Технічні науки

УДК 004.852

Магас Валентин Васильович
студент
Національний технічний університет України
«Київський політехнічний інститут»
Магас Валентин Васильевич
студент
Национальный технический университет Украины
«Киевский политехнический институт »
Mahas V.
student
National Technical University of Ukraine “Kyiv Polytechnic Institute”

КЛАСТЕРНИЙ АНАЛІЗ ПРОСТОРОВИХ БАЗ ДАНИХ ОНКОХВОРИХ
КЛАСТЕРНЫЙ АНАЛИЗ ПРОСТРАНСТВЕННЫХ БАЗ ДАННЫХ ОНКОБОЛЬНЫХ
CLUSTER ANALYSIS OF SPATIAL CANCER DATABASES

Анотація: Метою даної роботи є застосування засобів інтелектуального аналізу даних для виявлення прихованих закономірностей та зв’язків у просторових баз даних онкохворих пацієнтів. Дана робота зосереджується як на дискретних так і неперервних просторових медичних базах даних, до яких застосовуються кращі методи кластеризації для визначення максимально корисних кластерів.

Ключові слова: k-means, інтелектуальний аналіз просторових даних, алгоритми кластеризації, дендограма, просторовий розподіл.

Анотация: Целью данной работы является применение средств интеллектуального анализа данных для выявления скрытых закономерностей и связей в пространственных базах данных онкобольных пациентов.

Ключевые слова: k-means, интеллектуальный анализ пространственных данных, алгоритмы кластеризации, дендограма, пространственное распределение.

Summary: The aim of this research is the application of data mining for detecting hidden patterns and relationships in spatial cancer databases. This study focuses on discrete and continuous spatial medical databases on which clustering techniques are applied and the efficient clusters were formed.

Key words: k-means, spatial data mining, clustering algorithm, dendogram, spatial distribution.

Вступ

Сучасний період розвитку суспільства характеризується значним впливом на нього інформаційних технологій. Не стала винятком і медицина, яка сьогодні набула абсолютно нових рис. Поміж всієї ширини застосування новітніх технологій та засобів у даній сфері, чільне місце посідає інтелектуальний аналіз даних.

Інтелектуальний аналіз даних - збірна назва, що використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності.

У зв’язку бурхливим накопиченням медичних даних, особливо гостро постає проблема застосування аналізу даних для їх опрацювання та подальшого прогнозування, на їх основі. В свою чергу необхідність визначення прихованих тенденцій поширення певного захворювання на тій чи іншій території, виявлення неочевидних зв’язків між факторами що його спричиняють та подальшого прогнозування можливого розвитку подій з кожним роком набуває все чіткіших рис. Дуже актуальним виглядає дослідження просторової бази даних онкохворих, адже дана хвороба відзначається значною смертністю серед хворих, як результат, кожного року помирає понад 15 мільйонів людей у цілому світі.

Постановка задачі

Основні задачі, що були поставлені в рамках даної статті — підбір алгоримтів та програмних інструментів для аналізу бази просторових медичних даних, а також порівняння отриманих результатів для різних алгоритмів, як демонстрація можливостей застосування просторового аналізу даних в областях “e-Health” та “e-Medicine”, що нині активно розвиваються.

Вибір алгоритмів для аналізу

Кластеризація (або кластерний аналіз) - це задача розбиття множини об'єктів на групи, які називаються кластерами. Усередині кожної групи повинні виявитися «схожі» об'єкти, а об'єкти різних групи повинні бути якомога більш відмінні. Головна відмінність кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий і визначається в процесі роботи алгоритму.

Кластер має наступні математичні характеристики: центр, радіус, середньоквадратичне відхилення, розмір кластера.

Алгоритм k-means

Найбільш поширений серед ітеративних методів алгоритм k- середніх, також званий швидким кластерним аналізом. На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо числа кластерів, для можливості використання цього методу необхідно мати гіпотезу про найбільш ймовірну кількість кластерів. Алгоритм k-середніх будує k кластерів, розташованих на великих відстанях один від одного. Основний тип задач, які вирішує алгоритм k-середніх, - наявність припущень (гіпотез) щодо числа кластерів, при цьому вони повинні бути різні настільки, наскільки це можливо. Вибір числа k може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції.

$$\sum_{i=1}^{k} \sum_{k \in D_i} \|x_i - c_i\|^2$$

де D_i- набір векторів що належать до і-го кластеру, а c_i - середнє значення цих векторів

$$c_i = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, x_k \epsilon D_i$$

Основна ідея полягає в тому, що на кожній ітерації заново вираховується центр мас, для кожного кластера, потім вектори розбиваються на нові класи, відповідно до того який з отриманих центрів виявився ближчим за метрикою

Алгоритм BIRCH

Завдяки узагальненому вигляду кластерів, швидкість кластеризації збільшується, алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) при цьому володіє великим масштабуванням.

У цьому алгоритмі реалізований двоетапний процес кластеризації.

В ході першого етапу формується попередній набір кластерів. На другому етапі до виявлених кластерів застосовуються інші алгоритми кластеризації - придатні для роботи в оперативній пам'яті.

Аналогія, що описує цей алгоритм: якщо кожен елемент даних уявити собі як намистину, що лежить на поверхні столу, то кластери намистин можна "замінити" тенісними кульками і перейти до більш детального вивчення кластерів тенісних кульок. Число намистин може виявитися досить велике, проте діаметр тенісних кульок можна підібрати таким чином, щоб на другому етапі можна було, застосувавши традиційні алгоритми кластеризації, визначити дійсну складну форму кластерів.

Алгоритм DBSCAN

Алгоритм DBSCAN - один з перших алгоритмів кластеризації щільнісним методом. В основі цього алгоритму лежить кілька визначень:

- ε-околицею об'єкта називається околиця радіуса ε деякого об'єкту.
- Кореневим об'єктом називається об'єкт, ε-околиця якого містить не менше деякого мінімального числа MinPts об'єктів.
- Об'єкт p безпосередньо щільно-досяжний з об'єкта q якщо p знаходиться в ε-околиці q і q є кореневим об'єктом.

- Об'єкт p щільно-досяжний з об'єкта q при заданих ε і MinPts, якщо існує послідовність об'єктів p1, ..., pn, де p1 = q і pn = p, така що pi + 1 безпосередньо щільно досяжний з pi, $1 \leq i \leq n$.
- Об'єкт p щільно-з'єднаний з об'єктом q при заданих ε і MinPts, якщо існує об'єкт o такий, що p і q щільно-досяжні з o.

Для пошуку кластерів алгоритм DBSCAN перевіряє ε-околиця кожного об'єкта. Якщо ε-околиця об'єкта p містить більше точок ніж MinPts, то створюється новий кластер з кореневим об'єктом p. Потім DBSCAN ітеративно збирає об'єкти безпосередньо щільно-досяжні з кореневих об'єктів, які можуть привести до об'єднання кількох щільно-досяжних кластерів. Процес завершується, коли ні до одного кластеру не може бути додано жодного нового об'єкта. Хоча, на відміну від методів розбиття, DBSCAN не вимагає заздалегідь вказувати число одержуваних кластерів, виникне потреба у вказівках значень параметрів ε і MinPts, які безпосередньо впливають на результат кластеризації. Оптимальні значення цих параметрів складно визначити, особливо для багатовимірних просторів даних.

Отже, з цілого різноманіття методів та алгоритмів кластеризації для експериментальних досліджень було відібрано три . З ітеративних алгоритмів — k-means, з ієрархічних зупинимося на алгоритмі BIRCH, а з щільнісних оберемо DBSCAN. Такий підбір дозволить нам максимально якісно опрацювати дані та провести їх аналіз.

Дані для експерименту

В якості просторової бази онкохворих пацієнтів використаємо SEER Database , яка визначається значною кількістю записів, що являється ключовим фактором для успішного аналізу та подальшого прогнозування а його основі, а саме у ній мітяться дані про онкохворих жителів США, що збиралися протягом понад 40 років. Дані були поділені на дві групи просторові і непросторові.

Набір просторових даних містить координати розташування, зображення дистанційного зондування та іншу географічну інформацію. До основних непросторових даних відносились: стать, вік, сімейний стан, висота, вага.

Досліджувались основні різновиди захворювання такі як: рак легень, рак нирок, рак горла і т.д.

Основний інтерес у вирішенні даної задачі полягає у безпосередньому застосування методів ІАПД у медицині. Таким чином, основний акцент буде зроблено на географічну прив'язку пацієнтів, їх розташування та закономірності розподілу по території.

Реалізація та результати

Для реалізації алгоритмів кластеризації були застосовано мову програмування Python з цілою низкою доповнень та бібліотек, зокрема : Pandas, Numpy, SciPy, Scikit-learn, Matplotlib. Перші три бібліотеки застосовувалися переважно для зручного представлення вхідних даних. Scikit-learn став нам у нагоді для безпосереднього аналізу даних. Остання з бібліотек стала незамінною для візуалізації результатів, а саме, при побудові діаграм та графіків. Результати роботи кожного алгоритмів наведено на рисунках 1-3, осями яких виступають географічна довгота і широта.

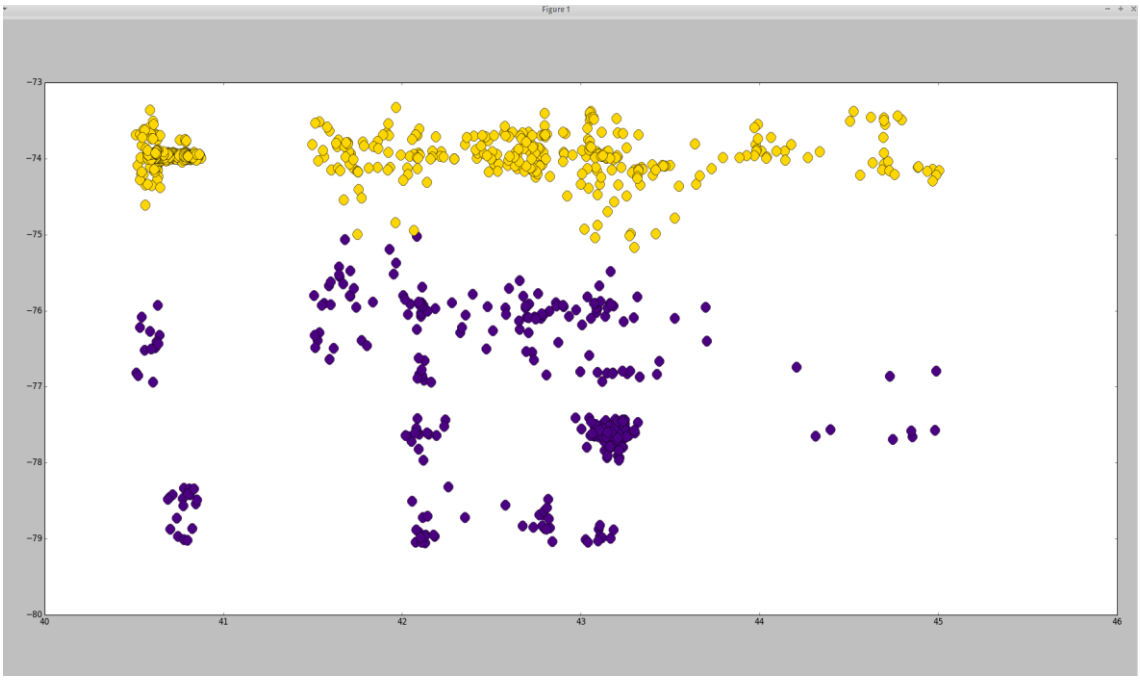

Рисунок 1.— Результати кластеризації методом k-means

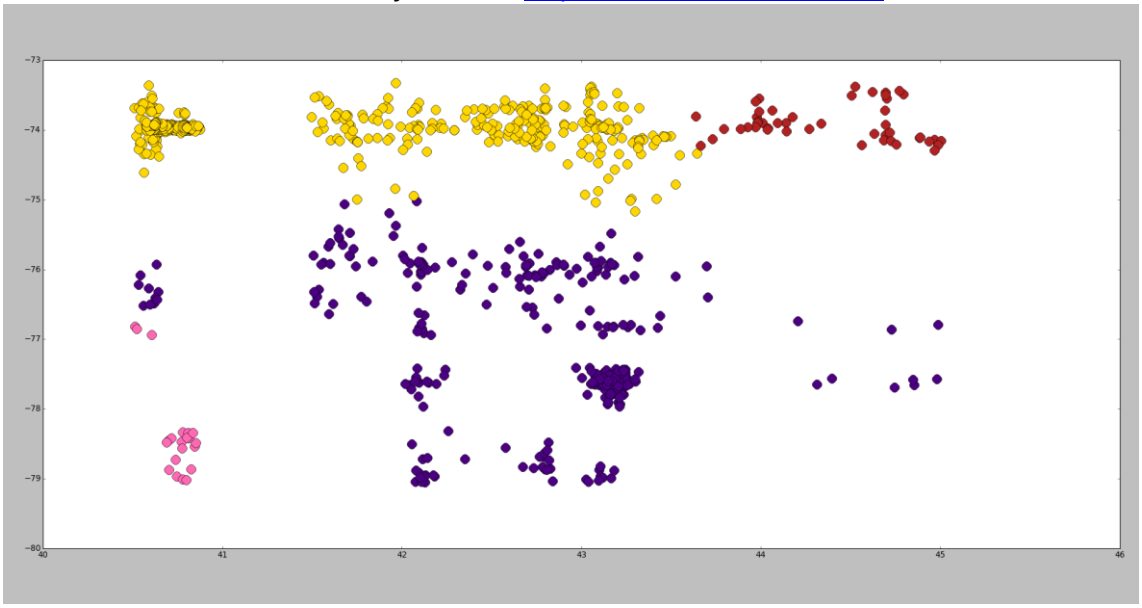

Рисунок 2. — Результати кластеризації методом BIRCH

Як можна бачити на малюнку вище , методом k-means весь набір даних був поділений на 2 кластери. У свою чергу, методом DBSCAN було створено 3 кластери. Алгоритм ієрархічної кластеризації виділив 4 кластери на тому ж наборі вхідних даних.

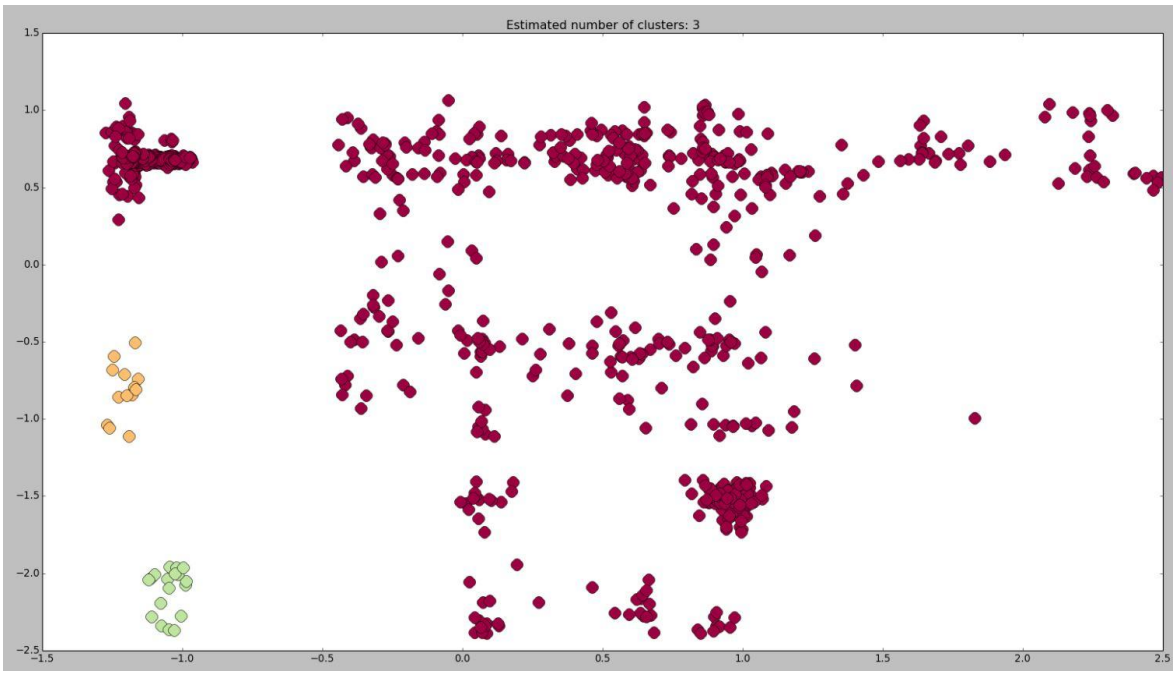

Рисунок 3. — Результати кластеризації методом DBSCAN

Для оцінки роботи алгоритмів застосуємо наступні критерії порівняння:

- **Однорідність**(homogeneity) – характеристика, що показує в якій мірі кожен кластер містить члени єдиного класу. Визначається за наступною формулою.

- **Повнота**(completeness) - характеристика, що показує в якій мірі всі члени даного класу присвоєні одному кластеру.
- **V-міра** (V-Measure) – міра заснована на ентропії, яка в явному вигляді визначає наскільки критерії однорідності та повноти були задоволені. Дана характеристика обчислюється, як середнє гармонійне різниці однорідності й повноти. Дана
- **Скоригований індекс** (Adjusted Rand Index) – міра подібності між двома кластерами даних. З математичної точки зору скоригований індекс пов'язаний с точністю і має місце навіть тоді, коли мітки класу не використовуються.
- **Взаємна інформація** (Mutal information) - це функція, що вимірює узгодженість кількох варіантів ігноруючи перестановки. Тобто, скоригована міра взаємної інформації тісно пов’язана зі зміною даних в процесі кластеризації.
- **Коефіцієнт силуету**(Silhouette Coefficient) – міра створена для оцінки моделі з чіткіше визначеним набором кластерів.

Результати порівняння алгоритмів зведем у таблиці 1.

Характеристика	K-means	DBSCAN	BIRCH
Однорідність	0.917	0.953	0.960
Повнота	0.766	0.883	0.812
V-міра	0.881	0.917	0.855
Скоригований індекс	0.966	0.952	0.914
Скоригована взаємна інформація	0.855	0.883	0.896
Коефіцієнт силуету	0.722	0.626	0.691

Таблиця 1. — Результати порівняння алгоритмів

Отже, проаналізувавши отримані результати роботи(таблиця 1) алгоритмів можна сказати, що щільнісний алгоритм DBSCAN проявив себе на більшості дослідів дещо краще ніж ітеративний та ієрархічний.

В якості результатів наведемо і карту поширеності онкозахворювань серед населення США(Рисунок 4), щоб була змога краще з'ясувати закономірності розподілу.

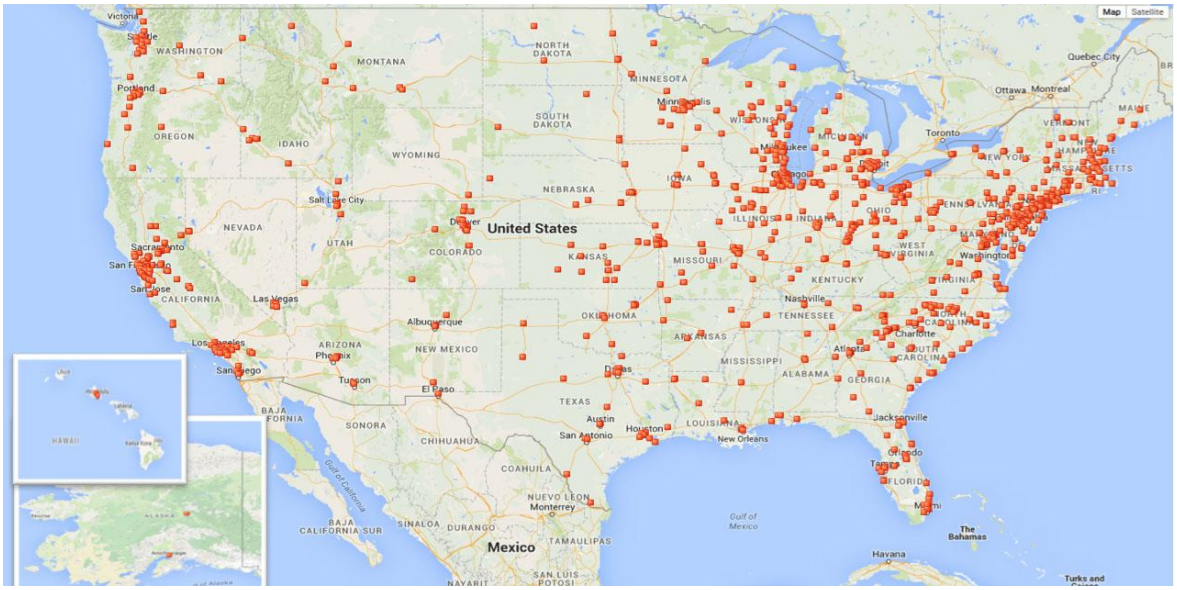

Рисунок 4. — Розподіл онкохворих по території США

Висновок

Завдяки інтелектуальному аналізу просторових даних виникла можливість пошуку взаємозв'язків та особливостей у великих наборах просторових даних і моделювання поведінки відповідних об'єктів дослідження.

В даній статті було показано можливість застосування засобів кластерного аналізу для досліджень бази просторових медичних даних ідсумовуючи зроблену роботу, зазначимо, що засоби кластерного аналізу чудово проявили себе для аналізу просторової бази медичних даних. Результати роботи кожного з алгоритмів були наведені у вигляді діаграм. Також була отримана карта розподілу онкохворих людей по території США. Результат порівняння алгоритмів кластеризації засвідчив, що обрані алгоритми добре зарекомендували себе на вхідному наборі даних. Окремо варто виділити алгоритм DBSCAN, який проявив себе дещо краще у вирішенні поставленої задачі.

Застосування як алгоритмів, так і програмних засобів аналізу просторових даних у сфері медицини представляє широке поле для подальших досліджень та вдосконалень..

Література

1. *Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining./ Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. – СПб.: БХВ – Петербург, 2008. — 13-22 с.*

2. *Дубров А. М. Многомерные статистические методы: Учебник. / Дубров А. М., Мхитарян В. С., Трошин Л. И. – К.: Финансы и статистика, 2000. — 8-16с.*

3. *Андрейчиков А.В. Интеллектуальные информационные системы. / Андрейчиков А.В., Андрейчикова О.Н. – К.: ФиС, 2004. — 22-45с.*

4. *Мандель И. Д. Кластерный анализ. / Мандель И. Д. — К.: Финансы и статистика, 1988. — 10 с.*

5. *Жамбю М. Иерархический кластер-анализ и соответствия. / Жамбю М. — К.: Финансы и статистика, 1988. — 345 с.*

6. *Дюран Б. Кластерный анализ. / Дюран Б., Оделл П. — К.: Статистика, 1999. — 128-84с.*