ДОДАТОК А

Графічний матеріал атестаційної роботи

Міністерство освіти та науки України

Харківський національний університет радіоелектроніки
Кафедра "Електронних обчислювальних машин "

Атестаційна робота на тему:
"Методи і моделі первинної обробки інформації з елементами самоподібності"

1

виконав: ст.гр. КСМм-19-1 Демяненко А.Є.
керівник: проф. Завизіступ Ю.Ю.

Актуальність проблеми

1) З кожним роком кількість інформації, яка розповсюджується у мережі Інтернет, дуже стрімко зростає.

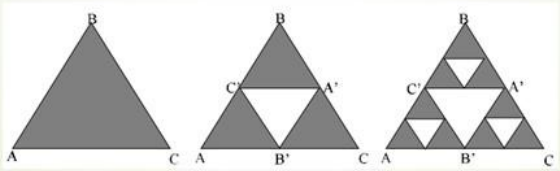
2) При використанні методів та моделей для первинної обробки інформації з елементами самоподібності, можна використати властивість самоподібності для розрахунку ключових атрибутів даних, які потім будуть передані в алгоритм кластеризації. Після цього алгоритм угрупує інформацію для її подальшого аналізу або використанні у машинному навчанні.

3) На даний момент не приділяється багато уваги створенню методів та інструментів для роботи з таким типом інформації.

2

Самоподібність інформації

Самоподібність та фрактали - це поняття, запроваджені Бенуа Б. Мандельбротом. Вони описують явище, коли певна властивість об'єкта - наприклад, природний образ, пакети які передаються в різних мережах, часовий ряд - зберігається щодо масштабування в просторі і часі.



3

Постановка задачі

Розробка модифікованого методу кластеризації даних, який буде використаний до текстових даних у розробленому веб-додатку. Дані для обробки повинні надходити до серверу через FTP-сервер або через клієнтський інтерфейс. Дана розробка повинна мати наступний функціонал:

- завантаження інформації для обробки;
- обробляти інформації у вигляд, до якого можна застосовувати дані методи;
- групувати інформацію по подібним ділянкам;
- повертати вихідні дані

4

Використані технології

Серверна частина:

- Scala
- Apache Spark
- Akka

Клієнтська частина:

- Javascript
- ReactJS

5

Структура даних у форматі CSV

Для первинної обробки було обрано взяти дані у форматі CSV з Міністерства охорони здоров'я щодо прогнозування кількості зайнятих ліжок з хворими на COVID-19.

https://healthdata.gov/dataset/covid-19-estimated-patient-impact-and-hospital-capacity-state

1	state	collection_date	total_beds_occupied_estimated	count_ll	count_ul	percentage_beds_occupied_estimated	percentage_ll	percentage_ul	total_beds	total_ll	total_ul
2	CW	22.10.2020	61,039	60,936	61,142	71.25	70.85	71.65	85,151	84,945	85,357
3	CW	23.10.2020	61,411	61,271	61,55	71.50	70.98	72.01	85,38	84,154	85,606
4	CW	24.10.2020	60,087	59,937	61,236	71.17	69.61	70.73	85,131	84,896	85,367

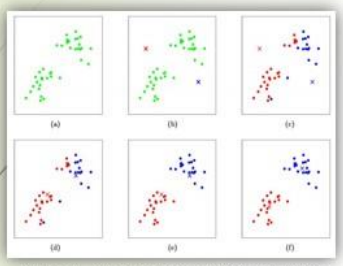
6

Опис полів даних

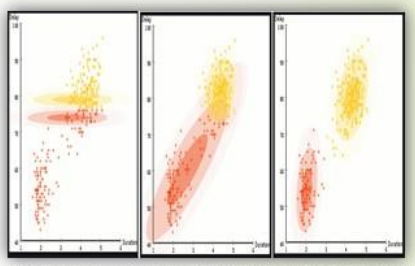
- state – двозначний код штату
- collection_date – приблизна дата
- inpatient_beds_occupied – прогнозована кількість зайнятих ліжок для даної дати та часу
- count_ll – прогнозована кількість зайнятих ліжок, нижня межа
- count_ul – прогнозована кількість зайнятих ліжок, верхня межа
- percentage_of_inpatient_beds_occupied_estimated – прогнозований процент зайнятих ліжок для даної дати та часу
- percentage_ll – прогнозований процент зайнятих ліжок, нижня межа
- percentage_ul – прогнозований процент зайнятих ліжок, верхня межа

7

Методи кластеризації даних



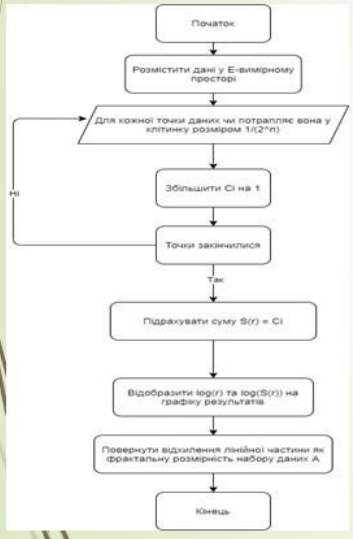
Метод кластеризації k-means

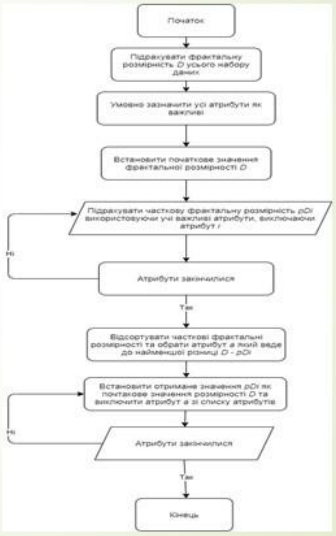


Метод кластеризації з використанням моделей Гаусса

8

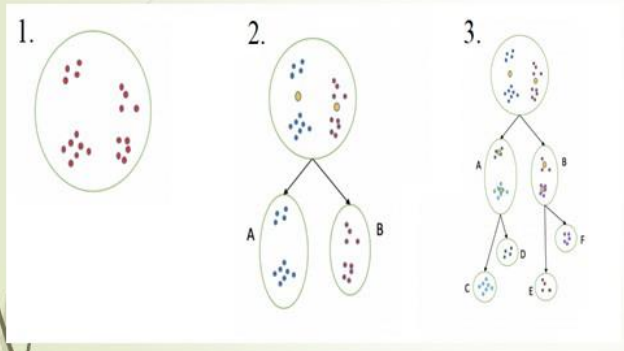
Алгоритм підбору ключових атрибутів





9

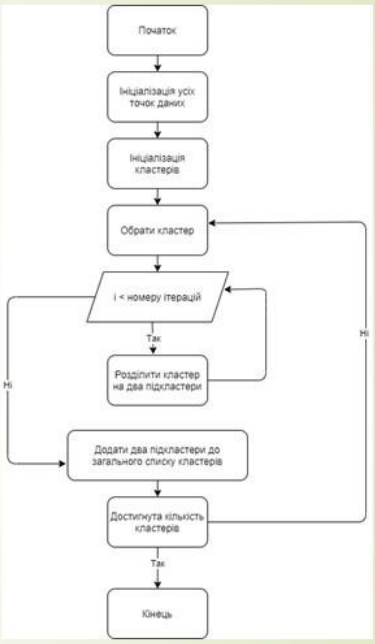
Метод кластеризації k-means з розподіленням



Ідея полягає в ітеративному розділенні на випадкове двійкове дерево, де кожне розбиття (вузол з двома дочірніми елементами) відповідає розбиттю точок кожного кластеру на 2.

10

Алгоритм виконання методу k-means з розподіленням



11

Приклад інтерфейсу користувача

Field	Data field type
state	string
collection_date	date
inpatient_beds_occupied_estimated	integer
count_ll	integer
count_ul	integer
percentage_of_inpatient_beds_occupied_estimated	float

12

Алгоритм роботи веб-серверу

Перша ітерація розподілення інформації. Як було описано вище користувач може задати одне поле основного розподілення, яке не використовує методи кластеризації.

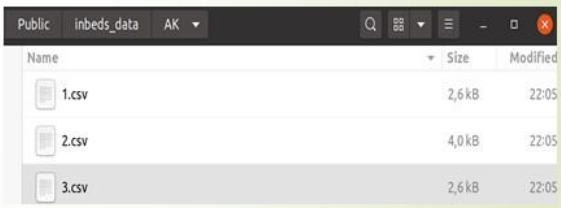
На другій ітерації до кожного розподіленого блоку даних застосовується модифікований метод кластеризації k-means з заданою кількістю кластерів K. Після виконання кластеризації усі кластери відокремлюються у свою групу. На програмному рівні це означає, що до даних додається ще одне додаткове поле під назвою “group_prediction_id”.

Після завершення другого етапу починається виконання етапу розділення сформованих даних по різним кластерам в залежності від їх значення поля “group_prediction_id”.

13

Результат роботи алгоритму

Збережені результати у наступній директорії: tmp/datasets_v1/user_id /inbeds_data_1606611904000/VT, VY … /1.csv, 2.csv, 3.csv, 4.csv, 5.csv, 6.csv



state	collection_date	total_beds_occupied_estimated	count_ll	count_ul	percentage_beds_occupied_estimated	percentage_ll	percentage_ul	total_beds	total_ll	total_ul	group_prediction_id
CW	22.10.2020	61,039	60,936	61,142	71.25	70.85	71.65	85,151	84,945	85,357	1
CW	23.10.2020	61,411	61,271	61,55	71.50	70.98	72.01	85,38	84,154	85,606	1
CW	24.10.2020	60,087	59,937	61,236	71.17	69.61	70.73	85,131	84,896	85,367	2

14

Висновки

У ході розробки даного методу первинної обробки інформації с самоподібністю був виконаний аналіз актуальності проблеми, та аналіз ефективного рішення проблеми. як результат було встановлено, що використання

Для вирішення цієї проблеми був проведений аналіз технологій та найоптимальніших технологій для даного рішення. за допомогою технологій роботи з великою кількістю даних (scala, spark) був створений веб-сервер, за допомогою якого можна обробляти інформацію для її подальшого використання в машинному навчанні та при статистичному аналізі.

Перспективи подальшого розвитку:

- створення більш детального інтерфейсу користувача (збереження результатів, авторизація тощо).
- обробляти дані, використовуючи різні методи кластеризації.

15