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.

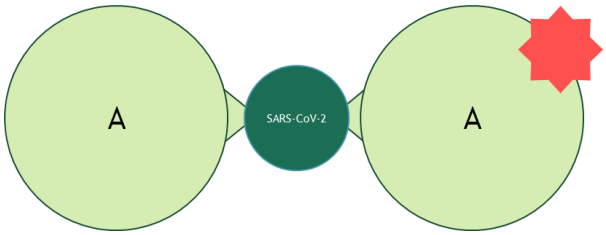


Рис. 1 Модель приєднання антитіл до вірусної молекули

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.

Labelled	NP1501*	NP1502*	NP1503*	NP1508*	NP1510*	NP1514*	NP1516*	NP1517*
blank	1.089	1.067	1.3664	1.412	1.67	1.07	1.1704	1.11
NP1501	0.449	0.715	1.0664	1.0248	1.136	0.26	0.4172	0.268
NP1502	0.893	0.425	0.336	0.196	0.349	0.3625	0.665	0.45
NP1503	0.768	0.309	0.068	0.052	0.095	0.305	0.7126	0.408
NP1507	0.422	0.856	0.7216	0.6088	0.785	0.1825	0.4256	0.154
NP1508	0.732	0.388	0.1456	0.0848	0.19	0.345	0.8456	0.411
NP1510	0.781	0.382	0.2152	0.1056	0.233	0.495	0.7826	0.527
NP1512	0.79	0.789	0.876	0.7504	0.979	0.735	1.015	0.737
NP1514	0.448	0.822	1.0968	1.0088	1.189	0.2475	0.4858	0.253
NP1516	0.385	1.034	0.9832	0.9304	1.053	0.1775	0.3052	0.152
NP1517	0.425	0.644	0.7952	0.7304	0.885	0.155	0.2758	0.079

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

$$\frac{-(cell[i][j]-blank[j])}{blank[j]} > 0.75 \quad (1)$$

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

$$0.5 < \frac{-(cell[i][j]-blank[j])}{blank[j]} \leq 0.75 \quad (2)$$

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

label	NP1501*	NP1502*	NP1503*	NP1508*	NP1510*	NP1514*	NP1516*	NP1517*
NP1501	2	1	1	1	1	3	2	3
NP1502	1	2	3	3	3	2	1	2
NP1503	1	2	3	3	3	2	1	2
NP1507	2	1	1	2	2	3	2	3
NP1508	1	2	3	3	3	2	1	2
NP1510	1	2	3	3	3	2	1	2
NP1512	1	1	1	1	1	1	1	1
NP1514	2	1	1	1	1	3	2	3
NP1516	2	1	1	1	1	3	2	3
NP1517	2	1	1	1	1	3	3	3

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:

$$d_1(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i), \text{де } \delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases} \quad (3)$$

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.

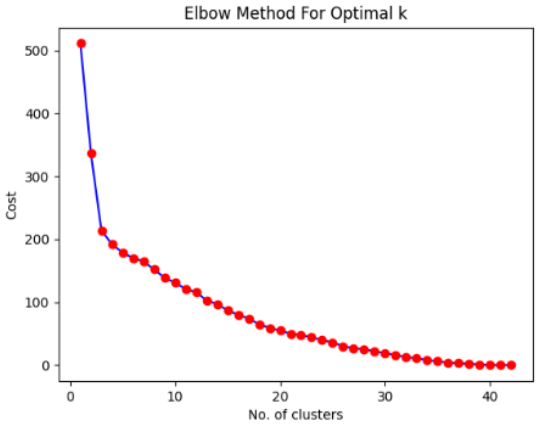
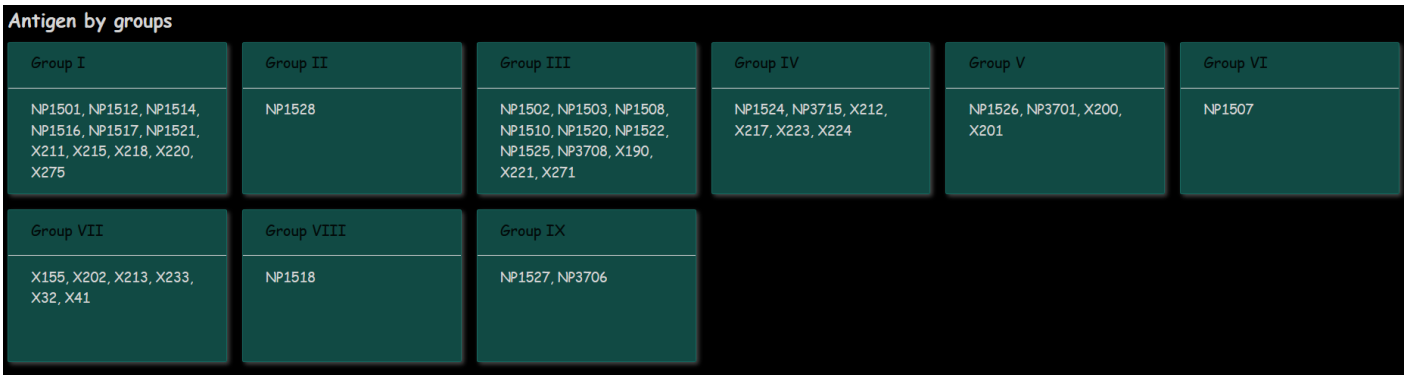


Рис. 4 Фрагмент позначених даних



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 I майже відповідає групі 1A в об'єднанні з 1B/2 та 4B, 4C, 5 і 6. Група III майже відповідає групі 2, Група II 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.

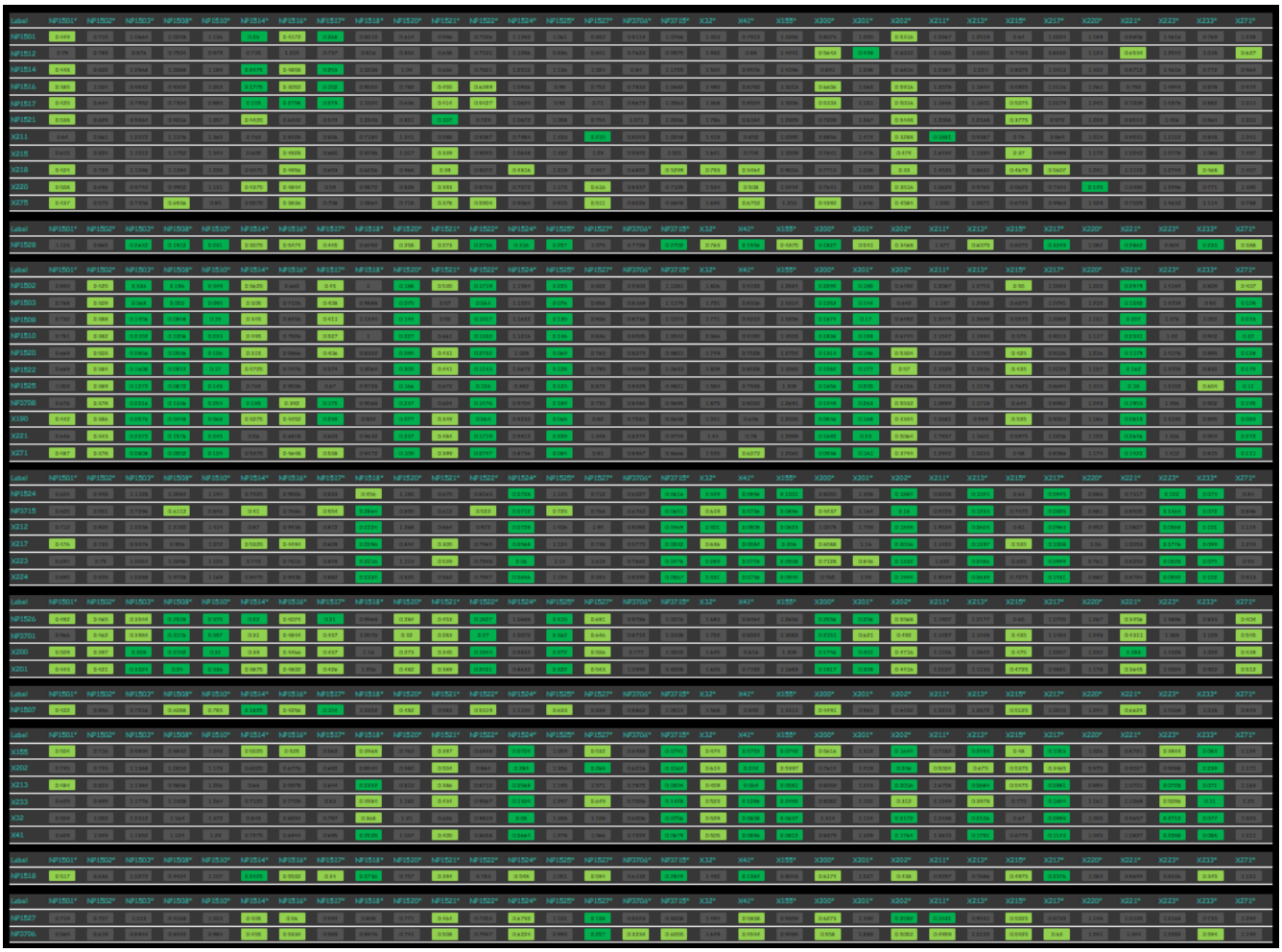


Рис. 6 Візуалізація результатів розбиття

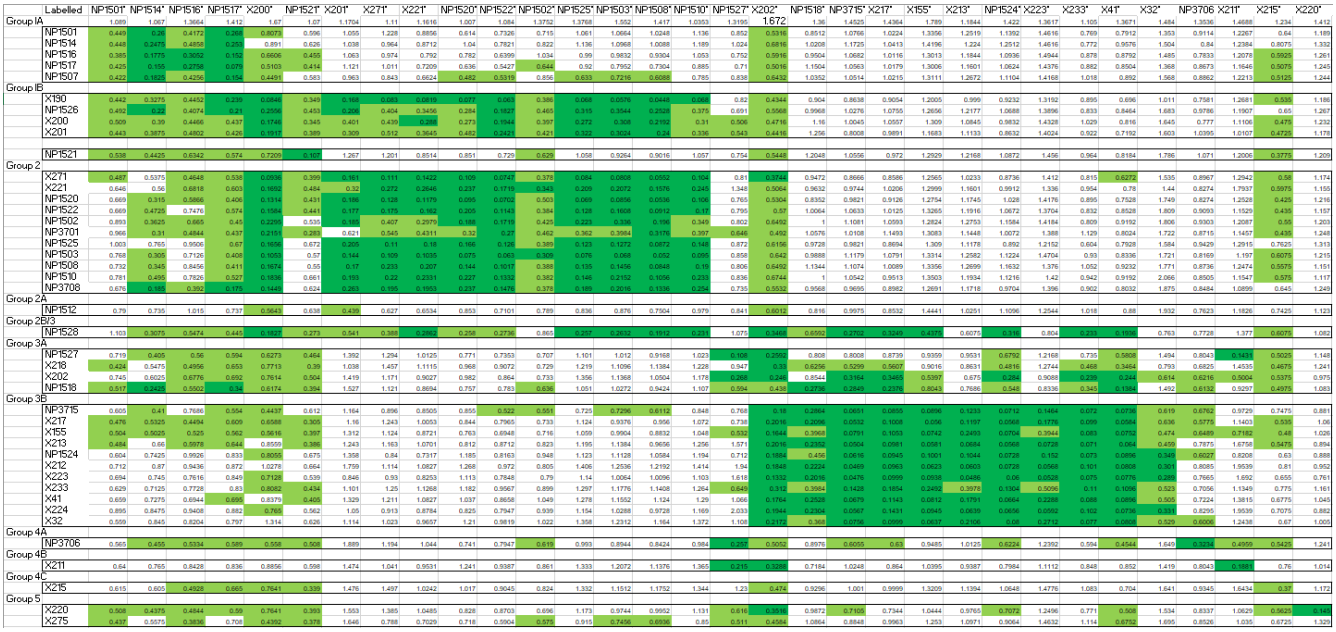


Рис. 7 Візуалізація очікуваного результату розбиття

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