наприклад, поділитися).

Спільнота або кластер формується вершинами, які щільно пов'язані та навряд чи пов'язані з вершинами інших громад. Спільноти або кластери допомагають розділити велику мережу на меншу колекцію вузлів з певних точок інтересу. Одна і та ж мережа може бути використана для виявлення спільнот різних типів. Наприклад, спільнота користувачів, які люблять футбол або підтримують конкретного політичного лідера.

Представлення складних реальних систем у вигляді вершин і країв може служити важливим інструментом для розуміння та аналізу загальної системи після виявлення спільнот. Всі вузли зі схожими властивостями та поведінкою утворюють спільноту. Наприклад, загальна тенденція людей в соціальній мережі, наприклад, подобається подібним пунктам / постам / сторінкам, обмін думками на певні теми, переваги та т.д., змушує їх потрапляти у віртуальні спільноти або кластери.

Виявлення цих спільнот або кластерів може бути значним для багатьох маркетингових та рекомендаційних додатків [1. Кім і Ан, 2008]. Наприклад, у соціальній мережі, повідомляючи футбольних фанатів про нову футбольну подію, пропонуючи предмети, засновані на історії перегляду людини, пропонуючи друзів, знаходячи впливових людей у групі тощо. Подібно до соціальних мереж, у виявленні спільноти знаходить своє застосування в багатьох інших областях, наприклад, у вивченні науково-дослідної співпраці [2. Jiyanthi and Priya, 2018], пошуку білкових взаємодій у біологічних мережах [3. Ванг та ін., 2010], вивченні карт повітряного та наземного трафіку в режимі реального

часу [4. Патанак та ін., 2018] тощо.

Мета і завдання дослідження. Метою дослідження є аналіз та порівняння двох методів для вирішення задачі кластеризації для визначення спільнот на множині елементів, визначення переваг та недоліків кожного.

Відповідно до мети були поставлені такі завдання:

- Провести аналіз алгоритмів кластеризації
- Провести аналіз існуючих робіт.
- Розробити систему для порівняльного аналізу.
- Виконати тестування двох алгоритмів щодо трьох наборів даних різного розміру.
- Проаналізувати отримані результати порівнюючи два методи та зробити висновок, який з методів є кращим для визначеної задачі.

Об’єкт дослідження. Об’єктом дослідження є процес вирішення задачі кластеризації для визначення спільнот на множині елементів

Предмет дослідження.

Предметом дослідження є аналіз ефективності методів Лувена та k-середніх для визначення спільнот на множинах елементів різного розміру.

1.ОГЛЯД ТА АНАЛІЗ НАЯВНИХ АНАЛОГІВ. ПОСТАНОВКА ЗАДАЧІ

1.1 Визначення проблеми

Завдання виявлення спільноти може бути складним процесом. Наприклад, у соціальних мережах це передбачає оцінку людей, їх взаємодію та прогнозування відсутньої інформації. З огляду на складність завдання виявлення спільноти, у літературі запропоновано великий обсяг даних та алгоритмів. Ці пропозиції є загальним внеском у дослідження виявлення спільноти або є специфічними для деяких вибраних областей застосування.

Алгоритми Лувена та K-середніх (K-means) є двома популярними алгоритмами в дослідницьких колах, які використовуються для пошуку спільнот або кластерів. Причиною вибору саме цих алгоритмів стало те, що алгоритм Лувена зазвичай є першим вибором для завдань виявлення спільноти у графічних мережах [Blondel et al. 2008]. Хоча K- середніх є дуже популярним вибором при роботі з програмами, які вимагають просторової кластеризації, наприклад зображень. У мережевому аналізі немає просторових зв'язків між вузлами. Тому міцність алгоритму спрямована на тестування на мережевих графіках. Алгоритм K-середніх був успішно застосований у кластерних мережевих системах [Vilcek, 2014]. Порівняння алгоритмів Лувена та K- середніх на тих самих наборах даних допоможе порівняти продуктивність та ефективність двох.

Алгоритм Лувена - це агломеративний ієрархічний алгоритм кластеризації, тоді як K-середніх - це спектральний алгоритм кластеризації. Обидва ці алгоритми допомагають розділити велику мережу на менші спільноти. Метою цієї статті є використання декількох реальних наборів даних для порівняння та аналізу продуктивності алгоритму Лувена та алгоритму K-середніх. Для проведення аналізу слід використовувати набір популярних показників ефективності. Набори даних збираються з реальних наборів даних із відомими спільнотами правди, а набори даних є загальнодоступними для дослідницьких цілей. Робота повинна довести, чи:

K-середіх s та Лувен можуть вловити суть мережі чи ні.

Основних заходів, тобто модульності для Лувена та SSE для K- середніх, достатньо для кластеризації мережі чи ні.

Близькість до основної істини достатня для того, щоб визначити ефективність алгоритму чи ні.

Покращення попередньої фази алгоритму (тобто виділення центроїда) покращує результати чи ні.

1.2 Огляд системи

Система, розроблена для проведення порівняльного аналізу, складається з модуля ідентифікації спільноти та модуля порівняння. Модуль ідентифікації складається з двох аналізованих алгоритмів, тобто алгоритму Лувена та алгоритму K-середіх. Набір мережевих даних служив вхідним сигналом для модуля ідентифікації. Карта, що вказує номер кластера кожного вузла, була результатом роботи модуля ідентифікації. Модуль порівняння порівнює карту спільноти виводу з “ground truth” спільнотами, таким чином допомагаючи оцінити загальну продуктивність алгоритму на основі певного набору даних. Загальна системна основа представлена на рис 1.1.

Рис. 1.1 Системна основа

1.3 Огляд та аналіз наявних аналогів та публікацій

З огляду на проблему ідентифікації спільноти або кластеризації, було опубліковано багато алгоритмів. Невеликі варіації або доповнення до спочатку запропонованих алгоритмів були запропоновані для поліпшення загальної продуктивності проти деяких конкретних наборів даних або для загального просування алгоритму

Артур і Васильвіцький [1] пропонують модифікацію початкового кроку K-середіх вибору центроїдів випадковим чином, називаючи його K-середіх ++. Спочатку алгоритм K-середіх вибирав будь-який вузол, який був обраний за допомогою випадкового вибору. Це призведе до того, що вузли утворять ті ж кластери в тих випадках, коли центроїди будуть відібрані близько один до одного. Запропонована модифікація гарантувала, що початкові випадкові центроїди добре відокремлені один від одного. Модифікація не тільки вдосконалила алгоритм, але і збільшила його швидкість. Перед застосуванням алгоритму K-середіх мережеві дані зіставляються з менш розмірним простором, який має всі корисні функції вихідних даних. Загальний метод, який використовується для виконання цього відображення - через нейронні мережі. Vilcek (2014) запропонував алгоритм під на}звою Deep K-середіх, в якому K-середіх служить багатошаровим автокодером, який рекурсивно розкладає високорозмірні дані на низькорозмірні дані. Запропонований алгоритм перевершив традиційний спектральний алгоритм кластеризації.

Ван і Коопман (2017) використовують алгоритми K-середніх і Лувена для кластеризації статей з набору даних великої мережі Astro, без визначених ground truth спільнот. Кластеризація заснована на пошуку семантичної подібності між статтями. Семантична інформація є метаданими набору даних. На основі результатів кластеризації різних дослідників для одного і того ж набору даних, наближення результатів дослідження розглядалося як основні спільноти істини для тестування двох алгоритмів. Застосування алгоритму Лувена було простим, оскільки алгоритму не потрібні були будь-які вхідні параметри для запуску і розрахунку кластерів з локальною максимальною модульністю. Однак для

використання K-середніх значення k визначалося прагматично. Автори використовували максимальну оцінку силуету, що відповідає k, яка становила 30. Знайдене число було близьке до 31, кількість суспільств, знайдених алгоритмом Лувена. Для початкового розміщення центроїдів алгоритм був запущений 10 разів для значення k від 10 до 60 і був обраний центроїдний вузол, який дав найнижчу суму квадратів.

Порівняльне дослідження було проведено Jianjun et al. (2014), в якому вони пропонують активний підхід до навчання "Must-Link Cannot-Link" для неорієнтованих графів і використовують його з напівконтрольованим алгоритмом виявлення спільноти. Комбінація покращила загальні результати кластеризації алгоритму. В цілому шість алгоритмів були зіставлені на основі їх показників ефективності при поділі чотирьох наборів даних відомих істин. Три метрики були обрані для вимірювання продуктивності алгоритму: модульність, точність і нормалізування взаємної інформації (HBI/NMI).

З метою оцінки алгоритмів виявлення спільноти для реальних малих і великих мережевих наборів даних, доступно багато опублікованих досліджень. Деякі пропонують різні показники показника ефективності, в той час, як інші обговорюють неправильні уявлення про певні фактори, які ігноруються, але можуть вплинути на продуктивність кластеризації. Робота висвітлює питання в самому процесі оцінки ефективності.

Лі та Каннінгем (2014) представляють аргумент, що алгоритми, які добре працюють на менших наборах даних з відомими спільнотам, можуть не працювати однаково добре, витягуючи спільноти у великих наборах даних соціальних мереж, де структура мереж невідома. Для кластеризації цих великих наборів даних алгоритми виявлення спільноти використовують деякі метадані. Ці метадані самі по собі можуть не ідеально зображувати структуру мережі.

Піл та ін. (2017) поділяють ту ж точку зору, що Лі та Каннінгем (2014). Алгоритми не працюють однаково в різних спільнотах різного розміру і складності. Автори запропонували теорему «No Free Lunch», яка доводить, що жоден алгоритм виявлення спільноти не є достатньо гнучким для обробки всіх

типів завдань виявлення спільноти. Не існує єдиної універсальної моделі для структури спільноти; жодні метадані не описують всі аспекти спільноти, і жоден алгоритм не підходить всім. Існують загальні алгоритми, які добре працюють в цілому, а потім є ті, які адаптовані для конкретних завдань. Згідно з дослідженням, навіть звичайна практика вибору метаданих з великих мереж (наприклад, стать, релігія, вік, етнічна приналежність і т.д.) в якості обґрунтованого показника істини не може бути ефективним інструментом для всіх типів громад. Вони вивчають взаємозв'язок між структурою спільноти та метаданими в декількох рамках виявлення, щоб оцінити, наскільки тісний зв'язок між ними до їх ролі в складних реальних мережевих системах. Іноді, коли алгоритм погано працює при витягуванні спільнот, помилка може бути пов'язана з вибраними метаданими. Метадані можуть або не повністю представляти всі групи, які можуть належати до мережі, або вони можуть не надто добре ставитися до загальної структури спільноти.

Jebabli et al. (2015) стверджує, що коли оцінювачі якості розглядаються як критерій оптимізації, структура спільноти може бути проігнорована. Це тому, що заходи якості не залежать від базової топології мережі. Дві несхожі спільноти, що мають різні зв'язки між вузлами, можуть давати подібні НВІ. Як передбачається, щоб називатися ефективним алгоритмом, він повинен забезпечити кластери, які згодні з топологією спільноти більше в порівнянні зі спільнотою землі істини. Попри зниження результатів показників оцінки, алгоритми повинні зосередитися на кодуванні топології мережевої спільноти. Щоб довести важливість підтримки топологічних структур під час кластеризації, робота використовує наземну структуру спільноти dataset даних Amazon і порівнює її з передбачуваною структурою спільноти, створеною популярним алгоритмом виявлення спільноти. Потім він вивчає топологічні властивості графа на макроскопічному рівні (через середній коефіцієнт кластеризації, діаметр, щільність і кореляцію ступеня, середній найкоротший шлях), мікроскопічний рівень (через відстань хмелю, розподіл ступеня вузла і пов'язаний з ним середній коефіцієнт кластеризації) і, нарешті, мезоскопічний

рівень (через розподіл розмірів спільноти).

Ця робота представляє методологію порівняльного аналізу з використанням модульності, NMI та показників точності, подібних до Jianjun et al. (2014), заявлених вище. Але з огляду на спостереження, представлені Lee and Cunningham (2003) і Peel et al. (2017), два алгоритми повинні оцінюватися не тільки на невеликих наборах даних з доступною наземною істиною, але і для великих мережевих спільнот, щоб отримати загальне уявлення про зміну результатів. Переконавшись в аргументах, представлених Jebabli et al. (2015), в рамках майбутньої роботи, результати, отримані двома алгоритмами, повинні бути оцінені відповідно до топологічних показників, щоб побачити, наскільки добре кожен алгоритм виконується проти підстав для топології мережі істинності. Крім того, бачачи вплив використання модульних ядер на основі модульності (Sommer et al., 2017) поряд з алгоритмом K-середіх на продуктивність кластеризації, такий підхід повинен бути вивчений в майбутньому.

Висновки до розділу 1

У цьому розділі було визначено проблему та напрямок роботи, оглянуто систему, на базі якої буде виконуватися вся подальша робота з алгоритмами та детельно оглянуті та проаналізовані аналогічні роботи та публікаціїї.

2 ПОХОДЖЕННЯ КОНЦЕПЦІЙ

2.1 Теорія графів

2.1.1 Граф

Граф формується набором вершин, з'єднаних по ребрах. Загальна нотація, яка використовується для позначення графа, - це G = (V, E), де V - список вершин / вузлів, а E - список ребер. На рис. 2.1 зображений простий граф, що складається з 20 вузлів. Ребер між вузлами 52.

Рис. 2.1 Компоненти графа

Графи можуть бути спрямованими або неорієнтованими, зваженими або незваженими.

а) орієнтований, зважений (б) неорієнтований, незважений

Рис. 1.2 Типи графів

Аналіз, представлений в цій роботі, стосується тільки неорієнтованих і незважених графів.

2.1.2 Суспільство

Група вузлів всередині графа, що демонструють подібні характеристики, може бути згрупована для формування спільноти. Для графа G, спільнота буде C = (Vc, Ec), так що C буде підмножиною G.

Рис. 2.2 Спільноти в межах графа

Спільнота також може бути визначена як згуртована група, де взаємодія членів групи між собою є більш інтенсивною у порівнянні з комунікаціями членів з суб'єктами, присутніми за межами групи [Jebabli et al., 2015].

Спільноти можуть перекриватися і не перекриватися. Спільноти, що перекриваються, - це ті спільноти, де вузли призначені більш ніж одному

кластеру / спільноті.

Рис. 2.3 Перекриття та спільноти, що не перекриваються

Аналіз у цій роботі в основному був зосереджений на оцінці ефективності алгоритму для реальних спільнот, що не перекриваються.

2.1.3 Матриця суміжності графа

Оскільки K-середіх працює над кластеризуванням даних 2D-простору, шукали спосіб відображення мережевих вузлів і ребер у формат 2D-простору. Для цього була створена нова матриця суміжності. Матриця відноситься до двовимірного (2-D) вектора, який представляє мережеву структуру в просторовому представленні, тобто представляє вузли та ребра у вигляді двовимірної матриці. Кластеризація K-середіх застосовується на цьому 2D-векторі. По суті, це матриця суміжності [Вайстейн, 2018], але на відміну від матриці суміжності, вона має матрицю по діагоналі.

Для заданого орієнтованого або неорієнтованого графа, якщо v є вершинами графа, матриця ребер є матрицею (0-1) розміром v x v, тобто v рядками та v стовпцями. Ребро графа буде представлене як 1 між двома вершинами в матриці Подібно до матриці суміжності, для неорієнтованих графів матриця ребер симетрична. Але на відміну від матриці суміжності, матриця краю має 1 в діагоналі. Нижче наведено кілька прикладів матриці ребер.

Рис. 2.4 Зразок матриці суміжності

На рисунку 6 зображено три різних графа, що складаються з 4 вершин. У всіх трьох графах матриця суміжності буде мати 4 рядки й 4 стовпці, що представляють інформацію про кожну вершину з іншими вершинами на графу.

Алгоритм K-середіх обробляє всі дані в кожній ітерації. Для дуже великих мереж, наприклад, мережа з більш ніж 10 000 вузлів, використовуючи матрицю

розміру nxn, де n - кількість мережевих вузлів, може зробити розмір даних надзвичайно великим і рідкісним. Виконання обчислень над такою великою матрицею стає надзвичайно дорогим з точки зору можливостей обробки, часу і пам'яті. Таким чином, для дуже великих мереж використання матриці розміром nxn під час кластеризації K-середіх є непрактичним підходом. Таким чином, як правило, при використанні K-середіх для кластеризації даних дуже великих мереж, зменшена розмірність використовується для представлення даних, наприклад, лапласівських векторів власного покоління, перетворення Фур'є, основного аналізу компонентів тощо. Кластеризація над цими зменшеними представленнями даних, забезпечує швидші результати та вимагає менше ресурсів, забезпечуючи еквівалентні результати. У цій роботі мережі, розглянуті для кластеризації, були відносно меншими розмірами; Найбільша з 1005 вузлів. Таким чином, не було виконано жодного попереднього процесу зменшення розмірності даних. Матриця суміжності представляла фактичні ребра мережі nxn.

2.2 Огляд методів Лувена та K-середніх

Основний процес навчання алгоритму кластеризації може бути або контрольованим, або неконтрольованим [Tzanakou, 2017]. І K-середіх, і Лувен є неконтрольованими алгоритмами навчання.

По суті, кластеризація мереж є неконтрольованою проблемою навчання.І K-середіх, і Лувен вбудовують логіку неконтрольованого навчання і можуть ефективно кластеризувати дані. Крім того, що найбільш попитні алгоритми кластеризації в літературі, обидва мають різний дизайн і риси. K-середіх враховує всі дані, тоді як Лувен зменшує розмір даних з кожною ітерацією. K-середіх має перевагу прогнозування відсутніх функцій під час процесу кластеризації. K-середіх вимагає ручного введення щодо вибору k, тоді як Лувен є повністю автоматизованим методом.

Ще однією мотивацією для вибору K-середіх і Лувена цієї роботи було те, що в суміжній роботі [Ван і Купмана (2017)] ці два алгоритми вважалися

кластерним семантичним представленням даних дуже великого набору даних Astro (з більш ніж 18 000 вузлами). Підставою для істини в цьому порівнянні було експериментальне наближення. Цю роботу можна розглядати як продовження порівняння двох алгоритмів з використанням менших наборів даних різних розмірів і закономірностей з відомими наземними істинами.

2.2.1Алгоритм Лувена

Лувен - це неконтрольований алгоритм (не вимагає введення кількості спільнот або їх розмірів перед виконанням), поділений на дві фази: оптимізація модульності та агрегація спільнот. Після завершення першого кроку слідує другий. Обидва будуть виконуватися доти, доки в мережі не припиняться зміни та не буде досягнуто максимальної модульності..[Блондель В.Д]

Один пробіг двох фаз утворює ітерацію. В кінці кожної ітерації знаходиться кластеризація (розділ) графа і новий рівень в дендрограмі (див. рис. 2.6). Корінь дендрограми являє собою остаточне кластеризування і має найвищу модульність.

Рис. 2.6 Дендрограма алгоритма Лувена

Фаза перша: Це етап вдосконалення, коли кожному вузлу присвоюється окрема спільнота. Вузли потім проходять випадковим чином. Для кожного пройденого вузла та алгоритм обчислює зміну модульності, коли вузол *i* переміщується з його поточної призначеної спільноти до будь-якої з сусідніх спільнот. Якщо потенційний хід дає більш високу модульність зміни, то вузол *i* присвоюється до цієї сусідньої спільноти. Якщо ніякі потенційні ходи не дають більш високої зміни модульності, то вузол залишається в його поточному суспільстві.

Нехай с буде сусіднім суспільством вузла і, в яке він зливається. Зміна модульності для вузла обчислюється за допомогою наступного рівняння;

Де,

$\sum_{in}$ це сума ваг ребер всередині спільноти с

$\sum_{tot}$ це сума ваг ребер, які прибувають до вузлів, що містяться в спільноті c

k_i це сума ваг ребер, які є інцидентом з вузлом i

$k_{i,in}$ сума ваг ребер від вузла i до вузлів в c

m сума ваг всіх ребер в мережі

Процес повторюється до тих пір, поки в ході розгортання всіх вузлів не буде більше переміщень, тобто не буде збільшено модульності для будь-яких потенційних ходів для кожного з вузлів для відповідних сусідніх вузлів. Коли модульність припиняється для покращення, це означає, що алгоритм знайшов локальну максимальну модульність.

Друга фаза: Це фаза де кожна спільнота, знайдене на першій фазі, розглядається як новий вузол для подальшої обробки. Ребра, які існують у раніше виявлених спільнотах, замінюються самоциклами, які з'єднані з новими вузлами. Вага самопетель визначається сумою ваг ребер, які були замінені. Одне ребро між новими вузлами замінює всі попередні ребра між відповідними спільнотами. Вага цього нового ребра дорівнює сумі ваг усіх ребер, які були замінені.

Дві фази повторюються, поки не буде досягнуто локальної максимальної модулярності.

Нижче показаний простий приклад, що показує вихід двох фаз на простому графі. Граф складається з 10 вузлів.Наприкінці першої фази було виявлено три спільноти. До кінця другої фази вони перетворюються на 3 вузли. Ребра всередині спільноти замінюються самоциклами, тоді як ребра між спільнотами замінюються одиночними ребрами між відповідними вузлами вузла..

(a) початковий граф (b) після першої фази (c) після другої фази

Рис. 2.7 Приклад алгоритму Лувена

Псевдокод алгоритму Лувена наведено нижче.

Рис. 2.8 Псевдокод алгоритму Лувена (Кім та ін., 2013)

Алгоритм має часову складність O (n log n), де m - кількість країв мережі. Лінійна складність робить алгоритм швидким. Значна частина обчислень проводиться на початковій фазі алгоритму. Після кількох первинних проходів кількість спільнот різко зменшується, що зменшує обчислення, проведені в пізніших проходах.

Алгоритм К-середніх (K-means)

Алгоритм Kmeans — це ітераційний алгоритм, який намагається розділити набір даних на K-попередньо визначених окремих підгрупи (кластери), що не перетинаються, де кожна точка даних належить лише одній групі. Він намагається зробити вузли всередині кластера якомога подібними, водночас зберігаючи кластери якомога різними (далекими). Він призначає вузли кластеру таким чином, щоб сума квадратів відстані між вузлами і центроїдом кластера (середнє арифметичне всіх точок даних, які належать цьому кластеру) була мінімальною. Чим менше варіацій у нас всередині кластерів, тим однорідніші (подібніші) вузли x в одному кластері. Мета завдання алгоритму - мінімізувати суму квадратичної помилки, E [Landman et al., 2018].

$$E = \sum_{i=1}^{k} \sum_{o \in C_i} d(o, cen_i)^2 = 1$$

Де d - евклідова відстань між вузлом, що розглядається, та обраним центроїдом.

Початковий етап виділення центроїдів впливає на загальні результати розподілу алгоритму K-середіх. Хороший вибір центроїдів призведе до кращих результатів кластеризації і навпаки.

2.2.2 Алгоритм K-середіх ++ (K-means++)

Для алгоритму K-середіх було запропоновано багато варіантів для покращення алгоритму (Fahim et al., 2006; Blondel et al., 2010), поліпшення початкового вибору центроїдів (Arthur and Vassilvitskii, 2007) або проведення м'якої нечіткої кластеризації для мереж, що перетинаються, де вузол може бути частиною ряду кластерів (Джеймс та ін., 1984).

Ідея полягає у тому, щоб уникнути розміщення початкових центроїдів в одному кластері, що може призвести до неоптимального рішення. Наприклад, на рис. 2.19 два центроїди виявляються дуже близько один до одного, що призводить до неоптимальної кластеризації.

(а) Неоптимальна кластеризація(б) Оптимальна кластеризація

Рис. 2.9 K-середіх чутливість до початкового вибору центроїдів

Псевдокод для алгоритму K-середіх представлений нижче.

Рис 2.10 Псевдокод для алгоритму K-значень (Ллойд, 1982)

Як може здатися з псевдокоду, алгоритм K-середіх має два важливих етапи. Спочатку алгоритм випадковим чином вибирає k об’єктів. Спочатку вважається, що ці випадково вибрані k об’єкти представляють середнє значення або центр кластера (центроїд). На першому кроці алгоритму всі об’єкти, що залишилися, обробляються таким чином, що на основі відстані між об’єктом і центром об’єкт призначається до кластера, до якого він є найближчим. На другому етапі для кожного кластера обчислюється нове середнє значення (центроїд). Ці два кроки повторюються, поки центроїд кластерів не зміниться.

На малюнку нижче показано інтерпретацію алгоритму K-середіх у вигляді блок-схеми.

Рис. 2.11 Рамки алгоритму K-значень

Найгірший випадок складності визначається тим, що $O\left(n^{\frac{k+2}{p}}\log(n)\right)$ де n - кількість вузлів вибірки, а p - кількість ознак. Середня складність є лінійною, тобто. O (k n T), де T - ітерації, а n - вибіркові вузли.

Порівняно з алгоритмом Лувена, K-середіх може однаково швидко досягати результатів. Ззазвичай запускається кілька разів. Оскільки алгоритм швидкий, повторювані пробіги не знижують його ефективність. Для невеликих

мереж алгоритм K-середіх може зайняти менше часу, ніж Лувена.

Висновки до розділу 2

Як згадувалося раніше, Лувен і K-середіх є двома з найбільш широко використовуваних алгоритмів для кластеризації даних. Проте формальної літератури, яка б порівнювала ефективність двох алгоритмів на загальних еталонних наборах даних, майже немає. Внесок цієї роботи полягає в тому, що вона порівнює два найбільш широко використовувані алгоритми кластеризації за трьома загальновживаними наборами базових даних. Порівняння допомагає визначити їхні сильні сторони та обмеження.

Вибрані набори даних по суті не перекриваються, але деякі мають окремі кластери, а інші-ні. Тож результати дослідження можуть визначити найбільш підходяще середовище кластеризації для обох алгоритмів.

Набори даних мають різний розмір. Це висвітлить поведінку алгоритмів у міру збільшення розміру даних від малого до великого. По-друге, оскільки два алгоритми реалізують різні методи кластеризації, тобто ієрархічні та спектральні, результати порівняння можна узагальнити, щоб представити їх основні методи кластеризації.

Ця робота має внести внесок в існуючу літературну базу даних порівняння алгоритмів для виявлення спільнотою. Такі порівняння можуть служити словником для вирішення ситуацій, коли певний алгоритм є найбільш підходящим для використання. Використовуючи набори даних, які розглядаються як еталони алгоритмів кластеризації в літературних колах, та оцінюючи ефективність двох алгоритмів на них, результати можна порівняти з існуючими результатами оцінки альтернативних алгоритмів кластеризації на тих самих наборах даних. Тож робота закладає основу майбутніх досліджень оцінки алгоритмів на основі опитування Такі опитування не існують з посиланням на ті самі набори даних.

Результати цієї роботи можуть бути використані для порівняння збільшення або зниження продуктивності K-середіх за допомогою інших методів

представлення даних мережі. Таким чином, ця робота також закладає основу для майбутніх досліджень щодо використання K- Means для кластеризації соціальних мереж.

3 ВИЗНАЧЕННЯ НАБОРУ ДАНИХ ДЛЯ ЕКСПЕРИМЕНТІВ ТА ВИЗНАЧЕННЯ МЕТРИК ТА ОЦІНОК МЕТОДІВ

3.1 Набори даних

Набори даних — це представлення реальної мережевої системи, представленої у вигляді графів, тобто вузлів і ребер. Ці відомі розділи або кластери є спільнотами наземної правди. Ці набори даних засновані на реальних даних системи, які можуть бути зібрані протягом тривалого періоду в рамках експерименту, спостереження або під час вирішення проблеми. Пізніше ці набори даних стануть доступними для громадськості, щоб служити тестовими прикладами для тестування продуктивності системи (SNAP, 2018).

Крім систем реального світу, також були запропоновані деякі синтетичні моделі, які можуть генерувати набори даних різної складності. Ці синтетичні моделі призначені для перевірки масштабованості алгоритмів виявлення спільнот [Fränti та Sieranoja, 2018].

Обсяг цієї роботи, однак, обмежується лише представленням аналізу реальних наборів даних.

Спільноти істини (Ground Truth Communities)

Для деяких мережевих наборів даних реальних систем є вже відомі кластери, в які мережа врешті-решт розділяється. Ці відомі розділи або кластери є ground truth спільнотами. Роботу будь-якого алгоритму кластеризації або ідентифікації спільноти можна оцінити на основі ground truth спільнот. Як правило, такі кластери та окремі вузли, які повинні знаходитися в кластері, відомі для невеликих мереж у реальному світі. Для великих мереж, як правило, лише невелика кількість спільнот відома як ground truth.

3.1.1 Набір даних мережі карате-клубу Захарія (Zachary's Karate Club)

Цей набір даних був обраний для аналізу не тільки тому, що набір даних являє собою невелику мережу, а й тому, що він широко використовується як

орієнтир для алгоритмів виявлення спільнот у дослідницькій спільноті [Грінк, 2014]. Набір даних був наданий у форматі GML компанією Newman (2013). Його основна істина також доступна.

Набір даних був зібраний Захарією (1977) протягом трьох років. Він представляє дані з університетського клубу карате, де учасники представлені вузлами. Соціальні відносини між будь-якими двома членами клубу, які виходять за межі клубного приміщення, представлені ребром між двома відповідними вузлами. Мережа складається з 34 вузлів (членів) і 78 ребер (канали зв'язку між членами). Через суперечку між інструктором та адміністратором клуб був поділений на дві групи. Зрештою, учасники, які мали більше спілкування з адміністратором, приєдналися до цієї команди. Решта приєдналися до інструктора.

Оскільки ground truth спільноти відомі цим набором даних, це гарантує простішу оцінку продуктивності та аналіз результатів виявлення спільноти двох алгоритмів.

Мітки для Ground truth

Зі знання ситуації в даній області відомо, що фактичних спільнот становить 2 для набору даних [Гірван та Ньюман, 2002]. Вузли, що входять до складу кожної спільноти, перераховані в табл. 3.1.

Таблиця 3.1

Спільноти наземної істини для набору даних клубу карате

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21
1	8, 9, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

Очікуваний графік кластеризації для основних спільнот істинності набору даних показаний на малюнку нижче. Усі мережеві діаграми були створені за

допомогою функції "plot" бібліотеки iGraph. Для представлення схеми розташування вузлів графу був обраний алгоритм Камада-Каваї. Вузли були налаштовані для відображення своїх міток. І всі вузли в спільноті отримали окремий колір, ніж інші. Решту графів данної роботи було побудовано аналогічним чином.

Рис. 3.1 Карта істини для набору даних клубу карате

Оцінка істини спільноти

Хороша модель спільноти повинна мати чіткі оцінки для таких показників стабільності спільноти, як модульність, силует та оцінка Калінського та Харабаша. Оцінки за цими трьома показниками наведені в Таблиці 3.2. Оцінки були розраховані на основі істиних спільнот правди й матриці суміжності. Два вектори були передані до відповідної функції Python для кожного з показників, що повернуло показник якості спільноти (Див. Додаток щодо відповідних функцій Python).

Таблиця 3.2

Оцінки наземних спільнот для набору даних клубу карате

Модульність	Силует	Калінський та Харабаш
0.37	0.16	7.83

Як видно з таблиці вище, основна істина має всі позитивні значення для показників добробуту громади, а це означає, що між даними немає перекриття.

3.1.2 Набір даних мережі «Дельфін» Dolphin

Набір даних був наданий у форматі GML компанією Ньюман (2013). Набір даних соціальних мереж дельфінів, зібраний [Лусьє 2003], був обраний, оскільки він удвічі більший за набір даних клубу карате Закарі. Також відомі ground truth

спільноти [Ченг та ін., 2014].

Набір даних складається з 62 вузлів, де кожен вузол являє собою дельфін (афаліна звичайна). Дельфіни мешкали у Новій Зеландії у місті Мілфорд Саунд. За період з 1994 по 2001 рік спостерігалося, що дельфіни часто спілкуються один з одним. Це спілкування представлено ребрами. Загальна кількість ребер становить 159.

Мітки ground truth

Загальна кількість спільнот для набору даних становить 4. Таблиця 3.3 показує вузли в кожному кластері

Таблиця 3.3

Спільноти наземної істини для набору даних Дельфін

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	1, 5, 6, 7, 9, 13, 17, 19, 22, 25, 26, 27, 31, 32, 41, 48, 54, 56, 57, 60
1	0, 2, 10, 28, 30, 42, 47
2	4, 11, 15, 18, 21, 23, 24, 29, 35, 45, 51, 55
3	3, 8, 12, 14, 16, 20, 33, 34, 36, 37, 38, 39, 40, 43, 44, 46, 49, 50, 52, 53, 58, 59, 61

Графік кластеризації ground truth спільнот показаний на малюнку нижче.

Рис 3.2 Карта ground truth кластеризації для набору даних соціальних мереж дельфінів

Оцінка ground truth спільноти

Оцінки для спільнот, тобто модульності, силуету та оцінки Калінського та Харабаша, для спільнот наземної істини наведені в Табл. 3.4.

Таблиця 3.4

Оцінки наземних спільнот для набору даних дельфінів

Модульність	Силует	Калінський та Харабаш
0.519	0.117	6.392

3.1.3 Набір даних мережі «Електронна пошта» (Email-Eu-Core)

Цей набір даних був вибраний, оскільки він представляє велику мережу реального світу і був використаний у літературі для аналізу характеристик алгоритмів [Венкатесарамані та Воробейчик, 2018]. Таким чином, набір даних буде служити еталоном для ефективності алгоритмів Лувена та K-середіх у великих мережах соціальних даних, де спільноти, як правило, перетинаються. Набір даних був наданий SNAP (2018).

Мережа Email-Eu-core представляє дані електронної пошти великої європейської дослідницької установи. Вузли представляють людей інституту, а ребра представляють будь-яку електронну пошту, яку вони могли надіслати іншій людині з університету. Будь-яке зовнішнє спілкування з рештою світу не є частиною набору даних.

Дані представлені незваженим, спрямованим графом із 1005 вузлами та 25571 ребрами.

Основну інформацію для набору даних надає SNAP (2018). Ground truth складається з 42 відділів (спільнот), і кожна особа (вузол графу) пов'язана саме з одним відділом (спільнотою). Основні дані для 42 громад також були надані SNAP (2018).

Ground truth мітки

Оцінки для показників доброти спільноти для ground truth спільнот наведені в таблиці 3.5. Як видно, набір даних має негативне значення для оцінки Силуету. Це означає, що між даними існує певне перекриття.

Таблиця 3.5

Показники узгодженості для основної істини набору даних «Електронна пошта»

Модульність	Оцінка силуету	Калінський та Харабаш
0.42	-0.197	7.129

У наведеній нижче таблиці показано порівняльний підсумок трьох наборів даних.

Таблиця 3.6

Короткий опис функцій наборів даних

	Клуб карате	Дельфіни	Електронна пошта
Вузли	34	62	1005
Ребра	78	159	25571
Ground спільноти	2	4	42
Ground Модульність	0.37	0.519	0.42
Ground Силует	0.16	0.117	-0.197
Ground Калінський та Харабаш	7.83	6.392	7.129

3.2 Показники ефективності

У ситуаціях, коли ground truth спільноти відома, процес оцінки ефективності здійснюється шляхом простого порівняння виявлених спільнот з відомими (Han et al., 2011). Різні показники продуктивності, запропоновані в дослідженні алгоритмів кластеризації, можна класифікувати на три основні категорії:

1.Заходи на основі підрахунку пар - на основі підрахунків точок, з якими пара не погоджується або погоджується, наприклад індекс Ранда, індекс Жаккарда, скоригований індекс ранду (ARI).

2.Заходи на основі зіставлення множин, на основі потужності множин,

спрямовані на пошук найбільших перетинів між парами вершин, що належать до різних кластерів, наприклад, чистота (Purity).

3.Заходи на основі теоретичної інформації - засновані на спільній інформації, що ділиться між двома кластерами для перевірки їх згоди, наприклад, Нормована взаємна інформація (NMI), Коригована взаємна інформація (AMI).

Вибір способів оцінювання важливий для типу аналізованих рішень. Наприклад, показники оцінки, точність та подібність Жаккарда (Jaccard similarity) не є ідеальними оцінювачами для двійкової або багаторівневої кластеризації. Це відбувається тому, що в ситуаціях, коли ідентифіковані кластери містять ті ж вузли, що і ground truth спільнот, але їх міткам присвоюється інша мітка, їм присвоюється нульовий бал. Що стосується двійкової та багаторівневої кластеризації, обидві метрики поводяться однаково, обидві вони не були обрані.

ARI підходить для ситуацій, коли результати кластеризації мають великі шанси мати великі кластери однакового розміру. Хоча AMI підходить для ситуацій, коли результати кластеризації є невеликими незбалансованими кластерами [Romano, 2016].

У ситуаціях, коли немає інформації про структуру суспільства, використовуються показники вимірювання якості спільноти [Han et al., 2011]. Ці показники базуються на різних характеристиках, які належать до стабільної структури спільноти. Ці показники допомагають гарантувати, що спільноти формуються наборами щільно з'єднаних вузлів, наприклад, за модульністю, оцінкою силуету, Калінським та Харабашем, тощо.

Для того, щоб зрозуміти якість сформованих кластерів за двома алгоритмами, слід було вивчити обидва оцінювачі продуктивності, тобто якість сформованих кластерів, а також точність щодо основної істини. На основі цих вимог системи було обрано десять змінних продуктивності з трьох категорій для оцінки двох алгоритмів. Короткий огляд десяти показників ефективності буде детально описано в цьому розділі. Ідея вибору діапазону оцінювальних метрик полягала в тому, щоб проаналізувати алгоритми з усіх пов'язаних точок зору.

3.2.1 Показнки узгодженості

Мережеві спільноти існують у різних формах. Вони можуть бути розрізненими, перекриваються, ієрархічними тощо. Залежно від типу застосунків існують різні евристичні засоби для вимірювання якості або узгоджентсть спільнот у мережі [Чакраборті та ін., 2017]. Ці евристичні показники визначають рівень якості сформованих спільнот.

3.2.1.1 Модульність

Модульність вимірює щільність вузлів, тобто взаємодію вузлів у межах спільноти проти взаємодій за межами спільноти. Значення коливаються від -1 до 1. Значення 1 вказує на дуже стабільну спільноту, тоді як від'ємне значення вказує на те, що вузли знаходяться в неправильному кластері. Метрика широко використовується в літературі для вимірювання якості кластерів, виявлених різними алгоритмами.

Модульність мережного графіка G визначається як:

$$Modularity = \frac{1}{2e}\sum_{i,j\in G}\left(AM - \left(\frac{d(i)*d(j)}{2e}\right)\right)*\delta(i,j)$$

Де,

- AM - матриця суміжності
- e - кількість ребер у мережі
- δ (i, j) дорівнює 0, якщо вузол і та вузол j не перебувають в одній спільноті, та дорівнює 1, якщо два вузли і та j знаходяться в одній спільноті
- d (i) - це ступінь вузла і, тобто кількість ребер, з'єднаних з вузлом і
- G - мережевий графік

Важливим аспектом показників модульності є те, що через межу роздільної здатності метрика не може точно оцінити спільноти малих розмірів (Fortunato and Barthelemy, 2007).

3.2.1.2 Оцінка силуету

Оцінка силуету повертає співвідношення внутрішніх та міжкластерних відстаней. Його значення знаходиться між -1 і 1. -1 вказує на неправильну ідентифікацію, 0 вказує на наявність спільнот, що перекриваються, а 1 вказує на здорову групу. Література показує, що якість кластерів була оцінена за допомогою цієї метрики (Фаліх, 2018; Фаньян, 2012).

Оцінка обчислюється шляхом знаходження відстані вузла від центроїда спільноти, до якої він належить, а потім порівняння його з відстанню вузла від центроїда найближчої сусідньої спільноти, до якої він не належить. Якщо присвоєння правильне, то перша відстань (призначена спільнота) має бути меншою за останню (сусідню) відстань. Ці відстані вузлів накопичуються і нормуються загальною кількістю вузлів мережі. Чим більше значення, тим краще кластеризація, оскільки це вказуватиме на те, що вузли розташовані ближче до центроїдів порівняно з центроїдами сусідніх спільнот.

Якщо $\overline{C}$ є центроїдом спільноти, до якого належить вузол i, та $C_{nearest}$ є центроїдом найближчої спільноти, до якої не належить вузол i, то, формально, оцінку силуету S можна визначити як:

$$S = \frac{1}{N}\sum_{i \in N} silhouette(i)$$

Де,

$$silhouette(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Де,

$$b_i = \frac{1}{|C_{nearest}|}\sum_{j \in C_{nearest}} distance(i, j)$$

Та

$$a_i = distance(i, \overline{C})$$

Найближчий сусід визначається шляхом оцінки відстані всіх сусідніх центроїдів, а потім вибору найближчого.

3.2.1.3 Оцінка Калінського та Харабаша

Оцінка Калінінського та Харабаша, яка також називається терміном Критерій коефіцієнта дисперсії, цей показник визначає співвідношення між внутрішньою та міжкластерною дисперсіями. Метрика була використана в літературі для оцінки якості кластерів за відсутності обґрунтованої істини (Калінський та Харабаш, 1974; Фаліх, 2018; Фаньян, 2012).

Формально критерій коефіцієнта дисперсії, VRC, визначається як:

$$VRC = \frac{CDist_{inter}}{CDist_{intra}} \times \frac{|N| - k}{|k| - 1}$$

Де,

- N - мережа
- K - це сукупність спільнот
- C - це вибрана спільнота з k
- $\bar{C}$ є центроїдом спільноти C
- $\bar{\bar{N}}$ є центроїдом цілої мережі N
- $CDist_{intra}$ є внутрішньокластерною дисперсійною або розсіяною матрицею:

$$CDist_{intra} = \sum_{C \in k} \sum_{i \in C} dist(i, \bar{C})^2$$

$CDist_{inter}$ є міжкластерною дисперсійною або розсіяною матрицею:

$$CDist_{inter} = \sum_{C \in k} |C| * dist(\bar{C}, \bar{N})^2$$

$\frac{|N|-k}{|k|-1}$ є терміном нормалізації, який запобігає монотонному зростанню показника VRC зі збільшенням кількості кластерів. Це робить VRC критерієм оптимізації. Велике значення VRC вказує на те що спільноти компактні. Більше значення означає, що спільноти чітко визначені, тобто їх внутрішні відстані невеликі (компактні) порівняно з зовніщніми відстанями.

3.2.2 Критерії ефективності методів

Заходи в цьому розділі допомагають порівняти продуктивність кластеризації з ярликами спільноти наземної правди. Ці заходи були використані в літературі для оцінки продуктивності алгоритмів кластеризації.

3.2.2.1 NMI (нормалізована взаимна інформація)

NMI базується на ентропії Шеннона в теорії інформації, це порівняння інформації про взаємну ентропію, що ділиться між двома кластерами (Zhang et al., 2018). Значення 1 означає високу кореляцію, тобто ідентичні результати кластеризації, тоді як значення 0 вказує на низьку кореляцію, тобто незалежні результати кластеризації. Метрика була використана для оцінки результатів кластеризації.

Якщо H є ентропією кластера, а два кластери для порівняння - це A і B, оцінка NMI визначається шляхом відношення їхньої взаємної інформації I до суми їх окремих ентропій H (A) і H (B).

$$NMI(A,B) = \frac{2 * I(A,B)}{H(A) + H(B)}$$

Ентропія кластеру ґрунтується на всіх складових кластерах. Якщо C - кластер у кластері A, N - розмір вузлів кластеру C, то ентропію A можна знайти як:

$$H(A) = -\sum_{C \in A} P(C) * \log_2 P(C)$$

Де $P(C) = |C|/N$.

Для іншого кластеризації B, де D - кластер B, а P (C, D) спільна ймовірність визначається як $P(C,D) = |C \cap D|/N$. , умовна ентропія H(A|B) оцінюється як:

$$H(A|B) = \sum_{C \in A} \sum_{D \in B} P(C,D) * \log_2 \frac{P(C)}{P(C,D)}$$

Взаємна інформація I (A, B) оцінюється як:

$$I(A,B) = H(A) - H(A|B)$$

3.2.2.2 Чистота (Purity)

Чистота дає частку правильно позначених членів. Кожен ідентифікований кластер узгоджується з одним кластером з еталонних кластерів, з яким він має максимальне перекриття. Тоді підрахунок подібних вузлів дає точність відповідності. Оцінка чистоти 1 дає повну відповідність. Метрика була використана для вимірювання результатів кластеризації алгоритмів (Rabbany et al., 2010; Hu, 2015).

Формально, чистота кластеризації (розділу), C, щодо істинної ґрунтовності, $\hat{C}$, визначається через:

$$purity\ (C,) = \frac{1}{N}\sum_{k=1}^{K} max_{l\in\{1,\dots,\widehat{K}\}}|C_k| \cap \left|\widehat{C_l}\right| \in [0,1]$$

Де,

N - мережа

C - це розділ N, що складається з неперекривних спільнот, тобто, C = {C1, C2, C3, ..., CK}

$\hat{C}$ - є основним розподілом істини, тобто, $\hat{C} = \{\widehat{C_1}, \widehat{C_2}, \widehat{C_3}, \dots, \widehat{C_{\widehat{K}}}\}$

Інтуїтивно, чистота вимірює частку вузлів, які були правильно позначені. Метрику не можна використовувати для визначення якості кластерів. Це пояснюється тим, що в ситуації, коли всі вузли виділені їх окремим спільнотам, показник чистоти буде 1.

3.2.2.3 AMI (Adjusted Mutal Information)

AMI вимірює подібність між кластерами і не залежить від абсолютних значень кластерів. Оскільки це показник, заснований на шансах, для мережі з великою кількістю кластерів AMI, як правило, високий. Для незалежних кластерів оцінка становить 0. Для подібних випадків значення становить 1. Метрика використовується в літературі для оцінки результатів кластеризації (Фенг, 2014).

Формально AMI визначається як,

$$AMI(A,B) = \frac{I(A,B) - E\{I(A,B)\}}{\sqrt{H(A)H(B)} - E\{I(A,B)\}}$$

Де

1. A - це мітки кластеризації
2. B - це основні мітки істини
3. H(A) - ентропія кластеризації A
4. H(B) - ентропія кластеризації B
5. I(A, B) - це взаємна інформація
6. E{I(A, B)} - очікуване значення взаємної інформації між усіма можливими кластерними парами

Чим вище значення AMI, тим кращі результати кластеризації.

3.2.2.4 ARI (Adjusted rand index)

ARI дає співвідношення кількості вузлів, які були правильно ідентифіковані. Метрика штрафує помилкові негативи та хибнопозитивні. Метрика змінюється від -1 до 1. Оцінка ARI 1 означає, що розділи/кластери, як очікується, ідентичні, -1 вказує на відсутність подібності, тобто відсутність згоди, тоді як оцінка 0 показує випадкову непереконливу згоду. Метрика була використана для оцінки алгоритмів кластеризації (Рабані та ін., 2010; Коллет, 2015; Тенг, 2017; Вагнер і Вагнер, 2007; Фагнан, 2012).

Якщо X = {X1, X2, X3, …, Xm} and Y = {Y1, Y2, Y3, …, Yn} - це два розділи (кластеризація), то їх перекриття (перетин) може спостерігатись за їхньою таблицею непередбачених обставин [ni,j]. Кожен запис таблиці представляє спільні вузли між двома кластеризаціями Xi і Yj тобто ni,j = |Xi ∩ Yj|. If {a1, a2, a3, …, an} представляє суми відповідних рядків матриці непередбачених ситуацій, а {b1, b2, b3, … , bm} - суми відповідних стовпців матриці непередбачених обставин, потім формально скоригований випадковий індекс, ARI можна визначити як:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

Де

n - загальна кількість вузлів мережі

$\sum_{i,j}\binom{n_{ij}}{2}$ є індексом

$\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}$ є очікуваним індексом

$\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right]$ є максимальним індексом

3.2.2.5 Індекс Фаулкса-Мэллоуса FMI (Fowlkes and Mallows Index)

Точність вимірює, наскільки точні виявлені кластери. Нагадаємо, вимірюється, скільки кластерів алгоритм міг виявити (Фенг, 2015).

$$Precision = \frac{TP}{TP + FP}$$

Поки

$$Recall = \frac{TP}{TP + FN}$$

Де

- TP - справжній позитив
- FP - хибнопозитивний
- FN - хибнонегативний

Розглянемо порівняння кластеризації C з основою істини, T. Вищевказані заходи можна визначити наступним чином:

1.Справжнє позитивне - це кількість пар вузлів, які знаходяться в одному кластері для виявленої кластеризації, C та наземної кластеризації, T.

2.Хибнопозитивне - це кількість пар вузлів, які знаходяться в одному кластері для виявленої кластеризації, C, але в іншому класі наземних міток, T.

3.Помилково негативне - це кількість пар вузлів, які знаходяться в різних кластерах як у виявленому кластері, C, так і в кластері істини, T.

FMI - це показник подібності, заснований на середньому геометричному рівні Precision and Recall (Fowkles and Mallows, 1983). Метрика була використана в літературі для оцінки результатів кластеризації (Вагнер та Вагнер, 2007).

Формально FMI визначається як:

$$FMI = \frac{TP}{\sqrt{(TP + FP) * (TP + FN)}}$$

Значення FMI коливається від 0 до 1, де 1 вказує на хорошу подібність між кластером та основою правди.

3.2.2.6 F1 (F-міра)

Оцінка F1 - це середньозважене значення (середнє гармонічне значення) Точності та Відкликання, яке враховує як хибнопозитивні, так і хибнонегативні. Він використовувався в літературі для вимірювання продуктивності алгоритмів класифікації (Лі та ін., 2008), а також алгоритмів кластеризації (Фенг, 2015; Вагнер та Вагнер, 2007).

Загальна формула для F-Оцінка така:

$$F_\beta Score = (1 + \beta^2) \frac{Precision.Recall}{\beta^2.Precision + Recall}$$

Де β - позитивне значення (зазвичай 0,5, 1 і 2), що додає ваги точності та відкликання.

Для оцінки F1 значення β дорівнює 1, тому формула набуває вигляду,

$$F_1\ Score = 2 * \frac{Precision.Recall}{Precision + Recall}$$

Найкраще значення балу F1 досягається на 1, а найгірше - на 0.

3.2.2.7 Однорідність (Homogeneity)

Ця міра не залежить від абсолютних значень етикетки. Враховуючи основну істину, кластер був би однорідним, якщо всі його точки містяться в одній мітці. Його значення коливається від 0 до 1, де 1 означає повну однорідність. Метрика була використана для оцінки результатів кластеризації (Розенберг та Хірчберг, 2007; Фаліх, 2018).

Припускаючи наступну інформацію:

1. N - загальна кількість вузлів у мережі

2. C представляє мітки класів такі, що C = {ci | i = 1, 2, …, n}

3. K представляє набір виявлених кластерів таким, що K = {ki | i = 1, 2, …, m}.

4. A - таблиця непередбачених ситуацій, що представляє два кластеризації C та K такі що A = {aij} де aij вузли, які є членами класу ci і виявлені в кластерах kj.

Тоді формально однорідність h визначається як:

$$h = \begin{cases} 1 & if\ H(C,K) = 0 \\ 1 - \dfrac{H(C|K)}{H(C)} & else \end{cases}$$

де

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|} a_{ck}}{n} log \frac{\sum_{k=1}^{|K|} a_{ck}}{n}$$

Для випадку ідеальної однорідності коефіцієнт нормалізації $\frac{H(C|K)}{H(C)}$ дорівнює 0.

3.2.3 Метод коефіцієнта силуету (Silhouette Coefficient)

Вхідний параметр алгоритму K-середіх це загальна кількость сформованих кластерів, k. Для будь-якого значення k алгоритм K-середіх породжує спільноти мережі. На основі метрики «Коефіцієнт силуету», де ідентифіковані мітки пов’язані з базовою позицією у вихідних даних, визначається зв’язок між вхідним значенням k та відповідним значенням для Коефіцієнта силуету. Метод полягає у виборі значення k, для якого коефіцієнт силуету є максимальним. Співвідношення k та оцінки силуету найкраще можна побачити на лінійному

графіку між двома значеннями. Встановивши значення k від 2 до загальної кількості вузлів та вивчивши відповідні показники силуету, можна отримати оптимальне значення k для мережі [Дестерк, 2018; Ковальчук, 2009].

3.2.4 Критерій ліктя (Elbow Criterion)

Використовуючи критерій силуету, може статися, що велике значення повертається лише кількома кластерами. Щоб уникнути виявлення дуже небагатьох кластерів, для визначення значення k також розглядається Критерій ліктя [Дестерк, 2018]. Сума квадратичної помилки (SSE) - це сума квадратної відстані кожного вузла, включеного в кластер від його центроїда. SSE оцінюється для кожного значення k. За допомогою евристичного критерію ліктя вибирається значення k, для якого відбувається різка зміна значення SSE. Відношення k та SSE найкраще можна побачити на лінійному графіку між двома значеннями. Загальна тенденція така: при збільшенні k значення SSE зменшується. SSE стає 0, коли k стає рівним загальній кількості вузлів на графі. Це пояснюється тим, що кожен вузол сам стає кластером, і різниця між вузлом кластера та центроїдом більше не існує. Метою критерію ліктя є вибір значення для k, яке є малим, а також має низьке значення SSE. Лікоть означає точку, після якої збільшення значення k спричиняє рівномірне зменшення SSE.

Висновки до розділу 3

У цьому розділі були описані всі набори даних різного розміру, які є загальнодоступними для дослідницьких цілей: Набір даних «Карате-клуб» «Захарія», набір даних соціальних мереж «Дельфін» та великий набір даних мережевих комунікацій «Електронна пошта». Усі три набори даних мають спільноти, що не перекриваються, і загальнодоступні в Інтернеті. Розділ також описує показники ефективності для методів.

4 АНАЛІЗ ОЦІНОК АЛГОРИТМІВ НА ОСНОВІ ПОКАЗНИКІВ ПРОДУКТИВНОСТІ

4.1 Оцінка методу Лувена

4.1.1 Застосування методу Лувена для набору даних «Карате-клуб»

За допомогою алгоритму Лувена кількість знайдених спільнот склала 4. Деталі вузлів у кожному кластері наведені в Таблиці 4.1.

Таблиця 4.1

Результати Лувена для набору даних «Карате-клуб»

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	4, 5, 6, 10, 16
1	0, 1, 2, 3, 7, 9, 11, 12, 13, 17, 19, 21
2	23, 24, 25, 27, 28, 31
3	8, 14, 15, 18, 20, 22, 26, 29, 30, 32, 33

Кластерна карта набору даних показана на малюнку нижче.

Рис. 4.1 Кластерна карта Лувена для набору даних клубу карате

Оцінки показників оцінки ефективності алгоритму Лувена наведені в табл. 4.2.

Таблиця 4.2

Оцінки спільнот Лувена для оцінки показників спільноти

Модульність	Силует	Калінського та Харабаша
0.418	0.146	5.54

Час, зайнятий алгоритмом Лувена, становив 0,0048 с. Показники для шести показників наведені в табл. 4.3.

Таблиця 4.3

Аналіз ефективності K Лувена (карате)

Чистота	NMI	AMI	ARI	Оцінка F1	Однорідність	FMI
0.98	0.61	0.44	0.49	0.8	0.83	0.63

4.1.2 Набір даних дельфінів

На відміну від очікуваних 4 спільнот, алгоритм знайшов 5 кластерів. Вузли в кожному кластері перелічені в табл. 4.4.

Таблиця 4.4

Результати Лувена для набору даних дельфінів

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	0, 2, 10, 42, 47, 53, 61
1	1, 7, 19, 25, 26, 27, 28, 30
2	12, 14, 16, 20, 33, 34, 36, 37, 38, 39, 40, 43, 44, 46, 49, 50, 52, 58
3	5, 6, 9, 13, 17, 22, 31, 32, 41, 48, 54, 56, 57, 60
4	3, 4, 8, 11, 15, 18, 21, 23, 24, 29, 35, 45, 51, 55, 59

Карта кластеризації для набору даних показана на малюнку нижче.

Рис 4.2 Кластерна карта Лувена для набору даних дельфінів

Оцінки для оцінки доброти спільноти алгоритму Лувена наведені в табл 4.5.

Таблиця 4.5

Оцінки спільнот Лувена для оцінки показників спільноти

Модульність	Силует	Калінського та Харабаша
0.518	0.108	5.75

Час, зайнятий алгоритмом Лувена, становив 0,0051 с. Оцінки за шість показників ефективності наведені в табл. 4.6.

Таблиця 4.6

Аналіз продуктивності Лувена (дельфіни)

	Чистота	NMI	AMI	ARI	Однорідність	Оцінка F1	FMI
Оцінка	0.887	0.73	0.64	0.64	0.79	0.032	0.74

4.1.3 Застосування методу Лувена для набору даних «Електронна пошта»

Подібно до попередніх двох наборів даних, алгоритм Лувена був безпосередньо застосований до мережного графіка. Час, необхідний алгоритму для формування спільнот, склав 0,0245с. Алгоритм Лувена знайшов 27 спільнот. А модульність для утворених громад склала 0,42. Таблиця 4.7. нижче показує показники доброти для спільнот, сформованих алгоритмом.

Таблиця 4.7

Показники добробуту спільноти Лувена для набору даних із Електронна пошта

Витрачений час (сек)	Створені спільноти	Модульність	Оцінка силуету	Калінський та Харабаш
0.0245	27	0.42	-0.295	3.96

Оцінки семи змінних показників наведені в табл. 4.8.

Таблиця 4.8

Аналіз ефективності K-середіх для k = 2 Електронна пошта

	Чистота	NMI	AMI	ARI	Однорідність	Оцінка F1	FMI
Оцінка	0.393	0.55	0.337	0.25	0.417	0.04	0.38

4.2 Оцінка методу K-середіх

4.2.1 Застосування методу К-середніх для набір даних «Карате-клуб»

Щоб знайти значення k, був виконаний критерій ліктя. Для кожного значення k записували відповідний бал SSE. Діаграма для SSE для значень k, починаючи з 1 до 34, потім була створена для пошуку ліктя, як показано на малюнку нижче.

Рис 4.3 Критерій ліктя для набору даних клубу карате (K-середіх)

Оскільки якихось окремих ліктів не було, значення традиційних K-середіх SSEs ('144.18', '115.82', '99.18', '85.26', '75.83', '68.47', '69.12', '53.43', '54.38', '47.41', '48.17', '42.49', '40.90', '38.02', '31.96', '29.08', '28.20', '24.33', '21.50', '21.83', '16.67', '17.17', '16.50', '12.33', '11.83', '9.50', '7.50', '7.50', '5.50', '4.50', '3.00', '2.00', '1.00') відповідно до значень k = 1 до k = 34. Найбільше падіння значень було для значення, що відповідає k = 2, тобто від 115,8 до 99,1. Отже, відповідно до критерію ліктя, було вибрано значення k = 2.

Для підтвердження такого рішення також були враховані оцінки показників ground truth спільноти, тобто оцінка силуету, модульність та Калінський та Харабаш.

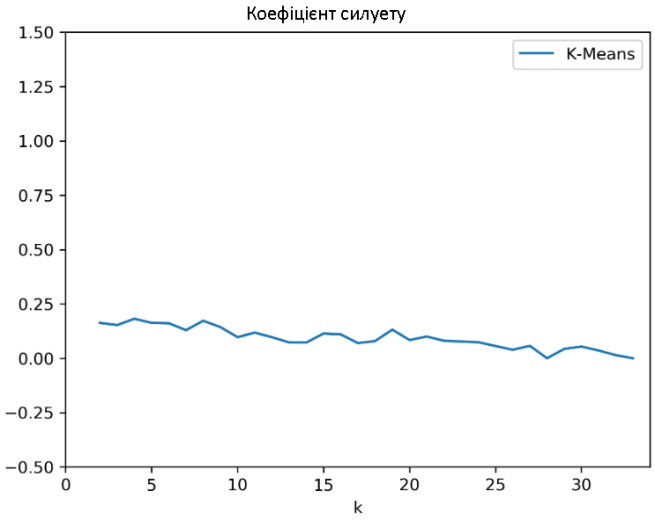

Рис. 4.4 Графік силуетів для клубу карате (K-значення)

Рис. 4.5 Графік модульності для клубу карате (K-середніх)

*Р*ис. 4.6 Оцінки Калінського та Харабаша для клубу карате (K-значення)

Щоб виділити процес вибору k, у табл. 4.9. наведені значення трьох показників ефективності для діапазону k від 2 до 10. Діапазон був вибраний, оскільки він містив максимальні значення трьох показників.

Таблиця 4.9

Аналіз показників добробуту громади для клубу карате (K-значення)

K	Силует	Модульність	Калінського та Харабаша
2	0.16	**0.37**	**7.83**
3	0.15	0.31	7.03
4	**0.18**	0.10	6.91
5	0.16	0.16	6.53
6	0.16	0.08	6.19
7	0.13	0.19	4.89
8	0.17	0.13	6.31
9	0.14	0.06	5.16

10	0.10	0.10	5.44

Значення модульності та індексу Калінського та Харабаша є найвищими при k = 2. Оцінка силуету показала максимальну продуктивність при k = 4. Але враховуючи більшість голосів за k = 2, вона була обрана як кількість кластерів для подальшого аналізу.

Після того, як було вибрано k, вивчалися оцінки семи показників ефективності для k від 2 до 8, як показано в таблиці 15. Як видно, для k = 2 алгоритм дав найкращі результати. Це показує, що вибране k справді формувало кластери найкращої якості, що також збігалося з ground truth спільнотами.

Таблиця 4.10

Показники ефективності K-значень

K	Чистота	NMI	AMI	ARI	Однорідність	FMI	Оцінка F1
2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
3	1	0.88	0.78	0.88	0.99	0.93	0.2
4	1	0.71	0.49	0.52	1	0.71	0.7
5	0.97	0.61	0.41	0.53	0.83	0.72	0.02
6	1	0.65	0.39	0.41	1	0.64	0
7	1	0.62	0.34	0.33	1	0.57	0.5
8	1	0.59	0.3	0.26	1	0.50	0.14

Деталі вузлів у двох кластерах наведені в табл. 4.11.

Таблиця 4.11

Результати K-значень для набору даних клубу карате

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21

1	8, 9, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

Карта кластера для набору даних показана на рис. 4.7.

Рис. 4.7 Кластерна карта K-значень для набору даних клубу карате

У таблиці нижче показано кількість ітерацій та час, необхідний алгоритму для конвергенції. Середній час K-значень становив 0,0054 с.

Таблиця 4.12

Час та ітерації K-Значення

K	Ітерації	Час (сек)
2	4	0.0072
3	7	0.0049
4	3	0.0047
5	4	0.0048
6	6	0.0047
7	5	0.0052
8	3	0.0066
9	2	0.0062
10	5	0.0059

Таблиця показує, що алгоритму знадобилося менше секунди для виконання кластеризації.

4.2.2 Набір даних дельфінів

Подібно до набору даних клубу карате, значення k було обрано прагматично. Для значень від 1 до загальної кількості вузлів графік для критерію ліктьового суглоба показаний на малюнку нижче.

Рис. 4.8 K-значення критерій ліктьового значення для набору даних дельфінів

Враховуючи значення SSE для K-значень (330.02, 300.85, 279.16, 259.53, 255.13, 239.67, 232.77, 226.52, 225.48, 213.27, 209.53, …, 11.50, 8.00, 5.67, 3.00, 1.00), найбільше падіння після цього зменшення стає більш плавним для значень k як 4 для 259,53 бала. Після цього зниження SSE було рівномірним.

Окрім критерію ліктя, усі показники ефективності за трьома показниками (оцінка силуету, модульність та бали Калінінського та Харабаша) також допомогли у виборі значення k.Діаграми K-середіх для варіації k щодо трьох показників показані на малюнках нижче.

Рис. 4.9 Графік силуету K-середніх «Дельфін»

Рис. 4.10 Графік модульності K-значень (дельфіни)

Рис. 4.11 Оцінка Калінського та Харабаша для K-значень (дельфіни)

Щоб пояснити процес вибору, значення трьох показників для діапазону k від 2 до 10 наведено у Таблиці 18. Цей діапазон був обраний, оскільки максимальні значення всіх показників знаходилися в цьому діапазоні.

Таблиця 4.13

Аналіз ефективності K для мережі «дельфін» (K-середніх)

K	Силует	Модульність	Калінського та Харабаша
2	**0.1**	0.3	5.8
3	0.09	0.23	**5.37**
4	0.08	**0.34**	5.25
5	0.07	0.21	4.18
6	0.06	0.24	4.22
7	0.06	0.26	3.8
8	0.05	0.2	3.52
9	0.02	0.18	3.07

10	0.04	0.09	3.16

Виходячи з перерахованих балів у таблиці, оцінка силуету була найвищою при k = 2. Калінський та Харабаш дали найвищий показник при k = 3. Модульність дала найвищий показник при k = 4. Враховуючи більшість подібності, використовуючи оцінки SSE та модульності, було вибрано k = 4.

Дані про ефективність щодо базової істини для k порівняно з сімома показниками показані на малюнку нижче.

Рис. 4.12 Показники ефективності K-значень для набору даних дельфінів

Підмножина цих показників для семи показників продуктивності щодо базової істини наведена у Таблиці 19. Щоб побачити зміну значень показників для зміни k, наведені значення для діапазону k від 2 до 8 у таблиці. Вибране значення k для K-значень було виділено.

Таблиця 4.14

Показники ефективності оцінки K-середніх для мережі «Дельфін»

K	Чистота	NMI	AMI	ARI	Однорідність	FMI	Оцінка F1
2	0.56	0.32	0.32	0.12	0.27	0.44	0.38
3	0.61	0.4	0.3	0.1	0.37	0.40	0.17
4	**0.56**	**0.33**	**0.25**	**0.07**	**0.32**	**0.37**	**0.29**
5	0.6	0.37	0.28	0.11	0.39	0.37	0.12
6	0.67	0.47	0.35	0.19	0.52	0.4	0.2
7	0.64	0.38	0.24	0.07	0.43	0.3	0.16
8	0.7	0.41	0.27	0.15	0.46	0.39	0.06

Порівняння кількості ітерацій за алгоритмом показано в таблиці нижче. Середній час K-значень становив 0,0056 с. Час вибраного значення k становив 0,0047 с.

Таблиця 4.15

Час та зміни K-середніх для мережі «Дельфін»

K	Ітерації	Час (сек)
2	12	0.0048
3	9	0.0066
4	**4**	**0.0047**
5	4	0.0046
6	6	0.0063
7	3	0.0066
8	4	0.0047
9	3	0.005
10	5	0.0047

Детальний результат кластеризації за алгоритмом K-значень для набору даних при k = 4 наведено в Таблиці 22.

Таблиця 4.16

Результати K-значень набору даних «Дельфін»

Мітка спільноти	*Ідентифікатори вузлів у спільноті*
0	0, 1, 2, 3, 4, 7, 8, 10, 11, 12, 15, 19, 20, 22, 23, 25, 26, 27, 28, 30, 31, 32, 35, 36, 39, 42, 44, 46, 47, 48, 49, 52, 53, 55, 58, 59, 60, 61
1	14, 16, 33, 34, 37, 38, 40, 43, 50
2	5, 6, 9, 13, 17, 41, 54, 56, 57
3	18, 21, 24, 29, 45, 51

Кластерна карта для набору даних для K-значень показана на малюнку нижче.

Рис. 4.13 Кластерна карта K-значень для набору даних «Дельфін» (k=4)

4.2.3 Застосування методу К-середіх для набору даних «Електронна пошта»

Щоб знайти оптимальне значення k, було розглянуто критерій ліктьового суглоба. Для діапазону k від 1 до 100 графік для SSE показаний на малюнку нижче.

Рис. 4.14 Ліктевий аналіз K-значень (Електронна пошта)

Враховуючи значення SSE для K-значень ('30681.44', '28057.21', '27226.72', '26367.67', '25542.75', '24915.37', '24526.53', '23639.54', '23559.98', '23124.40', '22780.39', …, '16451.01', '16501.98', '16289.76', '16527.40', '16138.34', '16248.95', '16071.48', '16180.69', '16181.18', '16005.31'), найбільше падіння після чого зменшення стає більш плавним для значень k як 7.

Щоб перевірити це значення k для мережі, аналіз силуету був виконаний для діапазону значень k від 2 до 100. На малюнку нижче показано графік коефіцієнтів силуету щодо кількості кластерів, вибраних для застосування K-середіх. Значення k обирали на основі сукупних балів SSE та інших показників доброти спільноти, тобто силуету, модульності та оцінки Калінського та Харабаша.

Рис. 4.15 Силует та модульність для «Електронна пошта»

Значення коефіцієнта силуету було найвищим 0,26 для k = 2. Хоча при цьому значенні k модульність була найвищою, тобто 0,26 при k = 15.

Показник подібності був загалом більш ніж у три рази вищим, ніж раніше вивчені набори даних клубу карате та дельфінів. При k = 2 значення було найвищим, тобто 93,8.

Рис. 4.16 Оцінки Калінського та Харабаша для набору даних «Електронна пошта»

Високі показники для трьох хороших змінних спільноти вказують на те, що вузли у сформованих спільнотах перебувають у стабільному положенні, тобто вони мають більш сильне відношення до кластера, до якого вони входять, а не до сусідніх кластерів. Низькі значення вказують на те, що вузли можуть бути частиною іншої сусідньої спільноти.

Як видно з графіків оцінки ефективності, спільноти, утворені в діапазоні від 2 до 100, є дуже нестабільними. Усі бали знижуються зі збільшенням значення k.

У таблиці нижче показано вибране значення k на основі максимальних значень оцінок.

Таблиця 4.17

Оптимальне значення k для K-значень з використанням показників якості

	SSE	Модульність	Силует	Калінський і Харабаш
k	7	15	2	2

Таким чином, виходячи з більшості значень коефіцієнта силуету та оцінки Калінського та Харабаша, значення k було вибрано як k = 2. Низьке значення модульності вказує на відсутність чітких меж між спільнотами, сформованими для мережі.

Таблиця 4.18

K-середіх показники спільноти за k = 2 «Електронна пошта»

	Силует	Модульність	Калінський і Харабаш
Оцінка	0.26	0.1	93.8

Сім показників продуктивності для вибраного значення k як 2 показано у таблиці нижче. Низьке значення для балів вказує на те, що оцінені кластери взагалі не збігаються з основною істиною.

Таблиця 4.19

Аналіз ефективності K-середіх для k=2 «Електронна пошта»

	Чистота	NMI	AMI	ARI	Однорідність	Оцінка F1	FMI
Оцінка	0.1	0.05	0.01	0.008	0.008	0.05	0.2

З первинної істини ми знаємо, що кількість кластерів у мережі становить 42. Аналіз результатів роботи K-середіх навколо цієї основної вартості істини покаже, чи могло б бути покращення в роботі алгоритму, якби інше k було обрали. Якщо існує таке значення k, це означало б, що аналіз коефіцієнта силуету та критерій ліктьового суглоба не змогли виявити оптимального k.

Як видно з результатів, значення k сильно коливається для всіх вищезазначених методів визначення k. Аналізуючи основу правди спільноти на значення k, показники добробуту спільноти при k = 42 наведені в таблиці 25. Як видно, силует та оцінка Калінського та Харабаша зменшуються, але модульність зростає. Збільшення модульності означає, що межа між кластерами стає більш чіткою.

Таблиця 4.20

Оцінки доброти спільноти K-середіх для k = 42 «Електронна пошта»

	Силует	Модульність	Калінський та Харабаш
Оцінка	0.09	0.19	14.5

оцінки семи показників ефективності при k = 42.

Таблиця 4.21

Аналіз ефективності K-середіх для k=42 «Електронна пошта»

	Чистота	NMI	AMI	ARI	Однорідність	F1 Оцінка	FMI

Оцінка	0.49	0.5	0.35	0.07	0.45	0.004	0.17

Усі показники продуктивності, крім оцінки F1 та FMI, продемонстрували поліпшення продуктивності, коли було використано число кластеру на основі істини. Це показує, що кластери, утворені K-середіх при k = 42, мали більшу схожість з основною істиною, ніж при k = 2. Це означає, що методам k-аналізу не вдалося виявити оптимального значення k. Крім того, загалом показники ефективності дуже низькі. Це означає, що K-середіх не кластеруються належним чином, щоб задовольнити вимоги спільнот наземної правди.

4.3 Оцінка методу K-середіх ++ (K-means++)

У цьому розділі наведено порівняння K-середіх ++ з алгоритмом K-середіх. Обидві версії реалізують Blondel та ін. Різниця між ними полягає в тому, як вибираються початкові центроїди. K-середіх вибирає їх випадковим чином, тоді як K-середіх ++ вибирає їх випадковим чином, але застосовує умову, що точки не близькі один до одного.

4.3.1 Застосування методу К-середніх для набору даних «Карате-клуб»

Щоб знайти значення k, був виконаний критерій ліктя. Розглянуто графік для SSE для значень k від 1 до 34, як показано на малюнку нижче.

Рис. 4.17 Критерій ліктя для клубу карате (K-значення проти K-значень++)

Вивчення значень K++ SSEs (144.18, 115.82, 99.18, 84.02, 74.99, 66.55, 60.79, 54.16, 49.59, 44.10, 40.53, 37.35, 32.63, 29.80, 26.80, 24.33, 21.33, 19.50, 17.33, 16.00, 14.00, 12.50, 11.00, 10.00, 9.00, 8.00, 7.00, 6.00, 5.00, 4.00, 3.00, 2.00, 1.00), що відповідають значенням k = 1 до k = 34, найбільше падіння значень було для значення, що відповідає k=2 тобто, від 115,8 до 99,1. Після цього падіння зниження балів було рівномірним. Отже, за критерієм ліктя було обрано значення k = 2. Значення k було подібним до значення алгоритму K-значень.

Щоб підтвердити наше рішення, оцінки добра спільноти за трьома метриками також вважалися силуетом, модульністю та Калінським та Харабашем. Графіки для трьох метричних балів як для K-середіх, так і для K-середіх ++ показані на рисунках нижче. Можна побачити, що K-середіх ++ дав більш високі оцінки силуету в порівнянні з K-середіх, але поведінка була протилежною Калінському і Харабашу. Для модульності K-середіх ++ дав більш плавну криву.

Рис. 4.18 Силуетний графік для клубу карате(K-значення проти K-значень++)

Рис. 4.19 Граф модульності для «Карате-клуб» (K-середніх проти K-середніх++)

Рис. 4.20 Шкала Калінського та Харабаша для «Карате-клуб» (K-середніх проти K-середніх++)

Між традиційними K-значень і K-значень ++, оцінки коливалися, але обидва поводилися дещо схоже на k = 2. При k=2 алгоритм KMeans++ прийняв більше ітерацій, ніж традиційні K-засоби, щоб сходитися. У таблиці нижче показано кількість ітерацій і час, який беруть два алгоритми для зближення. Середній час K-Значень++ становив 0,0056s на відміну від K-середіх, який становив 0,0054s.

Таблиця 4.22

Час і ітерації K-середіх і K-середіх ++ (Клуб карате)

K	Ітерації		Час (с)	
	K-середніх++	K-середніх	K-середніх ++	K-середніх
2	**7**	**4**	**0.008**	**0.0072**
3	3	7	0.006	0.0049
4	3	3	0.0045	0.0047

5	4	4	0.0046	0.0048
6	4	6	0.006	0.0047
7	3	5	0.0046	0.0052
8	2	3	0.0046	0.0066
9	3	2	0.0046	0.0062
10	3	5	0.005	0.0059

Графік ітерації для наведених вище даних показаний нижче.

Рис. 4.21 К-середніх ++ та. Ітерації К-середніх для набору даних клубу карате

У таблиці 28 наведено значення трьох показників ефективності для зміни значень k з 2 на 10. Значення модульності та індексу Калінського і Харабаша найвищі при k = 2. Оцінка силуету показала максимальну продуктивність при k = 11, але вона не збігалася з іншими показниками. Так, враховуючи більшість голосів за k = 2, він був обраний як кількість кластерів для подальшого аналізу.

Таблиця 4.23

Аналіз виконання варіацій К для мережі «Карате-клуб» (К-середніх ++)

К	Силует	Модуляція	Калінський і Харабаш
2	0.16	**0.37**	**7.83**
3	0.15	0.31	7.03
4	0.17	0.08	7.16
5	0.18	0.11	6.69
6	0.16	0.15	6.53
7	0.17	0.07	6.17
8	0.17	0.12	6.17
9	0.16	0.15	5.96
10	0.19	0.11	6.05
11	**0.20**	0.13	5.88

Оцінки за сім показників ефективності для k від 2 до 8 наведені в таблиці нижче. Як видно, для k = 2 алгоритм дав найкращі результати для К-середніх ++.

Таблиця 4.24

Оцінки продуктивності К-середніх ++

K	Чистота	NMI	AMI	ARI	Однорідність	FMI	F1 Оцінка
2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
3	1	0.88	0.78	0.88	0.99	0.93	0.058
4	1	0.71	0.49	0.52	1	0.71	0
5	0.97	0.61	0.41	0.53	0.83	0.72	0.65
6	1	0.65	0.39	0.41	1	0.64	0.27
7	1	0.62	0.34	0.33	1	0.57	0.27
8	1	0.59	0.3	0.26	1	0.50	0.25

Наведені нижче рисунки показують графіки оцінювання К-середніх++ на відміну від К-середніх. Порівняльні графіки показують, що, за винятком оцінки F1, оцінки для традиційних k-засобів були менш гладкими, тобто вони коливалися більш у порівнянні з К-середніх++. У формулі-1 обидва алгоритми показали коливання балів. Попри відмінності в балах, продуктивність обох алгоритмів була однаковою на k=2.

Рис. 4.22 Показники чистоти К-середніх (для k=2:10) для набору даних клубу карате

Рис. 4.23 К-середніх NMI Оцінки (для k=2:10) для набору даних клубу карате

Рис. 4.24 К-середніх AMI Оцінки (для k=2:10) для набору даних клубу карате

Рис. *4.25* К-середніх ARI Оцінки (для k=2:10) для набору даних клубу карате

Рис. 4.26 К-середніх Однорідність Оцінки (для k=2:10) для набору даних клубу карате

Рис. 4.27 К-середніх FMI Оцінки (для k=2:10) для набору даних клубу карате

Рис. 4.28 К-середніх F1 Оцінки (для k=2:10) для набору даних клубу карате

Деталі вузлів у межах двох кластерів показані в таблиці нижче.

Таблиця 4.25

К-середніх ++ Результати даних клубу карате

Community Label	*Node IDs in Community*
0	0, 1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21
1	8, 9, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33

Зіставлення кластера для набору даних показано на малюнку нижче.

Рис. 4.29 К-середніх ++ Карта кластеризації для набору даних клубу карате

4.3.2 Застосування методу К-середніх++ для набору даних «Дельфін»

Як і набір даних Клубу карате, значення k було обрано прагматично. Для значень k від 1 до загальної кількості вузлів були відзначені відповідні SSEs. А потім була сформована ділянка для ліктьової критерію, як показано на малюнку нижче.

Рис. 4.30 К-середніх Критерій ліктя для набору даних дельфінів

Лікоть не був ясним, тому враховуючи значення SSE для К-середніх++ (330.02, 310.88, 294.90, 259.30, 257.81, 246.62, 244.60, 228.30, ..., 10.67, 8.25, 6.17, 4.50, 3.00, 2.00, 1.00), найбільше падіння, після якого зниження стало більш плавним, було виявлено на рівні k 4 тобто 259,30.

Показники ефективності за трьома метриками (оцінка силуету, модульність і бали Калінського і Харабаша) також розглядалися, щоб допомогти вибрати значення k. Калінський і Харабаш і модульність за традиційними К-Засобами дали найвищий показник k = 4. У той час як модульність за традиційними К-засобами давала найвищу в k = 5. З огляду на більшість подібності, було обрано k = 4 К-середніх++ не вказував на послідовне k для всіх трьох показників. Отже, якби К-середніх++ розглядався для подальшого аналізу, значення k було б обрано як 5.

Значення для трьох показників для кожної ітерації наведені в таблиці нижче.

Таблиця 4.26

Варіаційний аналіз продуктивності К для мережі «Дельфін»

К	Силует	Модульність	Калінський і Харабаш
2	**0.17**	0.01	3.6
3	0.09	0.08	**3.5**
4	0.09	**0.38**	5.27
5	0.06	0.23	3.9
6	0.07	0.26	3.7
7	0.05	0.11	3.2
8	0.06	0.15	3.43
9	0.05	0.09	3.15
10	0.03	0.04	2.4

Графіки К-середніх++ для наведених вище метричних балів показані на рисунках нижче. Крива К-середніх також включена в графіки, щоб побачити загальні значення двох алгоритмів.

Рис. 4.31 Графік силуетів К-середніх та К-середніх ++

Рис. 4.32 Графік модульності К-середніх та К-середніх ++

Рис. 4.33 Граф оцінки Калінського і Харабаша К-середніх та К-середніх ++

Графіки значень для семи показників ефективності наведено на рисунках нижче. Відповідні значення для К-Засобів також були додані в графіки, щоб побачити порівняння балів двох версій алгоритму. Видно, що у всіх кривих оцінки К-середніх трохи вище, ніж К-середніх++. Але К-середніх++ утворюють спільноти, які більше пов'язані з наземною істиною в порівнянні з К-Засобами.

Рис. 4.34 Оцінки чистоти для набору даних «Дельфін»

Рис. 4.35 NMI Оцінка для набору даних «Дельфін»

Рис. 4.36 AMI оцінки для набору даних «Дельфін»

Рис. 4.37 ARI оцінки для набору даних Дельфіни

Рис. 4.38 Оцінки однорідності для набору даних «Дельфін»

Рис. 4.39 FMI оцінки для набору даних «Дельфін»

Рис. 4.40 F1 оцінки для набору даних «Дельфін»

Продуктивність К-середніх++ для семи показників показана на малюнку нижче.

Рис. 4.41 К-середніх++ оцінки виконання для набору даних «Дельфін»

Щоб побачити варіацію значень семи показників продуктивності для зміни k, у таблиці нижче показано підмножину значень діапазону k від 2 до 8. Виділено вибране значення k для К-середніх ++. Хоча криві К-середніх показують більш високі значення, але на k-4 оцінки К-середніх ++ вищі (див. таблицю порівняння в розділі Е або К-середніхоцінки ефективності з розділу С).

Таблиця 4.27

Шкала виконання К-середніх++ для дельфінів

K	Чистота	NMI	AMI	ARI	Однорідність	FMI	F1 Оцінка
2	0.45	0.18	0.07	-0.01	0.11	0.45	0.32
3	0.61	0.4	0.34	0.14	0.38	0.41	0.2
4	**0.59**	**0.38**	**0.32**	**0.14**	**0.37**	**0.4**	**0.09**
5	0.62	0.37	0.29	0.11	0.37	0.4	0.12
6	0.56	0.26	0.15	0.006	0.28	0.3	0.16
7	0.61	0.34	0.21	0.04	0.37	0.32	0.14
8	0.66	0.41	0.27	0.14	0.48	0.37	0.22

Порівняння кількості ітерацій за алгоритмами К-середніх та К-середніх ++ показано в таблиці нижче. Середній час К-середніх ++ склав 0,0056, а К-середніх – 0,0053. Для k = 4 як К-середніх, так і К-середніх++ виконували кластеризування в аналогічних часових рамках.

Таблиця 4.28

Час і порівняння ітерацій К-середніх і К-середніх++

K	Ітерації	Час

	К-середніх ++	К-середніх	К-середніх ++	К-середніх
2	3	12	0.0071	0.0048
3	3	9	0.0047	0.0066
4	**4**	**4**	**0.0047**	**0.0047**
5	6	4	0.0046	0.0046
6	5	6	0.0046	0.0063
7	4	3	0.008	0.0066
8	3	4	0.0047	0.0047
9	2	3	0.0048	0.005
10	3	5	0.0064	0.0047

Графік для порівняння ітерацій показаний на малюнку нижче.

Рис. 4.42 К-середніх та К-середніх++ Ітерації для бази даних дельфінів.

Детальний результат кластеризації за алгоритмом К-середніх++ для набору даних на k=4 наведено в таблиці нижче.

Таблиця 4.29

К-середніх ++ Результати бази даних «Дельфін».

Мітка спільноти	*Ідентифікатор вузла*
0	0, 1, 2, 3, 4, 7, 8, 10, 11, 12, 19, 20, 22, 25, 26, 27, 28, 30, 31, 32, 35, 36, 39, 42, 44, 46, 47, 48, 49, 53, 55, 58, 60, 61
1	5, 6, 9, 13, 17, 41, 54, 56, 57
2	14, 16, 33, 34, 37, 38, 40, 43, 52
3	15, 18, 21, 23, 24, 29, 45, 50, 51, 59

Зіставлення кластера для набору даних для К-середніх++ показано на рисунку нижче.

Рис. 4.43 К-середніх++ Кластеризаційна Карта бази даних дельфінів.(k=5)

4.3.3 Застосування методу К-середніх++ для набору даних «Електронна пошта»

Для пошуку оптимального значення k був розглянутий критерій ліктя. Для діапазону k від 1 до 100 графік для SSE показаний на малюнку нижче.

Рис. 4.44 К-середніх Аналіз ліктя «Електронна пошта»

Враховуючи значення SSE для К-середніх++ ('30681.44', '28057.21', '27079.87', '26247.93', '25540.68', '24835.12', '24234.69', '23866.06', '23826.29', '23346.10', '22764.11', '22641.56', ...,'15048.27', '15068.05', '14914.29', '14892.09', '14961.61', '14796.61', '14892.09', '14961.61', '14796.61', '1 4816.99', '14731.14', '14715.31', '14672.57', '14509.05'), найбільші падіння, після яких падіння стає більш плавним, призначені для значень k як 8.

Для перевірки значення k для мережі був виконаний аналіз силуету для діапазону значень k від 2 до 100. На рисунку 49 показано графік коефіцієнтів силуету щодо кількості кластерів, вибраних для застосування К-середніх++. Значення k було обрано на основі комбінованих показників силуету та інших показників, тобто модульності та Калінського та Харабаша.

Рис. 4.45 К-середніх++ Модуляція для «Електронна пошта»

Як К-середніх, так і К-середніх ++ мають найвищу модульність значення 0,26. Але оцінка досягається на рівні k=15 для К-середніх і k=9 для К-середніх++.

Рис. 4.46 К-середніх та К-середніх++ Шкала силуету «Електронна пошта»

Для оцінки силуету, як К-середніх, так і К-середніх++ розділили найвищу

оцінку силуету 0,34 при k = 2.

Для оцінки Калінського і Харабаша розглянемо сюжет на малюнку нижче. Як для К-середніх, так і для К-середніх++, оцінка була найвищою, тобто 93,8, при k= 2.

Рис. 4.47 Оцінки Калінського та Харабаша для набору даних

Наведена нижче таблиця показує порівняння оптимального значення k з урахуванням показників SSE та доброти спільноти.

Таблиця 4.30

Оптимальна k для К-середніх++ «Електронна пошта»

	SSE	Модульнисть	Сылует	Калінський і Харабаш
k	8	9	2	2

Загальна тенденція, що спостерігається в показниках доброти спільноти, полягає в тому, що оцінки зменшуються зі збільшенням k. Для всіх трьох показників найвищий бал дає найкращу структуру спільноти. Таким чином, виходячи з близькості значень k між SSE і модульністю, значення k було обрано як k = 9. У таблиці нижче наведено оцінки доброти спільноти для k=9.

Таблиця 4.31

К-середніх++ Якість спільноти для k=9 «Електронна пошта»

	Силует	Модульність	Калінський і Харабаш
Результати	0.15	0.26	35.8

Сім оцінок ефективності для вибраного значення k як 9 показано в таблиці нижче.

Таблиця 4.32

К-середніх Аналіз продуктивності для k=9 «Електронна пошта»

	Чистота	NMI	AMI	ARI	Однорідність	F1	FMI
Результати	0.29	0.33	0.18	0.04	0.22	0.015	0.2

Вивчення показників добра спільноти та відповідних балів алгоритму на k = 42 (від землі істини) покаже, чи є сформовані кластери ближче до істини чи ні. Наведена нижче таблиця показує оцінки добра спільноти та відповідні оцінки з наземною істиною на k= 42.

Таблиця 4.33

К-середніх++ Оцінки добробуту спільноти для k=42 «Електронна пошта»

	Силует	Модульність	Калінський і Харабаш
Результати	0.15	0.22	15.4

У таблиці нижче наведено бали за сім показників ефективності на k = 42.

Таблиця 4.33

Аналіз продуктивності К-середніх++ для k=42 «Електронна пошта»

	Чистота	NMI	AMI	ARI	Однорідність	F1	FMI
Результати	0.45	0.49	0.33	0.07	0.42	0.07	0.19

Усі показники продуктивності, крім FMI, показують покращення продуктивності, коли використовувався номер кластера з землі. Це показує, що кластери, утворені К-середніх при k=42, мали більшу схожість з наземною істиною, ніж при k=9. Це показує, що k-аналіз не ефективно виявив кількість кластерів, тобто k. По-друге, показники ефективності в цілому дуже низькі. Це означає, що К-Кошти недостатньо об'єднуються, щоб задовольнити вимоги наземних спільнот істин.

4.4 Тестове середовище

У Python було створено середовище тестування для аналізу продуктивності двох алгоритмів. Це сталося тому, що Python створив аналізатори, які можуть

читати різні формати набору даних. Потрібно надати тільки шлях до файлу, і Python сам створює граф. Крім того, як методи Лувен, так і К-середніх, а також майже всі показники оцінки, необхідні для проведення аналізу продуктивності двох алгоритмів, попередньо існують в Python (див. Додаток для деталей кожного методу). Дев'ять з десяти показників продуктивності мають реалізацію в бібліотеках Python. Тільки метрика чистоти була реалізована з нуля.

iGraph - це безкоштовна високопродуктивна бібліотека графів, розроблена в інтерфейсах C. Python з цією бібліотекою і полегшує складні мережеві дослідження та аналіз (Csardi, and Nepusz, 2005). Бібліотека полегшує процес створення мережевих вузлів, ребер, графіків, маркування і т.д. iGraph також має метод оцінки модульності утворених кластерів.

Реалізація алгоритму Лувена Blondel et al. (2010) доступна в пакеті iGraph. Використовувалися налаштування алгоритму Лувена за замовчуванням. Оскільки всі графіки набору даних були не переважливими, всім ребрам призначалися рівні ваги. Алгоритм був встановлений, щоб дати остаточне членство і модульність після виконання всіх ітерацій.

Scikit-learn - це інструмент машинного навчання та інтелектуального аналізу даних на Python. Інструмент має реалізації різних класифікацій, кластеризації, зменшення розмірності, попередньої обробки, регресії та алгоритмів відбору моделі (Педрегоса та ін., 2011). Для реалізації алгоритму К-середніх було використано інструмент KMeans з пакета Scikit-learn. Інструмент використовує метод К-середніх++ за замовчуванням, де гарантується, що початкові центроїди, хоча і генеруються випадковим чином, ніколи не будуть занадто близькими (Артур і Васильвіцький, 2007). Крім встановлення традиційної версії алгоритму К-середніх або К-середніх ++, для експериментів використовувалися налаштування алгоритму за замовчуванням. Це означає, що для кожного проходу алгоритму центроїди обчислюються десять разів, а алгоритм працює десять разів. Після цього в якості виходу алгоритму вибирається вихід, який має мінімальне значення інерції (SSE). Це означає, що час виконання, перелічений для К-середніх і К-середніх ++ проти кожного

набору даних, в десять разів перевищує час виконання.

Для кожного мережевого графа обчислюється матриця суміжності, а потім на цю матрицю застосовується фільтр k-середніх. Методи показників продуктивності також використовують цю матрицю суміжності для обчислення відповідних оцінок.

Всі випробування проводилися за допомогою macbook air. У системі використовується M1 chip 8 ядер. Система мала 16 ГБ оперативної пам'яті.

Висновки до розділу 4

У цьому розділі представлені результати тестування двох алгоритмів щодо трьох наборів даних різного розміру, які є загальнодоступними для дослідницьких цілей; Набір даних клубу карате Захарі, набір даних соціальних мереж «Дельфін» та великий набір даних мережевих комунікацій «Електронна пошта». Усі три набори даних мають спільноти, що не перекриваються, і загальнодоступні в Інтернеті. Розділ також містить оцінку версії алгоритму К-середніх++, щоб побачити, як покращення процесу вибору центроїдів покращує загальні результати алгоритму К-середніх.

5 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТІВ

5.1 Огляд результатів експериментів

Оцінки добробуту спільноти та оцінки ефективності Лувена, алгоритмів К-середніх та К-середніх++ для набору даних, відносно ground truth результатів, можна побачити в таблицях 41 та 42.

Таблиця 5.1

Результати експериментів

Набір даних	Метод	Модульність	Силует	Calinski and Harabasz	Кластери
Карате	Ground	0.37	0.16	7.83	2
	К-середніх	0.37	0.18	**7.83**	**2**
	К-середніх ++	0.37	**0.195**	**7.83**	**2**
	Лувен	**0.418**	0.146	5.54	4
Дельфін	Ground	0.519	0.117	6.392	4
	К-середніх	0.34	0.08	5.25	**4**
	К-середніх++	0.38	0.09	5.27	**4**
	Лувен	**0.518**	**0.108**	**5.75**	5
Електронна пошта	Ground	0.42	-0.197	7.129	42
	К-середніх	0.35	0.09	93.7	2
	К-середніх++	0.15	0.26	35.8	9
	Лувен	**0.42**	**-0.295**	**3.96**	**27**

З оцінки ground truth можна побачити, що для набору даних Клубу карате, як К-середніх, так і К-середніх++ не тільки дали правильні структури спільноти, їх оцінки також відповідали істиним оцінкам даних. Лувен, з іншого боку, дав дві додаткові спільноти. Модульність кластерів була навіть вище, ніж структура ground truth, але інші два заходи постраждали, і алгоритм не зміг правильно отримати кластери. Між К-середніх і К-середніх++, К-середніх++ дав кращу оцінку коефіцієнту силуету.

Результат для набору даних «Дельфін» показує, що знову ж таки, і К-середніх, і К-середніх++, алгоритми отримали правильний номер спільнот. Лувен виявив ще одну спільноту. Хоча була виявлена додаткова спільнота, але оцінки алгоритму Лувена не надто відхилялися ground truth. К-середніх++ знову дав кращі результати продуктивності в порівнянні з К-середніх для набору даних «Дельфін». Було невелике зниження в модульності міри.

Результат для набору даних «Електронна пошта» показує, що Лувен хоча і не зміг виявити всі спільноти, але його результати виявлення були ближче до дизайну спільноти істини в порівнянні з К-середніх. Хоча Лувен поділяв міру модульності з ground truth, але силует і Калінський і Харабаш оцінки були різними. Тим часом і К-середніх, і К-середніх++ не змогли переконливо показти себн. К-середніх виявив лише 2 спільноти, тоді як К-середніх++ виявив 9. Оскільки оцінка силуету є від'ємним значенням, це означає, що деякі вузли скоріше були б у різних кластерах.

Таблиця 5.2

Результати для вимірювань ефективності

Набір даних	Алгоритм	Чистота	NMI	AMI	ARI	F1	Однорідність	FMI
Карате	**К-середніх**	**1**	**0.71**	**0.49**	**0.52**	**1**	**1**	**0.71**
	К-середніх++	0.97	0.61	0.41	0.53	1	1	0.71
	Лувен	0.98	0.61	0.44	0.49	0.8	0.83	0.63
Дельфін	К-середніх	0.56	0.33	0.25	0.07	0.29	0.32	0.37
	К-середніх++	0.59	0.38	0.32	0.14	0.09	0.37	0.4
	Лувен	**0.887**	**0.73**	**0.64**	**0.64**	**0.03**	**0.79**	**0.74**
Електронна пошта	К-середніх	0.109	0.054	0.012	0.008	0.019	0.054	0.2
	К-середніх++	0.29	0.33	0.18	0.04	0.22	0.015	0.2
	Лувен	**0.393**	**0.55**	**0.337**	**0.25**	**0.417**	**0.04**	**0.38**

З заходів ефективності, отриманих для трьох наборів даних, можна

побачити, що між К-середніх і К-середніх++, К-середніх дав кращі відповідності громад до землі істини. Його оцінки були значно кращими, ніж у К-середніх++.

Для невеликого набору даних Клуб карате, як К-середніх, так і К-середніх++ перевершили Лувен. Для всіх семи кластеризійних відповідних заходів оцінки К-середніх і К-середніх++ були вищими, ніж Лувен. Отже, К-середніх і К-середніх++ не тільки сформували кращі спільноти, вони навіть добре збігалися з наземною правдою.

Але для бази даних «Дельфін» алгоритм Лувена дав кращі оцінки, ніж К-середніх, хоча кількість спільнот, визначених К-середніх, була такою ж, як і наземна істина, в той час як Лувен визначив додаткову спільноту. Всі відповідні заходи спільноти для Лувена дали кращі результати, ніж К-середніх і К-середніх++. Це показує, що, попри те, що К-середніх і К-середніх++ сформували кращі спільноти, Лувен сформував громади, які більше збігалися з наземною істиною. За винятком оцінки F1, К-середніх++ дав кращі результати, ніж К-середніх. Для набору даних «Електронна пошта», заснованого на оцінках всіх семи показників зіставлення кластеризації, алгоритм Лувена перевершив алгоритм К-середніх. Різниця між результатами була значно високою. І знову ж таки, за винятком F1 Оцінка, К-середніх++ дав кращі показники продуктивності, ніж К-середніх.

5.2 Порівняльна характеристика для набору даних «Карате-клуб»

Для набору даних «Електронна пошта», заснованого на оцінках всіх семи показників зіставлення кластеризації, алгоритм Лувена перевершив алгоритм К-середніх. Різниця між результатами була значно високою. І знову ж таки, за винятком F1 Оцінка, К-середніх++ дав кращі показники продуктивності, ніж К-середніх.

Рис. 5.1 Показники оцінки спільноти для клубу карате

Рис. 5.2 Показники оцінки ефективності для набору даних клубу карате

5.3 Порівняльна характеристика для набору даних «Дельфіни»

Високий рівень перегляду результатів алгоритмів наборів даних «Дельфін» представлений на рисунках нижче. Можна побачити, що К-середніх++ не тільки сформував кластери кращої якості, ніж К-середніх, але і показав кращі результати для всіх показників продуктивності. Лувен знайшов додаткову спільноту, але крім цього, вона не тільки сформувала спільноти кращої якості, але і наблизилася до землі в порівнянні з К-середніх++.

Рис. 5.3 Показники оцінки якості для набору даних «*Дельфін»*

Рис. 5.4 Показники оцінки продуктивності для набору даних «Дельфін»

5.4 Порівняльна характеристика для набору даних «Електронна пошта»

Високий рівень перегляду порівнянь оцінки добра спільноти показаний на рисунках нижче. Як видно з графіків, К-середніх++ працює краще, ніж К-середніх. Лувен показав себе краще, ніж К-середніх++, але він не працював занадто добре в порівнянні з наземною правдою.

Рис. 5.5 Показники якості для мережі «Електронна пошта»

Рис. 5.6 Показниеи ефективності для мережі «Електронна пошта»

Виходячи з ефективності двох алгоритмів відповідно до двох критеріїв оцінки, тобто обґрунтованих спільнот істини та показників істини спільноти, наведена нижче таблиця рейтингу на основі порівняння алгоритмів для трьох наборів даних.

Таблиця 5.3

Підсумок ефективності Лувена проти К-середніх

Evaluation Criterion	Карате-клуб	Дельфін	Електронна пошта
Community Goodness	К-середніх++ К-середніх	К-середніх ++ К-середніх	Лувен
Matching with Ground Truth	К-середніх К-середніх++	Лувен	Лувен

Висновки до розділу 5

У цьому розділі представлені результати тестування двох алгоритмів щодо трьох наборів даних різного розміру, які є загальнодоступними для дослідницьких цілей; Набір даних клубу карате Захарі, набір даних соціальних мереж «Дельфін» та великий набір даних мережевих комунікацій «Електронна пошта». Усі три набори даних мають спільноти, що не перекриваються, і загальнодоступні в Інтернеті. Розділ також містить оцінку версії алгоритму К-середніх++, щоб побачити, як покращення процесу вибору центроїдів покращує загальні результати алгоритму К-середніх.

ВИСНОВКИ

Ця робота є внеском у існуючу літературу порівнянь алгоритмів виявлення спільноти. Два популярних алгоритми кластеризації, Лувена і К-середніх, були порівняні з трьома реальними наборами даних різної складності для порівняння їх результатів. Підхід, обраний для цієї роботи, поляв в тому, щоб почати оцінку з невеликої реальної системи, а потім спостерігати за поведінкою алгоритму, оскільки тестові мережі ставали більшими і складними.

У найменшому наборі даних Клубу карате алгоритм К-середніх показав кращі результати, ніж Лувен, тоді як К-середніх++ показав кращі результати, ніж К-середніх. Для відносно більшого набору даних «Дельфін», К-середніх++ дав кращі результати, ніж К-середніх. Ці два алгоритми сформували кращі спільноти, ніж Лувен. Однак, у порівнянні з основною правдою, Лувен дав кращу продуктивність, ніж обидва. І, нарешті, у великому складному мережевому наборі даних «Електронна пошта» Лувен перевершив К-середніх як за формуванням спільноти, так і за збігом з основною істиною.

З порівняльних тестів для всіх наборів даних К-середніх++ показав кращі результати порівнюючи зі звичайним К-середніх. Ця непередбачуваність у поведінці обумовлена випадковою ініціалізацією центроїдів.

Непослідовну продуктивність двох алгоритмів можна віднести до різниці у властивостях трьох наборів даних. Набори даних змінювалися за розміром та структурою їх інтернотних відносин. Продуктивність К-середніх погіршилася зі збільшенням розміру мережі. Спостережуваною причиною невдачі К-середніх були неправильні значення, досягнуті під час кроку аналізу критерію ліктя. Метрики, які вважаються такими, що представляють хорошу спільноту, неправильно оцінили результати.

У даній роботі при виконанні К-середніх використовувалася матриця суміжностт розміру nxn, де n - кількість вузлів. Представлення було ефективним, оскільки воно правильно кластеризувало менші набори даних, де в даних не було шаблонів, що перекриваються. Для більшого набору даних, електронної пошти Eu-Core, шаблон даних (наприклад, область перетину між кластерами) в мережі

була такою, що К-середніх сам по собі не підходить для конкретної мережі. К-середніх виявився найбільш ефективним у випадках, коли в даних немає шаблонів, що перекриваються.

Представлення даних через матрицю було однаково ефективним у представленні вузлів усіх трьох наборів даних. Для меншого набору даних «Карате-Клуб» і «Дельфін» К-середніх дав більш швидкі результати, незважаючи на обчислення nxn в кожній ітерації. Але для більшого набору «Електронна пошта» Лувен перевершив К-середніх у всіх пробігах. Це показує, як використання матриці для великих наборів даних не є практичним підходом.

На основі порівняльного аналізу двох популярних алгоритмів на наборах даних різного розміру, робота доводить:

1.К-середніх і Лувен поодинці не відображають суть мережі.

2.Деякі топологічні характеристики також повинні бути враховані при використанні цих алгоритмів кластеризації. Модульность (Лувен) і SSE (К-середніх) недостатньо.

3.Ефективність алгоритму не може бути визначена простою близькістю до істини.

4.К-середніх++ — покращена версія алгоритму К-середніх як щодо швидкості, так і продуктивності.

5.На продуктивність алгоритму К-середніх сильно впливають початкові положення центроїдів. У цьому сенсі алгоритм Лувена більш надійний.

6.Рішення для здійснення початкового відбору центроїдів і кількості кластерів k повинні бути автоматизовані і засновані на топологічних особливостях мережі. Існуючий критерій ліктя і силуету може пропустити з'ясування оптимального k.

7.Ефективні алгоритми – це ті, які залишаються близькими до структури мережі, а також до істини.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Артур Д. та Васильвіцький С. К-середніх++: Матеріали вісімнадцятого щорічного симпозіуму ACM-SIAM з дискретних алгоритмів. 2007. стор. 1027-1035. URL: http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

2. Блондел, В. Д., Ламбіот, Р. Швидке розгортання спільнот у великих мережах. Журнал статистичної механіки P10008. 2008.

3. Вагнер С. та Вагнер Д. Порівняння кластеризації - огляд. Технічний звіт 2006-04. 2007. URL: https://i11www.iti.kit.edu/extra/publications/ww-cco-06.pdf

4. Вайсштайн, Є. В. . Матриця суміжності. 2018. URL: http://mathworld.wolfram.com/AdjacencyMatrix.html

5. Ван Дж, Лі М, Ден Ю, Пан Ю. Останні досягнення в методах кластеризації мереж взаємодії білків. 2010. BMC Genomics. DOI: doi:10.1186/1471-2164-11-S3-S10

6. Блондель В.Д., Ж.-Л. Гійом, Р. Ламбьот та Е. Лефевр, «Швидке розгортання спільнот у великих мережах», J. Stat. Хутро. (2008) P10008, стор 12, 2008.

7. Ван С. і Купман Р. «Кластеризація статей на основі семантичної подібності». Журнал Scientometrics, 2017. Том 111, випуск 2, стор. 1017-1031. DOI: https://doi.org/10.1007/s11192-017-2298-x

8. Венкатесарамані Р., Воробейчик Ю. Виявлення спільноти за допомогою моделювання інформаційного потоку. 2018. CoRR, abs/1805.04920.

9. Гірван, М. та Ньюман М. Е. Структура спільноти в соціальних і біологічних мережах. Proc Natl Acad Sci U S A. 2002. Том 99, випуск 12, стор. 7821-6.

10. Гуд, Б.Х., Ів, А.М. та Аарон К. Виконання максимізації модульності в практичних контекстах. Фізичний огляд. Е, Статистична, нелінійна та фізика м'якої матерії. 2010. Том 81, стор. 046106.

11. Дестерке, С. . Функції переконань: теорія та застосування. 2018. Стор. 263.

12. Джебаблі М., Черіфі Х., Черіфі К. і Хамуда А. Виявлення спільноти,

що перекривається, проти Ground-Truth у мережі спільної закупівлі AMAZON. Матеріали 11-ї Міжнародної конференції IEEE SITIS, Семінар комплексних мереж та їх застосування. 2015. стор. 328 - 336.

13. Джеймс Б., Роберт Е. та Вільям Ф. . Алгоритм кластеризації нечітких К-середніх, Комп'ютери та геологія. 1984. Том 10, випуск 2-3, стор. 191-203.

14. Джіянті, С.К. і Прія, Ч.К. . Кластеризаційний підхід до класифікації дослідницьких статей на основі пошуку за ключовими словами. Міжнародний журнал передових досліджень у галузі комп'ютерної інженерії та технологій (IJARCET). 2018. Том 7, випуск 1, ISSN:2278–1323

15. Елкан К. Використання нерівності трикутника для прискорення k-середніх. На Міжнародній конференції з машинного навчання. 2003. Стор. 147–153.

16. Зоммер Ф., Фоус Ф., Саренс М. K-засоби ядра на основі модульності для виявлення спільноти. 26-а Міжнародна конференція з штучних нейронних мереж, Конспект лекцій з інформатики. 2017. стор. 423-433.

17. Каваджі Х., Ямагучі Ю., Мацуда Х. і Хашимото А. Метод кластеризації на основі графів для великого набору послідовностей з використанням алгоритму розбиття графа. Геномна інформатика. Міжнародна конференція з геномної інформатики. 2001. Том 12, стор. 93-102.

18. Калінський, Т. і Харабаш, Дж. Дендритний метод для кластерного аналізу. Комунікації в статистиці - теорія і методи. 1974. Том 3. Стр. 1-27.

19. Кауфман Л. і Руссю П. Дж. Пошук груп у даних – Вступ до кластерного аналізу. 1990.

20. Кім Лім Юн «Два застосування методів кластеризації в Twitter: виявлення спільноти та вилучення проблем». Дискретна динаміка в природі і суспільстві, вип. 2013, ідентифікатор статті 903765, 8 сторінок, 2013. URL: https://www.hindawi.com/journals/ddns/2013/903765

21. Кім, Кьон Чже та Ан, Хенчуль. Система рекомендацій з використанням GA K-означає кластеризацію в інтернет-магазині. Експертні системи з додатками. 2008. Том 34, видання 2, стор. 1200-1209.

22. Ковальчик, Р. Обчислювальний колективний інтелект. Семантична мережа, соціальні мережі та мультиагентні системи: перша міжнародна конференція, ICCCI. 2009. Стор. 198-199.

23. Ландман Н., Панг Х., Вільямс К. і Росс Е. Кластеризація k-середніх. 2018. URL: https://brilliant.org/wiki/k-means-clustering

24. Лі К. та Каннінгем П. Виявлення спільноти: ефективне оцінювання у великих соціальних мережах, Журнал складних мереж. 2014. Том 2, випуск 1, стор. 19–37. https://doi.org/10.1093/comnet/cnt012

25. Лойд, С.П. Квантування за найменшими квадратами в пкм. Теорія інформації, IEEE Trans. 1982. Том 28, випуск 2, с.129–137.

26. Паттанайк В., Сінгх М., Гупта П., Сінгх С. К. Розумна оцінка заторів у режимі реального часу та техніка кластеризації для міських автомобільних доріг. 2016. 3420-3423. 10.1109/TENCON.2016.7848689.

27. Педрегоза Ф., Вароко Г., Грамфорт А., Мішель В., Тіріон Б., Грізель О. Машинне навчання на Python. Журнал досліджень машинного навчання. 2011. Стор. 2825−2830.

28. Райхардт, Дж., і Борнхольдт, С. Статистична механіка виявлення спільнот. Фізичний огляд. E, Статистична, нелінійна та фізика м'якої матерії. 2006. Том 74. 016110.

29. Розенберг А. і Хіршберг Дж. Умовна ентропійна оцінка зовнішнього кластера. 2007. URL: http://aclweb.org/anthology/D/D07/D07-1043.pdf

30. Романо, С., Він, Н. Х., Бейлі, Дж. і Верспур, К. Коригування для показників порівняння кластеризації шансів. 2016. Журнал досліджень машинного навчання 17. Стор. 1-32. URL: http://jmlr.csail.mit.edu/papers/volume17/15-627/15-627

31. Сю Р. і Вунш Д. Кластеризація. Джон Вайлі і сини. 2009. стор. 32

32. Тан Д. Скоригований індекс рандів. 2017. URL: https://davetang.org/muse/2017/09/21/adjusted-rand-index/

33. Фахім, А.М., та ін. Ефективний розширений алгоритм кластеризації k-середніх. Журнал Чжецзянського університету «Наука» A 7. 2006. Стор. 1626-

1633 pp.

34. Фокс, Е. Б. & Меловс, К.Л. Метод порівняння двох ієрархічних кластеризацій. Журнал Американської статистичної асоціації. 1983. Том 78, випуск 383, стор. 553-569. URL: http://wildfire.stat.ucla.edu/pdflibrary/fowlkes.pdf

35. Фортунато, С. і Бартелемі, М. Межа роздільної здатності при виявленні спільноти. Праці Національної академії наук. 2007. Том 104, випуск 1, стор. 36–41.

36. Франті П. та Сіераноя С. К-значення властивості шести контрольних наборів даних кластеризації. 2018. Том 48, випуск 12, стор. 4743-4759. URL: http://cs.joensuu.fi/sipu/datasets/

37. Хан Дж., Пей Дж. та Камбер М. Дата майнінг: концепції та прийоми. 2011. стор. 487-488.

38. Хріч Д., Дарст Р.К. & Фортунато, С. Виявлення спільноти в мережах: структурні спільноти проти реальної правди. 2014. Том 90, випуск 6. URL: https://arxiv.org/pdf/1406.0146.pdf

39. Цанаку, Є.М. Розпізнавання шаблонів під наглядом і без нагляду: вилучення ознак і обчислювальний інтелект. 2017. CRC Видання. Стор. 203.

40. Цзяньцзюнь, Ч., Мінвей, Л. Лунцзе, Л., Ханьхай, З. та Сяоюнь, Ч. Активне виявлення спільноти з напівнаглядом на основі обмежень обов'язкового і неможливого зв'язування. 2014. PloS один. 9. e110088. 10.1371/journal.pone.0110088.

41. Чакраборті, Т., Далмія, А., Мукерджі, А. та Гангулі, Н. Показники для аналізу спільноти: опитування. Журнал ACM Комп'ютерні системи (CSUR). 2017. Том 50 Випуск 4, стор. 54.

42. Чарді, Габор і Непуш, Тамаш. Програмний пакет Igraph для комплексних мережевих досліджень. InterJournal. Складні системи. 2005.

43. Чжан Дж., Чжу К., Пей Ю., Флетчер Г. та Печенізький М. Висновок про приналежність до кластеризації із зразків графіка. 2018. MLG@KDD'18. ISBN ACM 123-4567-24-567/08/06. URL: http://www.mlgworkshop.org/2018/papers/MLG2018_paper_37.pdf

44. Шапіро Л.; Стокман Г. Комп'ютерний зір. Прентіс Хол. 2002. Стор. 69–73.

45. Ши Дж. та Малік Дж. Нормовані вирізи та сегментація зображення. Транзакції IEEE щодо аналізу шаблонів та машинного інтелекту. 2000. Том 22, випуск 8, стор. 888–905.