Розробка ситеми кластеризації антитіл на основі коефіцієнту перехресного зв’язування

Олександр Зелінський, Віталій Горлач, Юрій Лебедін

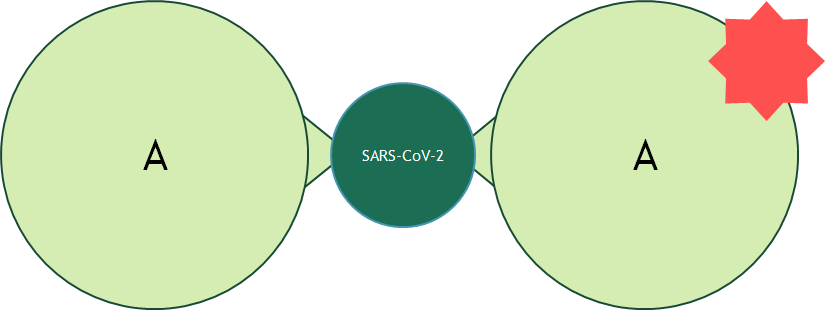
Факультет прикладної математики та інформатики

Львівський національний університет імені Івана Франка

Oleksandr.Zelinskyi@lnu.edu.ua

В умовах пандемії надзвичайно важливими є дослідження які несуть безпосередню користь для виявлення, запобігання та лікування вірусних захворювань, а якщо точніше вірусу Covid-19 або SARS-CoV-2. А в умовах поширення комп’ютерів та іншої потужної обчислювальної техніки зручним та важливим є використання комп’ютерних алгоритмів для виконання завдань пов’язаних з дослідженнями вірусів.

Дано молекулу вірусу SARS-CoV-2, до якої приєднуються два антитіла, для того щоб можна було їх відрізнити одне з них помічається \*. Для простоти вважатимемо, що все відбувається на площині, а антитіла це два круги однакового розміру, що приєднуються до меншого круга який представляє молекулу вірусу.

  
Рис. 1 Модель приєднання антитіл до вірусної молекули

Задача полягає у знаходженні двох антитіл (можуть бути однаковими), таких що знаходяться на оптимальній відстані одне від одного (не перетинаються, не знаходяться занадто близько). Зручним теоретичним способом для цього є розбиття списку антитіл на групи (антитіла з різних груп взаємодіють краще ніж з однієї). Далі буде наведено опис алгоритму розбиття.

Дані з експерименту подані у вигляді таблиці, де кожна комірка це коефіцієнт перехресного зв’язування міченого антитіла (зі стовпця) та не міченого (з рядка). В рядку позначеному як “blank” надані максимальні значення коефіцієнтів перехресного зв’язування для відповідного міченого антитіла.

Зображення, що містить стіл

Автоматично згенерований опис  
Рис. 2 Фрагмент початкових даних

Для подальшої роботи з даними їх позначають за наступним алгоритмом:

1. Цифрою 3 (темно-зеленим кольором) антитіла з хорошим зв’язуванням, якщо

(1)

1. Цифрою 2 (світло-зеленим кольором ) антитіла з середнім зв’язуванням, якщо

(2)

1. Цифрою 1 (білим кольором) антитіла майже без зв’язування, у всіх інших випадках.

Зображення, що містить текст, шафа

Автоматично згенерований опис  
Рис. 3 Фрагмент позначених даних

Тепер задача полягає у розбитті матриці перехресного зв’язування антитіл на групи за ознакою подібності раніше створеного показника зв’язування для полегшення виявлення оптимальних пар та приблизної локалізації місця зв’язування. Для цього використовують методи кластеризації, а саме k-modes.

k-modes – це алгоритм, який базується на алгоритмі k-means і використовується для кластеризації даних на основі якісних змінних. k-modes визначає кластери на основі відповідності категорій між точками даних. В даному алгоритмі відстань між двома точками даних X та Y описується як сума не схожих елементів:

(3)

Для визначення оптимальної кількості кластерів використовується elbow метод, який для різних значень k буде вибирати значення k у тій точці точці, де значення істотно не зменшується зі збільшенням значення k.

Для обробки даних та кластеризації використовувалась бібліотека kmodes, pandas, matplotlib та kneed з Python. В результаті виконання програми отримано оптимальне розбиття на 9 кластерів.

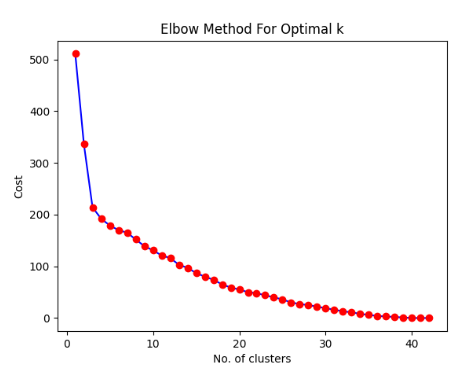
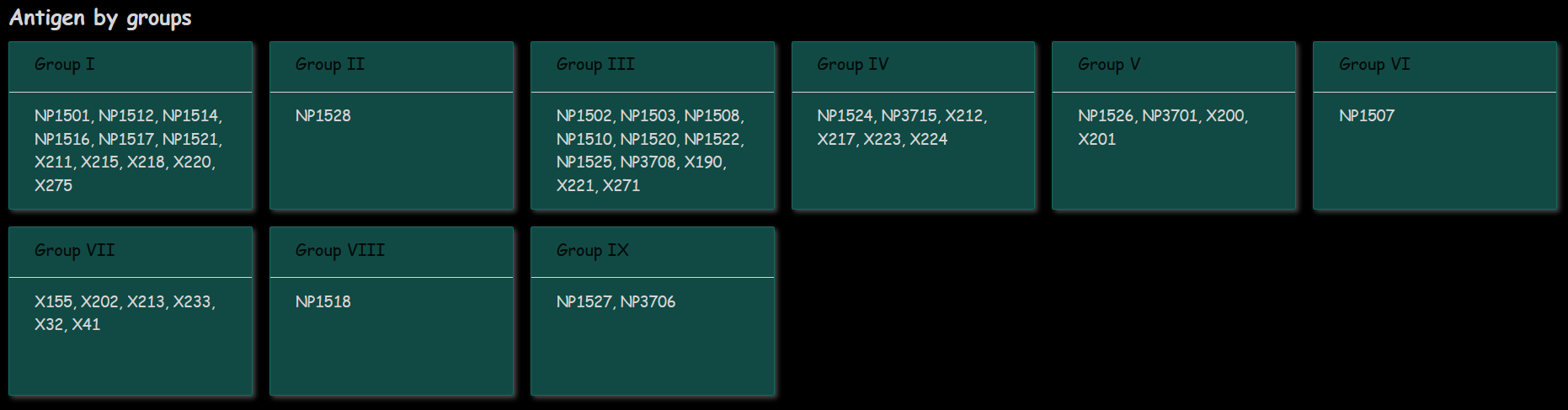


Рис. 4 Фрагмент позначених даних



Рию 5 Результат розбиття на кластери

|  |  |  |
| --- | --- | --- |
| Групи | Під групи | Елементи |
| 1 | A | NP1501, NP1514, NP1516, NP1517, NP1507 |
| B | X190, NP1526, X200, X201 |
| 1B/2 | | NP1512, NP1521 |
| 2 | | NP1502, NP1503, NP1508, NP1510, NP1520, NP1522, NP1525, X221, X271, NP3701, NP3708 |
| 2B/3 | | NP1528 |
| 3 | A | X202, X218, NP1518, NP1527 |
| B | X32, X155, X41, X212, X213, X217, X223, X224, X233, NP1524, NP3715 |
| 4 | A | NP3706 |
| B | X211 |
| C | X215 |
| 5 | | X220 |
| 6 | | X275 |

Табл. 1 Очікуваний результат розбиття

З результатів видно, що група І майже відповідає групі 1А в об’єднанні з 1B/2 та 4B, 4С, 5 і 6. Група ІІІ майже відповідає групі 2, Група ІІ відповідає групі 2B/3. Група 3B відповідає групі VII та групі IV. Зважаючи на те що кількість елементів які мають бути в однакових групах 30 з 43 елементів то можна вважати, що похибка становить близько 30%.

На рисунках 6 та 7 наведена візуалізація реального та очікуваного розбиття на групи за допомогою кольорів.

Зображення, що містить відкритий об’єкт, табло

Автоматично згенерований опис

Рис. 6 Візуалізація результатів розбиття

Зображення, що містить стіл

Автоматично згенерований опис

Рис. 7 Візуалізація очікуваного результату розбиття

**Список літератури**

1. Satyam Kumar Clustering Algorithm for data with mixed Categorical and Numerical features [Electronic resource]. – 2021. – URL:<https://towardsdatascience.com/clustering-algorithm-for-data-with-mixed-categorical-and-numerical-features-d4e3a48066a0>
2. Z. Huang*.* Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values(1998). Data Mining and Knowledge Discovery. 2(3): 283–304.
3. Python: K-modes explanation [Electronic resource]. – 2017. – URL :

https://stackoverflow.com/questions/42639824/python-k-modes-explanation