УДК 004.62:528 DOI: https://doi.org/10.30837/IYF.IIS.2024.107

КЛАСТЕРИЗАЦІЯ ТА ДИСКРЕТИЗАЦІЯ ГЕОДАНИХ

Котелевець К. А.
Науковий керівник – доц. Чала Л. Е.
Харківський національний університет радіоелектроніки, каф. ШІ,
м. Харків, Україна
e-mail: kyrylo.kotelevets@nure.ua

This work is devoted to geodata clustering using discretization and its practical relevance in today's geographic information environment. The importance of geodata clustering for identifying patterns and trends in geographic data is analyzed, and the problems associated with non-discrete geodata, such as large data volume and the presence of noise, are highlighted. It is experimentally shown on real GPS data using clustering methods K-Means, DBSCAN and OPTICS that the use of discretization techniques can effectively address these problems by reducing data volume, removing noise and improving the quality of analysis. The advantages of using geodata discretization for clustering are highlighted and the importance of further research and development of these methods to expand their applications in various fields is emphasized.

У сучасному світі геодані стали невід'ємним аналітичним ресурсом для різних галузей, що включають картографію, маркетинг, міське планування та багато інших. Популярність геоданих посилюється завдяки розвитку сучасних технологій, що сприяють збору, обробці та візуалізації великих обсягів даних. Метою роботи є аналіз існуючих проблем кластеризації даних та способів їх вирішення з використанням дискретизації, що дозволить покращити якість кластеризації та сприятиме подальшому розвитку цього напряму аналізу геоданих.

Одним з основних аналітичних інструментів аналізу геоданих є кластеризація – метод групування географічних об'єктів за схожими просторовими характеристиками в окремі групи, де об’єкти всередині кожної групи максимально подібні один до одного та мінімально подібні до об'єктів з інших груп. Основною метою кластеризації є виявлення прихованих патернів і структур в геоданих. Такий підхід широко застосовується у різних сферах, наприклад, у транспорті для оптимізації маршрутів транспортних засобів та аналізу трафіку з метою зменшення заторів.

Проблема, яка виникає під час проведення кластеризації геоданих, полягає у тому, що такі дані зазвичай мають неперервний характер. Незважаючи на те, що для більшості сфер використання геоданих достатньо обмежитись лише 6-ма знаками після коми для зберігання і роботи з довготою та широтою, вже така відносно невелика точність породжує майже 65 квадрильйонів унікальних пар координат. Це може

створювати складнощі у визначенні схожості між об'єктами та подальшому групуванні їх у кластери, оскільки методи кластеризації зазвичай вимагають визначення відстані у якості міри схожості між об'єктами [1].

Проте велика варіативність геоданих зовсім не завжди корелює з їх інформативністю, оскільки вони можуть містити шум або непередбачувані аномалії через технічні особливості їх збору. Такою особливістю може слугувати наявність похибки локації близько 2-3 метрів для систем GPS або GLONASS. Подібні похибки ускладнюють процес кластеризації та можуть призвести до виникнення неточностей у формуванні кластерів, що впливає на якість результатів та їхню подальшу інтерпретацію [2].

Дискретизація представляє собою процес перетворення неперервної величини, що може мати нескінченну кількість значень в момент часу або простору, на дискретну, що складається зі зліченної кількості окремих значень, відомих як вибірка. У контексті геоданих дискретизація означає перетворення неперервного простору геоданих на дискретний формат шляхом розділення їх на окремі одиниці, якими можуть виступати сітки чи області. Впровадження такого підходу може компенсувати негативний вплив перелічених проблем на якість результатів кластеризації.

Існує кілька методів дискретизації геоданих, що можуть бути використані в залежності від конкретних потреб та характеристик даних, наприклад, методи сегментування та обрізки, дозволяють розділити географічний простір на сегменти або області з урахуванням певних критеріїв, меж чи особливостей даних. Також подібним методом є використання сіток, де географічна область розділяється на певну кількість трикутних, квадратних або шестикутних сегментів.

Одними з найбільш відомих прикладів сіток для дискретизації є BingTiles та H3 – ієрархічні системи географічної індексації, що використовують квадратні та гексагональні сітки з різними рівнями деталізації відповідно. Наведені системи надають кілька десятків рівнів деталізації з середньою довжиною сторони сегменту від одного метру до сотень кілометрів на найменшому та набільшому рівнях відповідно [3, 4].

Практичний експеримент проводився на основі реальних даних GPS локацій викликів таксі у найбільших американських містах, таких як Сан-Франциско та Нью-Йорк, зібраних у період з 2020 по 2022 роки.

Після застосування методів кластеризації K-Means, DBSCAN та OPTICS до зазначеної вибірки без попереднього використання дискретизації, було отримано 8 розмитих і нечітко розділених кластерів локацій викликів таксі, що містили помітний рівень зашумленості та недостатньо чітко виражені географічні залежності.

Застосування дискретизації на основі гексагональної сітки H3 з 9-м рівнем деталізації та середньою довжиною сторони шестикутника у 200 метрів дозволило попередньою відфільтрувати шуми та похибки

визначення локацій, відсікаючи малоінформативні та зашумлені сегменти, дані в яких становили близько 12% від загального об'єму вибірки.

Повторне використання дискретизації на очищеній вибірці за допомогою 11-го рівня деталізації BingTiles з довжиною сторони квадрата у 76.4 метри скоротило розмір вибірки з більш ніж 4 мільйонів точок до приблизно 45 тясяч сегментів та дозволило сформувати значно виразніші кластери для всіх трьох методів. Отримані кластери набагато більш чітко окреслюють ділові та густонаселені райони міст, де виклики таксі відбувались відчутно частіше, що дозволить сформувати більш ефективні стратегії розміщення таксі та планування маршрутів. Такі дані допоможуть зменшити час очікування для пасажирів, підвищити ефективність роботи таксистів та покращити загальне обслуговування в цих районах.

Таким чином, завдяки тому що геодані стають все більш популярним джерелом інформації у різноманітних сферах, постійно зростаючий обсяг і неперервний характер створюють проблеми для аналізу та використання, зокрема, у задачах кластеризації. Було експериментально показано, що для вирішення цих проблем доцільно використовувати дискретизацію на основі сіток BingTiles та H3. Це дозволило попередньо відфільтрувати шуми та значно зменшити обсяг даних, що зробило кластерний аналіз більш ефективним та точним. Використання дискретизації робить геодані дедалі більш доступними та використовуваними у широкому спектрі завдань, сприяючи подальшому розвитку геоінформаційних технологій.

Список використаних джерел:

1. Demystifying Location Data Accuracy: [Електронний ресурс]. URL: https://www.mmaglobal.com/files/documents/location-data-accuracy-v3.pdf (дата звернення: 01.03.2024).

2. GPS accuracy: Lies, damn lies and statistics: [Електронний ресурс]. URL: https://www.gpsworld.com/gps-accuracy-lies-damn-lies-and-statistics (дата звернення: 02.03.2024).

3. Bing Maps Tile System: [Електронний ресурс]. URL: https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system (дата звернення: 03.03.2024).

4. Дудінова О.Б. Інтелектуальна обробка просторових даних в ГІС ландшафтно-екологічного моніторингу / О.Б. Дудінова, С.Г. Удовенко, Л.Е. Чала // Біоніка интелекту. – 2020. – Вип. 2 (95). – С. 43-50.