Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Модель деревовидної штучної імунної мережі для кластеризації даних

(тема)

Виконав:

студент II курсу, групи СПм-20-2

Носова О.Є.
(прізвище, ініціали)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва освітньої програми)

Керівник: ст. викл. Фомічов О.О.
(посада, прізвище, ініціали)

Допускається до захисту

В.о. зав. кафедри ЕОМ _____ Волк М.О.
(підпис) (прізвище, ініціали)

2022 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Носовій Олександрі Євгенівні
(прізвище, ім'я, по батькові)

1. Тема роботи Модель деревовидної штучної імунної мережі для кластеризаії даних

затверджена наказом по університету від " 24 " березня 2022 р. № 413 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 18 травня 2022 р.

3. Вхідні дані до роботи

3.1 Документація мову програмування C#

3.2 Література про моделі та методи кластеризації об'єктів

3.3 Література про штучні імунні системи

3.4 Документація середовища розробки Visual Studio 2019 Community Edition

3.5 Документація фреймворку WindowsForms

4. Перелік питань, що потрібно опрацювати у роботі

4.1 Аналіз предметної області

4.2 Аналіз використаних технологій

4.3 Програмна реалізація

4.4 Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Слайд-презентація – 20 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Огляд існуючих методів вирішення задачі	29.03.21-05.04.22	
2	Вибір та обґрунтування методики дослідження	06.04.21-16.04.22	
3	Вибір інструментальних засобів	17.04.21-22.04.22	
4	Проведення експериментів	23.04.21-29.04.22	
5	Оформлення матеріалів кваліфікаційної роботи	30.04.21-07.05.22	
6	Подання кваліфікаційної роботи керівникові та	12.05.21-13.05.22	
	її попередній захист		
7	Подання кваліфікаційної роботи на	14.05.21-18.05.22	
	рецензування		

Дата видачі завдання 28 березня 2022 р.

Студент _____
(підпис)

Керівник роботи _____ ст. викл. Фомічов О.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 83 с., 15 рис., 2 табл., 1 дод., 26 джерел.

C#, .NET, WINDOWS FORMS, КЛАСТЕРИЗАЦІЯ, ШТУЧНА ІМУННА МЕРЕЖА, ІМУННА САМОРЕГУЛЯЦІЯ, АФІННІСТЬ, АНТИТІЛО, АНТИГЕН, АВІДНІСТЬ

Метою кваліфікаційної роботи є дослідження ефективності використання різних алгоритмів кластеризації та удосконалення існуючих методів на основі штучних імунних систем, що дозволить підвищити їх швидкість та забезпечити високу точність виділення кластерів.

У ході виконання кваліфікаційної роботи був проведений аналіз існуючих рішень, їх переваг та недоліків. Створено програмний застосунок з графічним інтерфейсом користувача, що забезпечує можливість дослідження впливу параметрів роботи різних методів та підходів на швидкість та точність кластеризації.

Дослідження проводилось з використанням зображень, на яких на білому фоні було випадково розміщено множину точок.

Розроблений програмний засіб виконує покладене на нього завдання кластеризації об’єктів різними способами.

ABSTRACT

Master’s thesis: 83 pages, 15 figures, 2 tables, 1 appendices, 26 sources.

C#, .NET, WINDOWS FORMS, CLASSIFICATION, CLUSTERIZATION, ARTIFICIAL IMMUNE NETWORK, TRAINING SAMPLE, IMMUNE TRAINING, AFHINNISM, ANTIBODY, ANTITIGENE, AVIDITY

The purpose of the qualification work is to investigate the effectiveness of the use of various clustering algorithms and to improve existing methods based on artificial immune systems, which will increase their speed and ensure high accuracy of cluster selection.

During the performance of qualification work, an analysis of existing decisions, their advantages and disadvantages was carried out. A software application with a graphical user interface has been created, which provides the ability to investigate the impact of the parameters of the work of different methods and approaches on the speed and accuracy of clustering.

The study was conducted using images on which a set of points was accidentally placed on a white background.

The developed software assigned to it the task of clustering objects in different ways.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

АНТИГЕН – це зовнішній вплив (вірус, бактерія).

АНТИТІЛО – це компонент штучної імунної системи, внутрішня клітина, що відповідає можливому рішенню задачі

ВИБІРКА – це множина об'єктів або подій, вибраних за допомогою визначеної процедури з генеральної сукупності для участі в дослідженніы

ГА – генетичний алгоритм

ГРАФ – сукупність об’єктів зі зв’язками між ними

ІА – імунний алгоритм

ІМ – імунна модель

МУТАЦІЯ – випадкові зміни в деякому антитілі; у штучній імунній системі відповідає випадковим змінам у можливому рішенні.

ПЗ – програмне забезпечення

ПОПУЛЯЦІЯ – це множина можливих рішень задачі.

СКВ – системи контролю версій

ШІС – штучна імунна система

aiNET – Artificial Immune Network

KNN – k Nearest Neighbor

MST – Minimum Spanning Tree

ВСТУП

Проблема кластеризації даних є однією з найбільш важливих проблем у сфері інтелектуального аналізу та має широкий спектр застосування у вирішенні прикладних та дослідницьких задач.

Метою кластеризації даних є розподіл набору об’єктів на специфічні групи – кластери за визначеним набором ознак та з використанням специфічних засобів визначення подібності цих об’єктів. Під час кластериації відбувається визначення кластерів та розподіл між ними набору вхідних даних. Широке розповсюдження інформаційних технологій призвело до створення великої кількості методів кластеризації, які широко застосовуються у машинному навчанні, інтелектуального аналізу даних, аналізу бізнес-процесів великих підприємств та задачах розпізнавання образів та звукових даних.

На сьогодні широке використання при вирішенні практичних задач класифікації мають ієрархічні методи, дерева прийняття рішень, методи C-means та машини опорних векторів. Однак, ці методи мають свої особливості, що ускладнюють їх застосування для вирішення деяких видів практичних задач, тому в останні роки широкого використання набули методи, що використовують біологічні принципи організації обчислювань, серед яких окремо можна виділити штучні імунні системи та мережі.

.

1 АНАЛІЗ ЗАДАЧІ КЛАСТЕРИЗАЦІЇ ДАНИХ ТА ПОСТАНОВКА ЗАДАЧІ ПРОЕКТУВАННЯ

1.1 Аналіз задачі кластеризації даних

Кластеризація даних відноситься до одного з основних методів інтелектуальної обробки інформації та є процесом розподілу сукупності об’єктів, які характеризуються набором ознак, на множину однорідних груп – кластерів [1-6]. Окрім того, кластеризація даних проводиться також для дослідження відносин між деякою множиною об’єктів, на основі однорідності їх ознак. При цьому об’єкти є однорідними та належать одному кластеру у разі, якщо вони мають спільні властивості, що виражені однаковими значеннями основних ознак. Для визначення однорідності (близькості) різних об’єктів між собою використовується певний критерій для вимірювання відстані між ознаками цих об’єктів [8].

Зараз існує кілька найбільш поширених видів поєднання сукупності об’єктів – це класифікація та кластеризація даних. Класифікація даних передбачає наявність вихідної множини заданих класів з відомими характеристиками, або використання множини класифікованих об’єктів, які визначаються як навчальна вибірка [3-10]. При цьому відбувається класифікація з контрольованим навчанням, або навчання із вчителем. Цей вид об’єднання даних є найпростішим видом класифікації. Під час класифікації з контрольованим навчанням набір початкових класів визначається явним чином за допомогою множини об’єктів навчальної вибірки. Навчальна вибірка використовується задля побудови деякої вирішальної функції, яка використовується для розподілу об’єктів, що розподіляються під час класифікації поміж існуючими класами. Для оцінки якості результатів класифікації можуть бути використані контрольні або перевіряючі вибірки об’єктів. Використання таких наборів класифікованих об’єктів відбувається наступним чином. Для всіх об’єктів перевіряючі

вибірки знову визначається належність до наданої множини класів, тобто проводиться повторна класифікація. Після цього результати повторної класифікації порівнюються із початковим вірним розподілом об’єктів між класами. Якість класифікації вважається високою, якщо результати повторної класифікації множини об’єктів контрольної вибірки не відрізняються від вихідних значень класифікації, або різниця між ними перебуває у межах статистичної похибки. У випадку, якщо результати цього порівняння повторної класифікації із початковою класифікацією різняться більш ніж на 25%, якість класифікації вважається низькою [9, 10] і потребує зміни алгоритму або його істотної модифікації.

Більшою складністю визначається завдання кластеризації, тобто класифікації із неконтрольованим навчанням, або навчанням без вчителя, або самонавчанням. При цьому завдання кластеризації полягає у розподілі множини вихідних об’єктів між кількома групами – кластерами [3-6]. Слід зазначити, що під час кластериації будь-яка інформація про вхідні класи або навчальну вибірку відсутня та об’єкти розподіляються між кластерами, що формуються в процесі самонавчання алгоритму. У деяких випадках, окрім множини початкових об’єктів, що підлягає групуванню, користувачем алгоритму визначається кількість кластерів, яка має бути отримана в результаті розподілу цих об’єктів. Зазвичай під час кластеризації кількість кластерів динамічно визначається під час самонавчання алгоритму. Для оцінки якості кластеризації той самий набір об’єктів підлягає кластеризації декілька разів. Після цього результати багаторазової кластеризації одного і того ж набору об’єктів порівнюються між собою таким самим чином, як це відбувається при роботі з контрольною вибіркою. При цьому якість кластризації вважається високою, якщо різницю між результатами багаторазового розподілу множини об’єктів міх визначеною кількістю кластерів, відрізняється на 8-10% [8, 9]. У випадку відмінності результатів кластеризації більш ніж на 30% якість алгоритму вважається низькою та формується висновок про необхідність зміни алгоритму, його модифікації

або модифікації способу його самонавчання.

Найбільш поширеним видом групування початкової множини об’єктів є автоматична класифікація. Головною особливістю цього методу розподілу об’єктів є поєднання основних особливостей класифікації та кластеризації даних [7-9]. Відповідно до цього при вирішенні задачі автоматичної класифікації можливе використання навчальної вибірки або наборів вхідних класів разом з механізмами самонавчання, який є властивим для методів кластеризації. Таким чином, функціонування методу автоматичної класифікації поділяється на два головні етапи:

– класифікація даних з урахуванням навчальної вибірки або набору початкових класів;

– формування кластерів для об’єктів, для яких не було визначено належності до одного з початкових класів.

Існуючі методи класифікації розподіляються між трьома групами [6-9]:

– евристичні методи – визначаються як найбільш простими методами групування даних. Характерною особливістю цих методів є використання деяких вхідних припущень про основні властивості класів та принципи розподілення між ними початкової множини об’єктів, що класифікуються. Окрім того, у цих методах використовується структурний підхід до вирішення задачі класифікації, тому такі методи часто називають методами прямої класифікації. Головною перевагою таких методів є простота реалізації та достатньо висока швидкість розподілення об’єктів між множиною визначених класів. Головним недоліком даних методів є зниження точності класифікації під час роботи з класами складної геометричної форми;

– оптимізаційні методи класифікації та кластеризації є методами пошуку оптимального розподілу об’єктів між визначеними класами або множиною кластерів, що формуються під час роботи алгоритму. При цьому процес класифікації зводиться до виділення можливих комбінацій розподілу початкових об’єктів на першому етапі, їх аналізу та виділення найбільш оптимального рішення на другому. Для таких методів властивим є

виникнення суто математичних проблем, таких як визначення шляхів досягнення глобального та локального оптимуму та інш. Головною перевагою цих методів є висока точність класифікації. Серед недоліків, характерних для оптимізаційних методів, виділяються висока чутливість до способу вихідного розподілу об’єктів та методу пошуку оптимуму, що виявляється у підвищенні складності даних методів та зниженні їх швидкодії;

– апроксимаційні методи є методами, в яких також як і в оптимізаційних методах, проводиться пошук оптимального розбиття вихідної множини об’єктів шляхом апроксимації. До цих методів належать методи, що функціонують на основі різних біологічних систем та принципів організації інтелектуальної обробки даних, таких як штучні нейронні мережі, генетичні алгоритми, штучні імунні системи та ін. Основною перевагою даних методів є висока точність класифікації при порівняно високій швидкодії. Такі методи за точністю розподілу та класифікації об’єктів не поступаються оптимізаційним методам, а за швидкодією аналогічні із евристичними методами.

1.2 Огляд існуючих методів кластеризації

Найбільш поширеними методами класифікації з контрольованим навчанням є [3-6,8]:

– метод k найближчих сусідів, kNN (англ. k Nearest Neighbors);

– метод вирішальних дерев, DT (англ. Decision Trees).

При цьому метод k найближчих сусідів належить до групи евристичних методів, а метод вирішальних дерев – до групи оптимізаційних методів. Головною особливістю роботи даних методів є використання навчальних вибірок або набору заданих класів, стосовно яких визначаються приналежності множини об’єктів, які досліджуються методом.

Залежно від типу ознак, що характеризують вихідні об’єкти та навчальну вибірку, дерева рішень поділяються на два основні види. У разі

коли об’єкти представлені матрицями дискретних ознак, для їх класифікації застосовуються стандартні вирішальні дерева. Якщо ознаки об’єктів, що класифікуються, описуються безперервними величинами – застосовуються дерева регресії [8].

У загальному випадку вирішальне дерево є графом, в якому вершинами є функції, які використовуються для розподілу об’єктів, що класифікуються, між заздалегідь визначеними класами. При цьому для такого розподілу може використовуватися як одна, так і група ознак, за якими об’єкти навчальної вибірки будуть розбиватися між класами. Листя графа, що представляє вирішальне дерево, є ідентифікатори класів, яких відбувається визначення власності класифікованих об'єктів. Ребра даного графа є зв'язками між функціями-вершинами, що здійснюють розподіл даних, що класифікуються, та ідентифікаторами вхідних класів. При класифікації даних за допомогою методу вирішальних дерев кожен об’єкт проходить за його вершинами, при цьому в кожній вершині відбувається перевірка ознак даного об’єкту на приналежність до одного з визначених класів. Оскільки з кожної вершини об’єкт потрапляє до іншої вершини, або відбувається перехід до листа вирішального дерева, що є ідентифікатором класу – використання методу вирішальних дерев призводить до повної класифікації початкового набору об’єктів. Найбільш поширеними є бінарні або булеві дерева рішень. У цих методах класифікації з кожної вершини виходить лише два ребра. Такий спосіб представлення дерева дозволяє провести швидке формування графа правил, якими проводиться класифікація вихідних даних, але збільшує його розміри.

Процес формування вирішального дерева полягає у послідовному розбитті навчальної вибірки на частини доти, доки у кожній частині сформованого дерева не буде виявлено об’єкти лише одного класу. При цьому можуть використовуватися різні критерії поділу об’єктів, що класифікуються. Наразі існує кілька видів критеріїв розподілу початкової множини об’єктів: критерії, що використовуються для виділення одного

класу, та критерії, що використовуються для групи класів (D-критерії) [7-8]. Використання критеріїв для виділення одиночних класів притаманно бінарним вирішальним деревам.

Серед переваг даного методу класифікації виділяються простота реалізації та модифікації, а також простота формування графа правил класифікації у разі використання бінарних дерев. Основними недоліками даного методу є висока чутливість до шумів, перешкод та різних спотворень у навчальній вибірці, а також необхідність перенавчання та повторної побудови дерева при виявленні помилки вибору критерію поділу, що призводить до зниження швидкодії алгоритму.

Найбільш поширеним методом класифікації, який може бути легко модифіковано, та який передбачає функціонування із контрольованим навчанням є метод k найближчих сусідів, kNN [1-6]. Особливістю даного методу класифікації є принцип визначення належності об’єкту тому чи іншому класу за результатами класифікації найближчих сусідів, кількість яких визначається деяким вхідним параметром k. Найближчими сусідами для об’єкта, що класифікується, є k класифікованих об’єктів навчальної вибірки. Приклад kNN класифікації об'єкта під час використання п'яти найближчих сусідів наведено на рисунку 1.1.

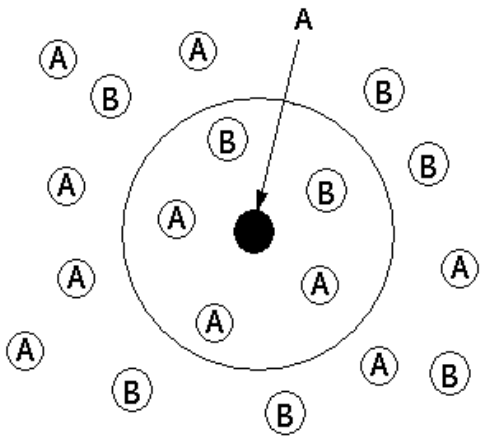

Рисунок 1.1 – Класифікація методом k найближчих сусідів

При цьому між об’єктом, що класифікується, та об’єктами навчальної

вибірки визначаються відстані відповідно до певної метрики. Після чого для кожного об’єкта визначаються k найближчих об’єктів навчальної вибірки. Належність об’єкту до того чи іншого класу визначається класифікацією більшості найближчих сусідів. Слід зазначити, що на практиці значення параметра k встановлюється непарним для того, щоб мінімізувати можливість виникнення ситуації неоднозначності при класифікації.

Формально вирішальна функція даного методу відповідно до [1-6] представляється так (формула 1.1):

$$a(u; X^l, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} [y_{i,u} = y], \qquad (1.1)$$

де u – об’єкт, що класифікується;

X^l – мнжина об’єктів навчальної вібірки;

k – ідентифікатор кількісті найближчих сусідів, за допомогою яких визначається належність об’єкту u до класу y_i, $y, y_i \in Y$.

Основними перевагами методу k найближчих сусідів є простота реалізації та модифікації, а також висока можливість інтерпретації результатів класифікації. Основним недоліком методу найближчих сусідів є низька точність порівняно з іншими методами класифікації з контрольованим навчанням.

Серед алгоритмів кластеризації даних найбільш поширеними є графові методи [5-7] та метод с-середніх (c-means) [1-4]. Функціонування даних методів можливе завдяки визначенню деякої апріорно відомої кількості кластерів, яке є вхідним параметром алгоритму, або використання деякого порогового значення, що обмежує зростання кластерів, що формуються алгортмом кластеризації.

Метод середніх (c-means) [1-4] є найбільш популярним та модифікованим методом кластеризації даних. Цей метод входить до групи

оптимізацинних методів угруповання даних. Основною відмінністю методу c-means є апріорно відома кількість кластерів, які мають бути отримані в результаті угруповання вихідних об’єктів. Слід зазначити, що в даному методі при виділенні кластерів велике значення набуває центр кластера, при цьому в ході кластеризації відбувається постійне формування кластерів і перевизначення їх меж. Відповідно до кожного кластера існує певний центральний об’єкт, званий центроїдом [2-4], стосовно якого виробляється визначення приналежності до даного кластеру того чи іншого об’єкта. При цьому початкові центроїди визначаються випадковим чином, і щодо даних об’єктів відбувається їх розподіл поміж виявленими центроїдами кластерів. Після цього центри сформованих кластерів уточнюються. В результаті уточнення центроїдів центром кластера може бути обраний інший об’єкт, що належить даному кластеру. На наступному етапі щодо нового центру кластера проводиться кластеризція групованих об’єктів та визначення центроїду знову повторюється. Уточнення центрів кластерів припиняється у разі, якщо результаті кластеризації об’єктів немає вибору ногового центроїда, тобто, в результаті угруповання та уточнення центрів кластерів на двох ітераціях алгоритму центроїди не змінюються. При цьому відбувається досягнення стану оптимального розбиття об’єктів, що розподідяються на кластери.

Слід зазначити, що завершення кластеризації може відбутися не тільки за відсутності відмінностей між центрами кластерів на двох розбиття об’єктів, але також у разі, якщо значення критерію середньої помилки розбиття перестануть відрізнятися.

Основними перевагами методу c-means є простота реалізації та модифікації алгоритму, висока швидкодія, можливість формування кластерів складної геометричної форми та низькі вимоги до обчислювальної системи.

При уточненні центрів кластерів у цьому методі виробляється визначення критеріїв помилки кластеризації [1-4], тобто. визначення центроїдів відбувається мінімізація цільової функції (формула 1.2):

$$E = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - c_i)^2 , \qquad (1.2)$$

де k – кількість кластерів;

S_i – кластер;

x_j – об’єкт, що належить кластеру S_i;

c_i – центр (центроїд) кластеру S_i.

Серед основних недоліків даного методу виділяють високу чутливість до помилок на етапі визначення початкових центрів кластерів та необхідність попереднього визначення кількості фінальних кластерів.

Одними з найбільш ранніх та широко використовуваних методів угруповання даних є методи, що функціонують на основі математичної теорії графів, що одержали назву графових методів [3-6]. Дані методи часто відносять до ієрархічних методів класифікації у зв'язку з деякою подібністю у принципах роботи та візуалізації результатів. Робота графових методів кластеризації даних здійснюється при використанні матриці суміжності (матриці близькостей) об'єктів, що групуються під час кластеризації. При цьому на початковому етапі роботи даних методів відбувається формування матриці суміжності для визначення ваг ребер графа, що формується, на основі якого проводиться кластеризація, а також для визначення центрів кластерів. Матриця суміжності формується шляхом визначення відстаней відповідно до деякої метрики між об'єктами, що групуються, при цьому для кожного об'єкта визначаються відстані до всіх інших об'єктів. Для виділення кластерів зі сформованого зваженого графа може використовуватися метод мінімального покриває дерева MST [6,7], або значення граничної відстані включення об'єкта, що обмежує збільшення розмірів кластера. Слід зазначити, що в результаті такого угруповання з багатьох вихідних об'єктів формуються кластери-стрічки [7-9], а не кластери з виділеним центром, як це відбувається в більшості існуючих методів кластеризації.

Головними перевагами даних методів угруповання об’єктів є висока масштабованість алгоритму, висока точність кластеризації, простота реалізації та модифікації алгоритму, а також можливість легкої інтерпретації результатів розподілу об’єктів між множиною сформованих кластерів.

Серед недоліків даного методу виділяють обмеження на способи представлення даних - дані методи кластеризації добре працюють з метричними ознаками, вираженими цілими або речовими значеннями, при цьому точність класифікації знижується при обробці символів неметричних ознак групованих об'єктів.

Існуючі методи автоматичної класифікації, що використовують контрольоване навчання та механізми самонавчання, є комбінацією різних методів класифікації та кластеризації даних. При цьому на першому етапі роботи з даних, що класифікуються, виділяються об’єкти, які можуть бути згруповані при використанні методів контрольованого навчання за допомогою навчальної вибірки або набору вихідних класів. Після цього для багатьох некласифікованих об’єктів відбувається кластеризація. В результаті автоматичної класифікації вихідні об’єкти розподіляються серед існуючих класів, або формують нові кластери, які будуть використовуватися для класифікації інших наборів даних.

1.3 Штучні імунні системи

В даний час у галузі моделювання систем штучного інтелекту велика увага приділяється системам, які функціонують на основі різних біологічних принципів організації розрахунків. Теорія штучної імунної системи є однією з останніх областей у галузі штучного інтелекту та м’яких обчислювальних систем [12-17]. У ієрархії систем, що використовуються для побудови штучного інтелекту теорія штучних імунних систем займає важливе місце поряд з такими теоріями, як штучні нейронні системи, генетичні алгоритми, алгоритми рою тощо (рисунок 1.2).

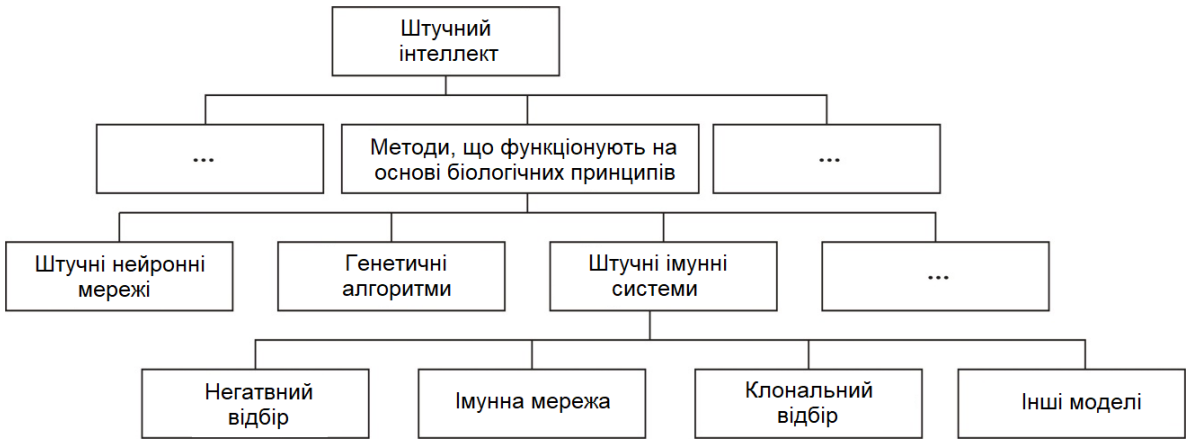

Рисунок 1.2 – Штучні імунні системи у ієрархії систем штучного інтелекту

Слід зазначити, що на відміну від інших систем штучного інтелекту, які функціонують на основі біологічних принципів організації обчислень, відокремлюється теорія штучних імунних систем заснована на основних принципах роботи природної імунної системи людини. Основне завдання природної імунної системи полягає в виявленні будь-яких чужорідних предметів, що входять у тіло, їх локалізацію та блокування, розпізнавання особливостей сторонніх предметів під час клонування та мутації клонів, а також їх руйнування на завершальній стадії [15-16] . Використання перелічених механізмів для вирішення практичних проблем дозволяє вирішити не тільки проблеми класифікації з контрольованою підготовкою та кластеризацією, але й завданням автоматичної класифікації даних. Слід зазначити, що для штучних імунних систем, а також для більшості систем інтелектуальної обробки інформації, характерним є використання спеціального термінологічного апарату. Основними об’єктами, що використовуються в штучній імунній системі, є лімфоцити, антитіла, антигени та фагоцити. У той же час антигенами є усі чужорідні об’єкти, визнані імунною системою.

Під час вирішення проблеми класифікації популяція антигенів формується задля визначення навчальних та контрольних вибірок об’єктів, що підлягають класифікації, за допомогою яких відбувається розподіл об’єктів поміж класами. Антигени нейтралізуються лімфоцитами, які, в свою

чергу, розподіляються на дві групи: В лімфоцити (В-клітини) та Т-лімфоцити (Т-клітини) [15]. Основна відмінність лімфоцитів – це особливості їх функціонування. Т-клітини використовуються системою для організації накопичення В-клітин, які, в свою чергу, вступають у пряму взаємодію із популяцією антигенів, беруть участь у їх розпізнаванні та нейтралізації. У той же час В-клітини пов’язані з антигенами шляхом виробництва антитіл, які під час клонування та мутацій здатні змінити свою структуру та адаптуватися до особливостей різних антигенів. Таким чином, процес клонування, мутації клонів та їх вибору є етапами самонавчання штучної імунної системи, що дозволяє автоматично класифікувати об’єкти.

Серед особливостей та властивостей природної імунної системи, що використовується для вирішення різних практичних проблем, виділяють [13]:

– здатність утворювати нові типи антитіл та відбір найбільш підходящих з них для взаємодії з антигенами;
– можливість самостійного навчання, під час якого спостерігається зміна концентрації антитіл та їх ознак при взаємодії з антигенами;
– формування різноманіття антитіл, що досягаються за допомогою мутації, тобто зміни в ознаках антитіл або клонів з метою формування багатьох різних антитіл для збільшення ймовірності розпізнавання ознак антигенів;
– використання порогового механізму, робота якого полягає в тому, що вироблення антитіл, їх клонування та мутація проводяться лише після подолання певного порогу, залежно від ступеня подібності ознак антитіл та антигенів;
– використання асоціативної імунної пам’яті, яка дозволяє зберігати інформацію про характеристики різних антигенів, розпізнані та нейтралізовані імунною системою.

На сьогоднішній день більшість існуючих імунних методів та алгоритмів інтелектуальної обробки інформації за принципом функціонування належать до однієї з чотирьох основних імунних моделей. Ці

моделі включають [13-17]: модель штучної імунної мережі, модель клонального вибору, модель негативного вибору, модель, що функціонує на основі теорії небезпечних сигналів.

Принцип функціонування моделі клонального відбору вперше був описаний у 1959 році М. Берентом, який є автором назви моделі та терміном «клональний відбір» [13-15]. Основою характерною рисою цього принципу є ідея поведінки В-клітин, коли чужорідні предмети виявляються природньою імунною системою, що полягає у трансформації популяції антитіл шляхом клонування, мутації та редагування популяції клонів та антигенів. Відповідно до цієї теорії, процес вибору клонів – це механізм, який забезпечує вироблення антитіл, що характеризуються унікальними значеннями ознак, тобто виключається імовірність утворення декількох антитіл, які мають однакові значення основних особливостей. У той же час, єдиною подією, від якої залежить процес відбору клонів, є реакція розпізнавання антигену, що виражається у відповідній спорідненості або афінності між ними. При цьому під терміном аффінність розуміють ступінь взаємодії між двома об’єктами, що визначається наступним чином [13-17]:

$$aff_{ij} = \left(1 + d_{ij}\right)^{-1}, \tag{1.3}$$

де d_{ij} – евклідова або манхеттенська відстань між ознаками i-го та j-го імунних об’єктів.

Принцип роботи моделі клонального вибору представлений нижче (рис. 1.3) і умовно поділяється на кілька основних етапів. На першому етапі функціонування моделі клонального відбору чужорідний антиген виявляється антитілами імунної системи. У той же час антитіла, які виявили цей антиген, починають розпізнавати його ознаки, внаслідок чого відбувається їх стимуляція та клонування.

Під час клонування кожне стимульоване антитіло утворює багато однакових клонів, які є копіями цього імунного об’єкту. Після закінчення

процесу клонування стимульованих антитіл відбувається мутація сформованої популяції клонів, ознаки цих клонів піддаються змінам випадковим або детермінованим способом. Ця трансформація клонів відбувається для розпізнавання антигенів та отримання інформації про їх структуру шляхом відтворення їх особливостей.

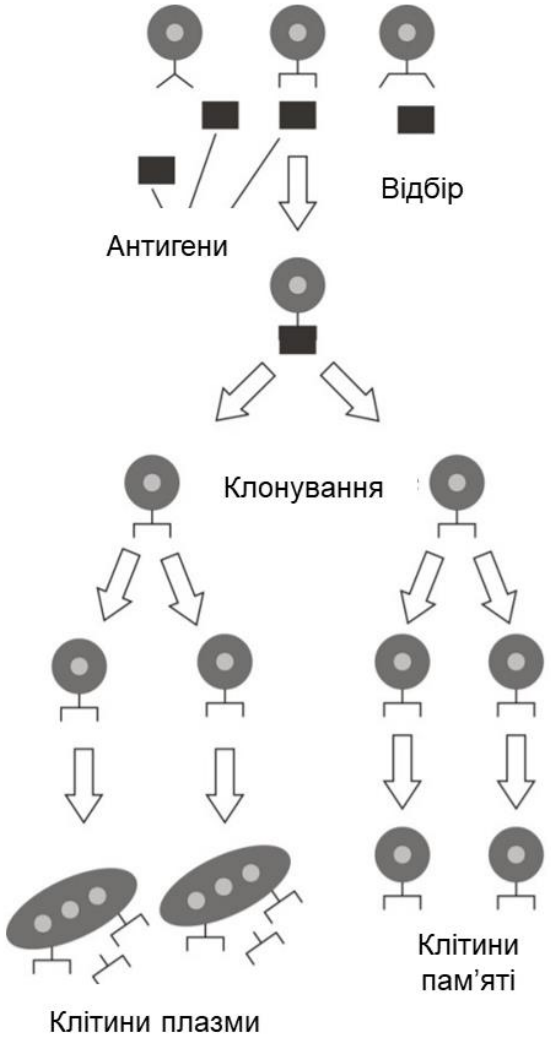

Рисунок 1.3 – Принцип клонального відбору

Етап відбору клонів починається з подання визнаних антигенів до багатьох клонів з параметрами, зміненими під час мутації. На цьому етапі спорідненість визначається між клонами антигенами як ступінь ідентичності їх ознак. Після цього популяція клонів редагується шляхом видалення об'єктів, що характеризуються найгіршими значеннями спорідненості, тобто

аффінностями до антигенів. Такі клони стають плазматичними клітинами і більше не використовуються імунною системою. Решта клонів замінюють батьківські антитіла, якщо вони характеризуються найкращими показниками афінності до стимулюючих антигенів. При цьому у випадку отримання повної інформації про антиген завдяки повному відновленню його (антигена) ознак, антитіла та клони стають специфічними клітинами імунної пам'яті [14-16]. Решта антитіл залишаються плазматичними клітинами та продовжують циркулювати через організм.

В даний час існує кілька методів, які функціонують на основі моделі клонального відбору: Clonalg та BCA [13-16], а також метод RLAIS [15,16], який використовує елементи цієї моделі в редагуванні популяції клонів. Ці алгоритми є універсальними і можуть використовуватися для вирішення проблем класифікації, кластеризації, розпізнавання зображень, оптимізації та управління роботизованними системами.

Модель штучної імунної мережі, запропонованої В. Ерне [13-17], надає більше можливостей організувати взаємодію як між так і внутрішньо у популяціях імунних об'єктів. У цій моделі всі антитіла та їх клони – це регульована мережа клітин, які взаємодіють не лише з антигенами, але і між собою. У цьому випадку взаємодія між антитілами відбувається не тільки тоді, коли вони стимулюються антигенами, але і за відсутності сторонніх об'єктів. Ця модель заснована на припущенні взаємодії лімфоцитів між собою шляхом встановлення взаємозв'язків між їх антитілами (рисунок 1.4). Таким чином, розпізнавання антигену здійснюється не одним антитілом або клоном, а організованою мережею лімфоцитів та їх антитілами.

У той же час, якщо антиген виявляється та визнається системою, відбувається стимуляція мережі антитіл, внаслідок чого обмежена кількість антитіл, що характеризуються високими значеннями аффінностей до антигенів, утворює популяцію клонів. Після цього сформовані клони піддаються мутаціям, щоб отримати більше інформації про визначені антигени. Для редагування популяції клонів та антитіл ця модель

використовує механізм супресії (стиснення мережі) [13-17]. При цьому механізм супресії може використовувати не тільки принципи моделі клонального відбору, але й виключно мережеві принципи для редагування кількості антитіл, в яких антитіла, що характеризуються низькими значеннями афінностей до інших стимульованих антитіл, та видаляються з поточної популяції імунних об’єктів.

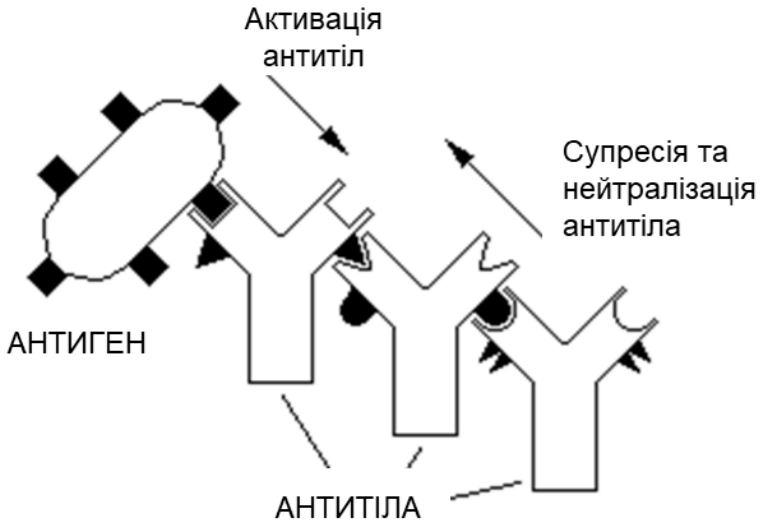

Рисунок 1.4 – Принцип мережевої взаємодії між антитілами

В даний час існує кілька основних методів, які впроваджують принципи моделі імунної мережі [15-17]: методи aiNET, opt-aiNET та метод штучних імунних шарів розпізнавання RLAIS. Ці методи в основному використовуються для вирішення проблем розпізнавання та аналізу даних та є основними відповідно до інших існуючих імунних методів, які функціонують на основі моделі штучної імунної мережі.

Модель негативного відбору та модель, що функціонує на основі теорії небезпечних сигналів, є вузько спеціалізованими імунними моделями, що використовуються для вирішення обмеженої кількості практичних проблем, пов’язаних з організацією захисту від вірусів, управління та виявлення аномалій [15]. Ці моделі функціонують на основі принципів розділення та розпізнавання системних клітин та сторонніх об’єктів, які надходять до

системи в обмеженій кількості. У цьому випадку методи негативного відбору випадковим чином виробляють набори детекторів антитіл та видаляють ті, що характеризуються високими значеннями аффінносітей до окремих антигенів або всієї популяції антигенів. Моделі негативного відбору – це найдавніші формалізовані імунні моделі. Найвідомішими є робота Фореста для створення алгоритмів виявлення аномалій. В результаті цього було розроблено один із найбільш ранніх імунних методів негативного відбору, що отримав назву NSA [15].

При роботі із різними імунними методами та моделями під час вирішення практичних завдань, ці методи формуються з багатьох імунних операторів. Таке формування алгоритму з операторів відбувається задля спрощення їх модифікації та реалізації [16]. У цьому випадку імунний оператор – це певна функція або набір функцій, що виконують ті самі дії задля перетворення популяції антитіл. При цьому модифікація імунного оператора може призвести до зміни поведінки всього імунного алгоритму. Серед найбільш модифікованих імунних операторів, визначаються оператор клонування, оператор мутації, оператор відбору клонів та оператор супресії антитіл. Слід зазначити, що використання певного оператора в імунному алгоритмі часто визначає належність цього методу до певної імунної моделі. Відповідно до цього, використання оператора відбору клонів або оператора старіння антитіл не підтримується алгоритмами, які функціонують на основі моделі штучної імунної мережі, а використання оператора супресії не підтримується в методах, які функціонують на основі моделі клонального вибору або моделей негативного чи позитивного вібдору.

Під час вирішення практичних проблем імунні моделі та методи можуть поєднуватися з іншими підходами до організації розрахунків, таких як нейронні мережі, генетичні алгоритми, нечітка логіка або використовувати суто статистичні чи математичні методи побудови систем штучного інтелекту. У таких випадках отримані моделі та методи називаються гібридними. Слід зазначити, що використання гібридних методів організації

обробки інтелектуальної інформації зазвичай призводить до найкращих результатів, ніж використання моделей, які функціонують в межах однієї з перелічених теорій.

1.4 Постановка задачі проектування

Головною метою цієї роботи э розробка методу автоматичної класифікації або кластеризації на основі принципів побудови штучної імунної системи. При цьому у якості базового методу визначається метод aiNET, що функціонує на основі моделі штучної імунної мережі. Це пов’язано з тим, що модель штучної імунної мережі дає більше можливостей щодо перетворення та регуляції популяції антитіл під час імунного навчання або процесу саморегуляції, властивого штучним імунним системам, ніж інші існуючі імунні моделі. Це відбувається завдяки механізму мережевої взаємодії популяції антитіл. Задля виконання задачі проектування необхідно вирішити ряд наступних задач:

– використати імунний метод aiNET для побудови алгоритму кластеризації, що буде відзначатися високою ефективністю;

– реалізувати можливості використання графічних файлів із різнокольоровими об’єктами як даних на кластеризацію та навчальних чи контрольних зразків;

– модифікувати оператори клонування, мутації та супресії задля організації оптимальної роботи методу кластеризації;

– підвищити точність кластеризації даних порівняно з методами, які функціонують на основі моделі клонального вибору, не втрачаючи високої швидкості, характерні для методу aiNET.

Окрім того, у якості критерію визначення близькості між об’єктами під час кластеризації передбачається застосовується критеріїв аффінності та рівня стимуляції антитіл [14-17]. У якості критерію, що обмежує розмір кластерів, утворених під час саморегуляції об’єктів штучної імунної мережі,

передбачається використання природного порогу аффінності NAT (Natural Affinity Threshold), наведений у [16], для антигенів, антитіл та клонів. При цьому початкові дані під час кластеризації представлені набором антитіл, утворених динамічно або шляхом завантаження із графічних файлів.

Слід зазначити, що основним недоліком aiNET, який є базовим методом, що використовується для формування нового методу кластеризації, є низька точність порівняно з іншими імунними алгоритмами. Визначений недолік є результатом використання оператора випадкового додаткового формування антитіл після впровадження мережевої супресії антитіл. Задля усунення цього недоліку у класичній реалізації методу aiNET, необхідно внести ряд змін щодо організації мутації, супресії та додаткового розсіювання імунних об’єктів. Крім того, потрібно приділяти багато уваги критеріям припинення імунного навчання та процесу саморегуляції системи, що визначає швидкість алгоритму кластеризації.

2 ОСОБЛИВОСТІ РОБОТИ МОДЕЛІ ДЕРЕВОВИДНОЇ ШУЧНОЇ ІМУННОЇ МЕРЕЖІ ПРИ КЛАСТЕРИЗАЦІЇ ДАНИХ

2.1 Особливості кластеризації даних на основі імунного підходу

Для вирішення задачі кластеризації даних на основі імунного підходу доцільне використання моделі штучної імунної мережі. При цьому інші імунні моделі, такі як модель клонального відбору, або моделі позитивного чи негативного відбору можуть використовуватися тільки ля вирішення задач класифікації об’єктів з контрольованим навчанням. Це обумовлено тим, що модель штучної імунної мережі передбачає аналіз не тільки взаємодії типу антитіло-антиген, але й взаємодії типу антитіло-антитіло. Слід зазначити, що моделі, які сфокусовані на дослідженні зв’язків тільки між антитілами та антигенами можуть використовуватися в задачах класифікації, прогнозування та розпізнавання образів, тобто в задачах з контрольованим навчанням та за умов використання навчальних та контрольних вибірок об’єктів, які зазвичай представляються популяцію антигенів. Під час кластеризації з набору початкових об’єктів формується популяція антитіл, які під час взаємодії між собою формують групи-кластери. Відповідно до цього, завдяки використанню моделей штучних імунних систем при розв’язанні задачі кластеризації об’єктів, популяція антигенів відсутня, тому що кластеризація не передбачає групування антитіл з контрольованим навчанням та роботи з тренувальною вибіркою антигенів. Тому при кластеризації всі обчислення відбуваються тільки у самій популяції антитіл, які презентують початковий набор об’єктів.

Задля підвищення якості кластеризації та підвищення швидкості процесу групування об’єктів доцільним є проведення додаткової роботи з набором цих об’єктів, з котрих формується початкова популяція антитіл. Під час цієї підготовчої роботи необхідно провести визначення шкал ознак об’єктів. Це відбувається у випадку, якщо ці шкали не визначені ще до старту

процесу кластеризації. При дослідженні шкал початкових даних відбувається визначення діапазонів можливих значень для всіх ознак, які характеризують набір об’єктів для кластеризації [9-11]. Відповідно до цього, для кожної групи ознак під час визначення шкал даних, відбувається визначення максимального та мінімального можливих значень, які може прийняти та чи інша ознака під час мутації імунного об’єкту. Це необхідно зробити для того, щоб під час мутації ознаки клонів не приймали значення, що виходять за межі допустимого діапазону значень, бо це може призвести у перспективі до значного падіння швидкості кластериації.

Під час виділення критеріїв групування об’єктів відбувається визначення порогової аффінності NAT (Natural Affinity Threshold) між усіма об’єктами, що підлягають кластеризації. Визначення порогу NAT забезпечується завдяки використанню механізму мережевої взаємодії між антитілами. Відповідно до цього, для кожного об’єкта з початкового набору антитіл відбувається обчислення аффінностей до всіх інших антитіл. Тому визначення параметру NAT відбувається за формулою 2.1 [15]:

$$NAT(AB) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n-1} aff\left(ab_i, ab_j\right)}{n(n-1)}, \tag{2.1}$$

де n – кількість антитіл у популяції;

$aff\left(ab_i, ab_j\right)$ – значення аффінності між антитілами.

Використання порогового значення NAT, що характеризує початкову популяцію антитіл, дозволяє проводити відбір антитіл у якості початкових центрів кластерів, що формуються в процесі роботи імунного методу.

Одним з найбільш поширених способів формування кластерів є визначення необхідної кількості кластерів у якості вхідного параметру для алгоритму кластеризації. Альтернативним способом вибору кількості

кластерів є автоматичне визначення цієї кількості в процесі роботи алгоритму. Однак, під час такого автоматичного визначення часто можуть відбуватися формування достатньо великої кількості невеликих кластерів у випадку використання значної кількості метричних ознак для опису початкової множини об’єктів. Також це може призвести до виділення невеликої кількості досить крупних кластерів у випадку зменшення кількості ознак, що використовуються для опису об’єктів, що кластеризуються.

Для забезпечення формування деревовидної структури мережі антитіл пропонується формування K-пов’язаного графу антитіл. При цьому у якості міри, що визначає силу зв’язку або подібності між антитілами штучної імунної мережі використовується поняття аффінності. Відповідно до цього, перед формуванням деревовидної мережі антитіл, поміж всіма початковими антитілами обчислюються аффінності у межах деякого простору ознак. Таким чином, на першому етапі мережа антитіл формується як граф, де кожна вершина пов’язана з усіма іншими вершинами цього графу.

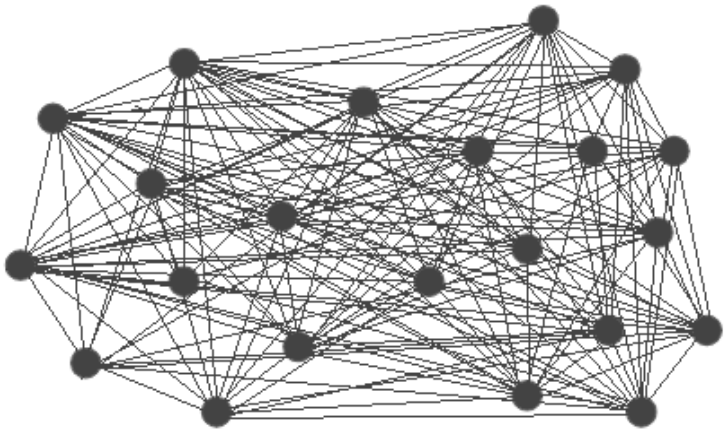

Рисунок 2.1 – Перший етап формування деревовидної імунної мережі

При цьому скорочення кількості зв’язків між антитілами відбувається завдяки видаленню зв’язків, що характеризуються мінімальними значеннями аффінностей. Відповідно до виразу, що використовується для визначення аффінності (1.3), діапазон значення аффінності вимірюється у

межах (0.0; 1.0]. При цьому, чим меншим є значення аффінності – тим слабший зв’язок між об’єктами, тобто, тим менша подібність між імунними об’єктами. Збільшення значення аффінності означає підсилення зв’язку між парою імунних об’єктів.

Слід зазначити, що при зменшенні кількості зв’язків між антитілами деревовидної імунної мережі передбачається використання деякого вхідного параметру К, який обмежує кількість зв’язків між антитілами мережі. Тобто, кожне антитіло може створювати не більше ніж К зв’язків з іншими антитілами мережі, інші зв’язки, які створює це антитіло видаляються. Ця ідея запозичена з алгоритму класифікації, що має назву К найближчих сусідів. Відповідно до цього, під час формування деревовидної мережі антитіл, кожне антитіло може формувати не більше ніж К зв’язків, що характеризуються максимальними аффінностіми. Слід зазначити, що параметр К може змінюватися у діапазоні від 2 до безлічі зв’язків. При цьому збільшення кількості активних зв’язків, які залишаються під час будування деревовидної імунної мережі, призводить до значного зниження швидкості роботи імунного алгоритму кластеризації. Однак, зниження кількості зв’язків між антитілами деревовидної імунної мережі призводить до зниження точності кластеризації та появі значної кількості помилок групування імунних об’єктів.

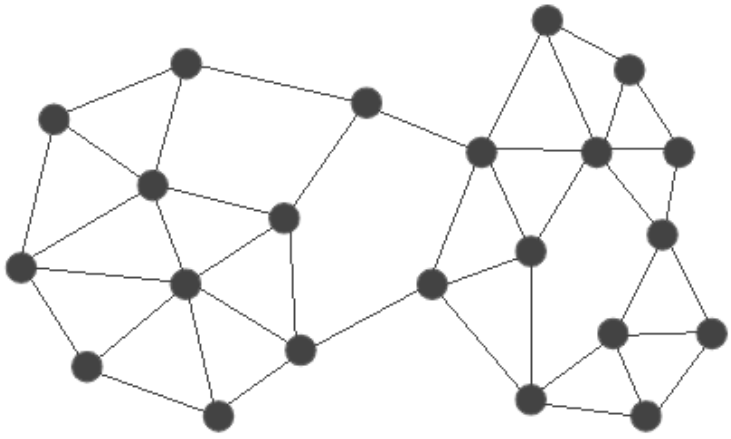

Рисунок 2.2 – Деревовидна К-зв’язана мережа антитіл, при К = 3

Слід зазначити, що в результаті формування деревовидної мережі деякі імунні об’єкти, що знаходяться у скупченнях антитіл, можуть мати велику кількість зв’язків, що перевищте значення, обмежене параметром К. Це відбувається тому, що таке антитіло самостійно формує К зв’язків та приймає ще додаткові зв’язки від інших антитіл, які знаходяться в безпосередній близькості відповідно до аффінності до нього.

Таким чином завдяки скороченню кількості зв’язків між антитілами мережі, з початкового графу з повними зв’язками між об’єктами, формується деревовидна структура мережі антитіл, з якої в подальшому відбувається виділення центрів кластерів. Слід зазначити, що в результаті кластеризації різні моделі та алгоритми формують кластери різного виду. Найбільш поширеними видами кластерів є кластери з чітким визначеним центром, в яких у якості центру кластеру обирається один з сукупності об’єктів, що досліджуються алгоритмом кластеризації. Деякі методи кластеризації передбачають формування кластерів центром, що формується незначною кількістю об’єктів, що мають подібні ознаки. Методи кластеризації, що функціонують на основі математичного апарату теорії графів, передбачають формування кластерів-стрічок, які не мають чітко визначеного центру кластеру, та формуються з безлічі пов’язаних між собою об’єктів. Запропонований метод кластеризації на першому етапі роботи формує кластери із визначеним центром, яким є одне з антитіл деревовидної мережі. Завдяки невеликій кількості зв’язків між імунними об’єктами К-пов’язаної деревовидної імунної мережі, визначення початкових центрів кластерів відбувається значно скоріше, ніж при використанні графу, де кожна вершина має безліч зв’язків з іншими вершинами.

Після формування К-пов’язаної мережі антитіл відбувається обчислення рівня стимуляції цього антитіла завдяки аффінностям з іншими імунними об’єктами, що формують мережу. Визначення рівня стимуляції відбувається за формулою 2.2 [16-17]:

$$s_i = \frac{1}{K}\sum_{j=1}^{K} aff\left(ab_i, ab_j\right), \tag{2.2}$$

де s_i – значення рівня стимуляції антитіла пов’язаними з ним антитілами;

K – параметр, що визначає кількість зв’язків антитіл у мережі.

На основі значення рівня стимуляції антитіла відбувається визначення кандидатів в центри кластерів. На цьому етапі формування деревовидної імунної мережі формується множина антитіл, які характеризуються високими рівнями стимуляції та кількістю зв’язків, що перевищує зазначений раніше параметр K. Отриманий таким чином набір антитіл формує сукупність об’єктів-кандидатів у центри кластерів. Слід зазначити, що при цьому кількість таких обраних клітин-кандидатів в центри кластерів може перебільшувати кількість кластерів, визначену на старті алгоритму.

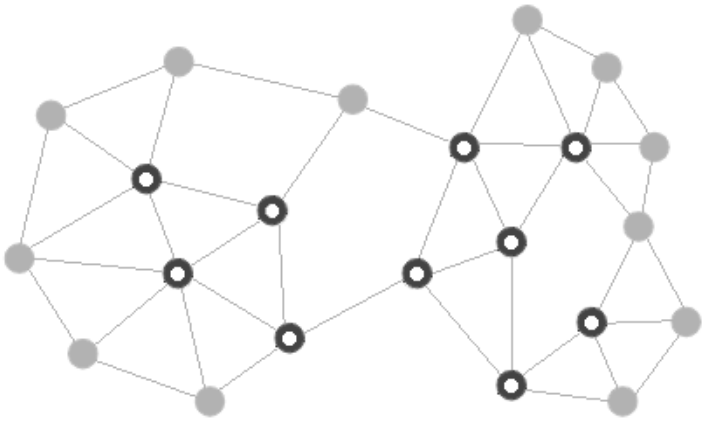

Рисунок 2.3 – Визначення антитіл з високим рівнем стимуляції

На зображенні 2.3, що відображає результат вибору антитіл-кандидатів в центри кластерів, об’єкти з мінімальним рівнем стимуляції, які не можуть бути обрані у якості центрів кластерів, позначені світло-сірим кольором, а

антитіла, які характеризуються великою кількістю міцних афінних зв’язків та високим рівнем стимуляції – відображаються білими вершинами и темно-сірим контуром. Таким чином у процесі обрання потенційних центрів кластерів не приймають участі антитіла, які знаходяться на периферії штучної імунної мережі, що призводить до підвищення швидкості роботи алгоритму кластеризації об’єктів.

Виділення центрів кластерів відбувається за допомогою параметру, який регламентує кількість кластерів, яка має бути виділена в результаті кластеризації. У цьому процесі також приймає участь значення порогової аффінності NAT, яке визначається на першому етапі формування імунної мережі. Під час вибору центрів кластерів з усіх антитіл-кандидатів у центри обирається антитіло, яке характеризується максимальною кількістю зв’язків з іншими антитілами мережі та максимальним рівнем стимуляції. Таке антитіло буде обрано у якості центра першого кластеру. Слід зазначити, що центри інших кластерів визначаються відносно обраного центру першого кластеру. Відповідно до цього, для того, щоб об’єкт-кандидат у центри кластеру міг бути обраним у якості центру нового кластеру, його аффінності із іншими антитілами-центрами кластерів, що були визначені раніше, мають не перевищувати значення порогової аффінності мережі NAT. У випадку, якщо аффінність між антитілом-кандидатом у центри кластеру перевищує порогове значення NAT хоча б з одним з обраних раніше центрів – це антитіло виключається із множини об’єктів-кандидатів у центри кластерів. Завдяки цьому, зі всього набору потенційних центрів кластерів обирається тільки невелика кількість найбільш далеких один від одного антитіл, які характеризуються між собою слабким афінним зв’язком.

На зображенні 2.4 представлено результат відбору центрів кластерів для деревовидної імунної мережі, що має k зв’язків. При цьому усі афінні зв’язки між антитілами мережі були позначені світло-сірим кольором, а об’єкти-кандидати у центри кластерів позначені темно-сірим кольором, об'єкти, обрані у якості центрів сформованих кластерів, позначені зеленим

кольором. Слід зазначити, що в цьому випадку в процесі кластеризації очікуються формування двох кластерів

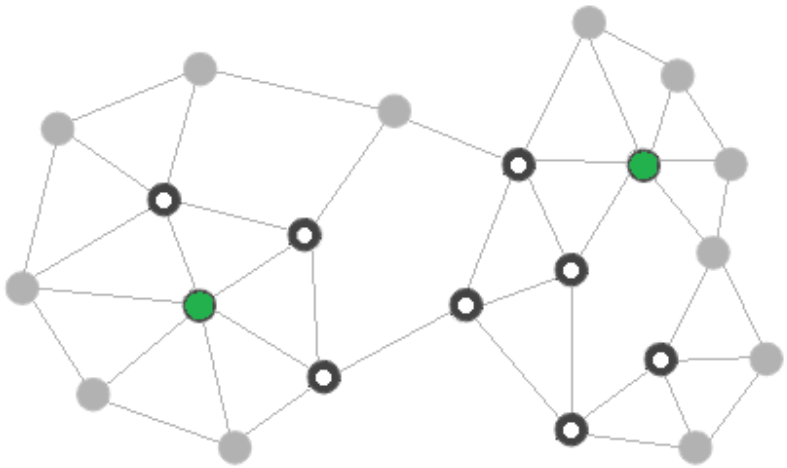

Рисунок 2.4. – Результат вибору центрів кластерів

Після розподілу центрів кластерів відбувається процес визначення приналежності до них антитіл імунної мережі. Цей процес потребує декількох етапів. На першому етапі кластеризуються імунні об’єкти, які характеризуються сильним афінним зв’язком з антитілами, які є центрами кластерів, тобто об’єкти, які мають прямі зв’язки з кластерами в деревовидній імунній мережі з K зв’язками. Слід зазначити, що на цьому етапі процес імунної модифікації мережі, який супроводжується клонуванням, мутацією та регуляцією популяції згрупованих антитіл не запускаються. Це пов’язано з високою обчислювальною складністю та значними тимчасовими витратами, які неминуче виникають при проведенні цих імунних процесів. Тому задля підвищення швидкості кластеризації, всі антитіла, що мають сильний афінний зв'язок к одним з обраних центрів кластерів, приєднуються до того ж кластеру, з центром якого вони пов’язані. Таким чином, після кластеризації антитіл, які наближені до центрів кластерів, утворюються групи, що складаються з обмеженого набору імунних об’єктів. Кластери цього виду також називаються кластерами з центрами

сильного згущення. Для проведення кластеризації інших антитіл, які не характеризуються міцним по аффінності зв’язком з жодним центром сформованих кластерів, відбувається визначення авідностей до кожного з кластерів та його антитіл, що формують центр сильного згущення. Визначена таким чином авідність відображає рівень міцності імунного зв’язку між об’єктами кластеру та цим антитілом. Поняття авідності між антитілами та антигенами, або між антитілами імунної мережі часто використовується в моделях штучних імунних мереж, які використовуються для вирішення задач класифікації, кластеризації та аналізу даних.

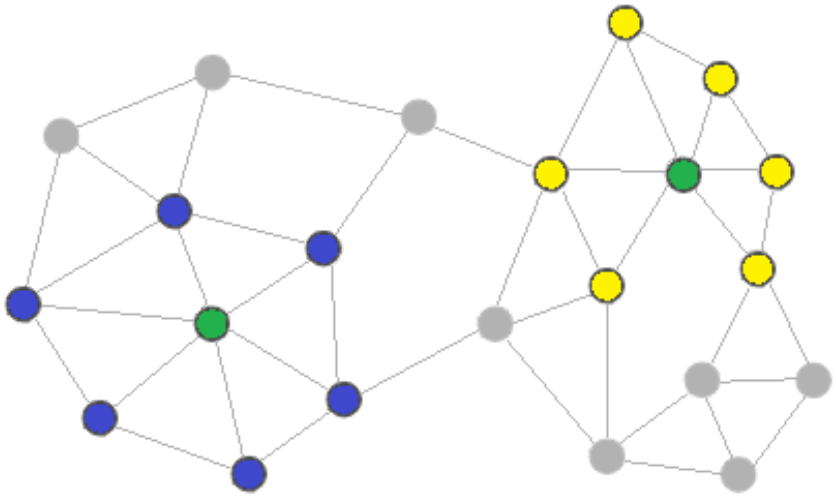

Рисунок 2.5 – Кластеризація об’єктів, пов’язаних з центрами кластерів

На рисунку 2.5. відображається результат групування імунних об’єктів, що мають прямий сильний афінний зв’язок з антитілами, обраними в якості центрів утворених кластерів у деревовидній імунній мережі. Зеленим кольором відзначаються антитіла-центри кластерів, синім та жовтим кольорами відзначаються антитіла, пов’язані із двома різними центрами кластерів, решта антитіл відображаються світло-сірим кольором.

Визначення авідності базується на спорідненості та аффінності між імунними об’єктами. У цьому випадку авідність антитіл з іншими антитілами, що належать до одного кластера, визначається як сума

аффінностей між ними:

$$av_i = \sum_{j=1}^{m} aff\left(ab_i, ab_j\right), \tag{2.3}$$

де av_i – значення авідності антитіла з іншими антитілами кластеру;

m – кількість антитіл у кластері;

$aff\left(ab_i, ab_j\right)$ – аффінність між антитілами одного кластера.

Після визначення авідностей між антитілами, що утворюють центр сильного згущення кластеру, ці значення усереднюються. Отримана таким чином середня авідність визначає кластер та буде використовуватися при кластеризації інших об’єктів деревовидної імунної мережі, що не мають безпосереднього сильного імунного зв’язку із центром будь-якого кластеру.

Кластеризація інших об’єктів відбувається шляхом запуску імунних процесів в деревовидній імунній мережі антитіл завдяки операторам клонування, мутації та використанню супресії клонів та мережі антитіл, що не належать до жодного кластеру. При цьому для кожного клону після його мутації відбувається визначення звідностей з цільовими об’єктами, якими є кластеризовані антитіла що формують кластери з центром сильного згущення. Під час відбору для кожного клонованого антитіла зі всієї множини його клонів обирається один об’єкт, що характеризується максимальною авідністю з антитілами, що формують центр кластеру сильного згущення. Цей клон замінює собою антитіло, від якого він був створений під час роботі оператору супресії популяції антитіл з невизначеною приналежністю до кластеру. Слід зазначити, що імунний процес клонування, мутації, супресії клонів та супресії антитіл деревовидної імунної мережі завершиться у випадку, коли у мережі не залишиться жодного антитіла, що має авініть до одного з центрів кластеру, яка буде меншою за авідність у кластерах з центрами сильного згущення. Таким

чином, процес кластеризації об’єктів, які не приймають участі у формуванні центрів сильного згущення, та не можуть бути обрані у якості центру нового кластеру, відбувається шляхом процесу імунної саморегуляції мережі антитіл та визначається шляхом оптимізації звідностей поколінь антитіл. При цьому такі антитіла створюються завдяки використанню імунних операторів, які визначають формування покоління антитіл на ітерації процесу імунної саморегуляції деревовидної мережі антитіл. На рисунку 2.6 наведено результат кластризації штучної імунної деревовидної мережі антитіл, розподілених між двома визначеними кластерами.

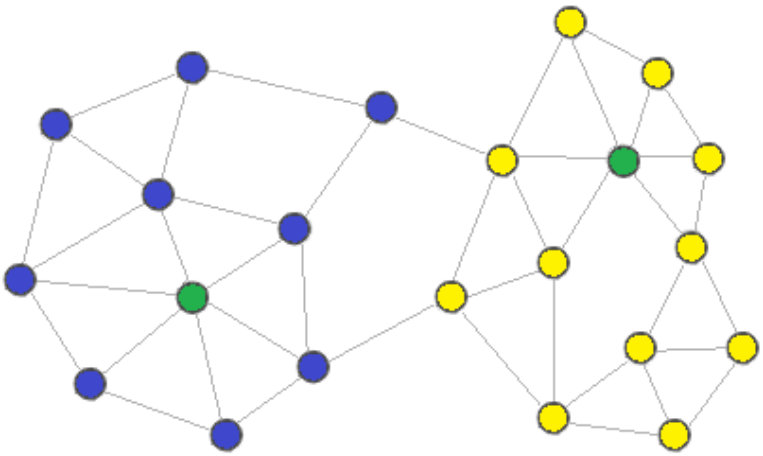

Рисунок 2.6 – Результат кластеризації мережі антитіл

У процесі саморегуляції деревовидної імунної мережі, що має K зв’язків між антитілами значну роль відіграє організація роботи імунних операторів клонування, мутації, редагування утворених антитіл, додаткового формування антитіл, супресії клонів та супресії мережі. Саме ці оператори забезпечують перетворення ознак об’єктів, що кластеризуются в процесі імунної саморегуляції до ознак цільових об’єктів.

Оператор клонування використовується для додавання нових об’єктів (клонів) до антитіл, які є точними копіями клонованих об’єктів. Існує кілька основних підходів до організації роботи оператора клонування,

які зазвичай використовуються у різних імунних моделях [15-17]:

– статичне клонування;

– пропорціональне клонування;

– зворотно-пропорційне клонування.

За допомогою статичного клонування кількість утворених клонів є фіксованим значенням і не залежить від спорідненості клонованих антитіл з цільовим об’єктом. Відповідно до цього, кількість створених клонів є вхідним аргументом імунного методу і не змінюється під час роботи. Такий аргумент може бути або цілим постійним, або деякою функцією загальної кількості антитіл, виражених у відсотковому співвідношенні. Цей метод організації клонування характеризується простотою та надійністю функціонування. Основним його недоліком є низька швидкість. Це пов’язано із тим, що антитіла з мінімальними спорідненістю на цільові об’єкти утворюють таку ж кількість клонів, що й антитіла, які характеризуються максимальними афінностями, що тягне за собою збільшення витрат часу з мутацією утворених клонів та організацією їх відбору.

З пропорційним клонуванням кількість утворених клонів залежить від спорідненості клонованих антитіл з цільовим об’єктом і визначається наступним чином [15-17]:

Слід зазначити, що максимально допустима кількість об'єктів у виразі 2.4 зазвичай є загальною кількістю антитіл або антигенів. Основна перевага цього підходу полягає у збільшенні швидкості імунного навчання, оскільки антитіла, що характеризуються максимальними афінностфми, утворюють максимальну кількість клонів, що призводить до збільшення швидкості відновлення ознак антигенів під час мутації.

Основним недоліком такого підходу є зменшення кількості сформованих клонів під час роботи з невеликими наборами об’єктів. Таким чином, якщо загальна кількість антитіл, що використовується як максимальна кількість об’єктів має невелике значення, ті кількість клонів буде ще меншою.

$$Nc_i = M \cdot aff\left(ab_i, ag_j\right), \tag{2.4}$$

де Nc_i – кількість клонів, що формується антитілом;

M – максимальна можлива кількість об’єктів;

$aff\left(ab_i, ag_j\right)$ – аффінність між антитілом та цільовим об’єктом.

Таким значення клонів, то кількість утворених клонів буде обмежена десяток об'єктів з максимальними афінностіми клонованих антитіл, що може призвести до збільшення кількості антитіл при формуванні наступної популяції.

У випадку зворотно-пропорційного клонування кількість утворених клонів має пропорційну залежність від спорідненості клонованих антитіл з цільовим об'єктом, однак ця залежність є пропорційною. Відповідно до цього, кількість клонів, утворених за допомогою цього методу клонування, визначається формулою 2.5:

$$Nc_i = M \cdot \left(1 - aff\left(ab_i, ab_j\right)\right) \tag{2.5}$$

Основною перевагою цього способу організації процесу клонування є формування великої кількості клонів, незалежно від максимально імовірної кількості імунних об’єктів мережі антитіл. Головним недоліком такого підходу є уповільнення швидкості імунного навчання та саморегуляції мережі, тому що антитіла, які характеризуються мінімальними афінностіми, формують максимальну кількість клонів. Це в подальшому має негативний вплив на швидкість роботи імунного алгоритму.

Оператор мутації відіграє важливу роль в роботі імунних методів та алгоритмів. Робота оператору мутації полягає у внесенні змін в ознаки клона, що підлягає мутації задля відтворення ознак цільових об’єктів та підвищення афінного зв’язку з ними. Завдяки мутації достягається стан специфічності клонів антитіл з одним із цільових об’єктів або антигенів. На сьогодні існує

декілька видів оператору мутації [15-17]:

– статична мутація;
– пропорційна мутація;
– звоторньо-пропорційна мутація;
– випадкова мутація.

Головна різниця поміж наведеними видами оператора мутації полягає у способі визначення коефіцієнту мутації [15-17]. Цей коефіцієнт мутації є значенням, яке використовується під час зміни ознак об’єкту, який підлягає мутації. Зазвичай таким об’єктом є клон.

При використанні статичного оператора мутації коефіцієнт мутації визначається константним значенням та є вхідним аргументом імунного метода [15-17]. Тому цей коефіцієнт мутації є незмінним впродовж роботи алгоритму та використовується для всіх об’єктів, незалежно від показників їх аффінностей з цільовими об’єктами. Головними перевагами такого підходу є можливість визначення кількості популяцій антитіл, необхідного для досягнення стану повної специфічності популяцій, а також простота реалізації імунного алгоритму дослідження даних. Основним недоліком цього підходу є збільшення часу, необхідного на проведення імунного самонавчання або процесу саморегуляції імунної мережі при мінімальних значеннях статичного коефіцієнту мутації. Також серед недоліків цього способу визначення коефіцієнту мутації є неможливість досягнення стану повної специфічності між популяцією антитіл на набором цільових об’єктів при збільшенні коефіцієнту мутації.

При використанні оператора випадкової мутації коефіцієнт мутації визначається випадковим чином у діапазоні від мінімуму до максимуму значення аффінності. Таким чином, коефіцієнт мутації не залежить від характеристик клонів, що підлягають змінам або ознак їх батьківських антитіл. Головною перевагою цього способу організації роботи оператора мутації є простота реалізації. Головним недоліком даного способу мутації є низька швидкість імунного навчання та процесу саморегуляції мережі.

При використанні оператору пропорційної мутації, значення коефіцієнту мутації прямо пропорційно значенню аффінності батьківського антитіла із цільовим об’єктом або набором цільових об’єктів. При цьому, відповідно до [15-17], значення коефіцієнта мутації визначається за допомогою виразу 2.6:

$$\mu = rand[0;\ aff(ab, AG)], \tag{2.6}$$

де μ – коефіцієнт мутації;

$aff(ab, AG)$ – аффінність між батьківським антитілом та набором цільових об’єктів.

Використання пропорційної мутації підвищує швидкість імунного навчання та процесу самогеруляції деревовидної імунної мережі. Це відбувається завдяки підвищенню імовірності досягання стану повної специфічності об’єкту що підлягає мутації одному зі своїх цільових об’єктів. Основним недоліком даного методу є мінімальні зміни ознак клонів в процесі мутації, які характеризуються невеликими значеннями аффінностей до своїх цільових об’єктів. Відповідно до цього при використанні такого типу мутації імовірне збільшення кількості популяцій клонів, необхідного алгоритму для досягнення стану повної специфічності з одним з цільових об’єктів.

При використанні оператора зворотньо-пропорційної мутації коефіцієнт мутації зазвичай визначається у зворотній пропорції до значення аффінності між батьківським антитілом та його цільовим об’єктом або набору таких цільових об’єктів. Відповідно до [16] значення коефіцієнту мутації визначається за формулою 2.7:

$$\mu = rand\left[0;1 - aff\left(ab, AG\right)\right] \tag{2.7}$$

Основною перевагою такого підходу є підвищення імовірності того, що

в процесі мутації один з клонів достягне стану повної специфічності з одним зі своїх цільових об’єктів. Це відбувається завдяки тому, що об’єкти, які характеризуються мінімальними аффінностями до цільових об’єктів, підлягають найбільшої кількості змін. Тому використання такого оператору мутації призводить до підвищення швидкості імунного навчання алгоритму або процесу саморегуляції штучної імунної мережі порівняно з іншими наведеними видами операторів мутації.

Задля редагування популяції клонів після мутації та популяції антитіл після формування штучної імунної мережі, різні імунні моделі використовують різні підходи. Найбільш поширеними методами редагування популяції антитіл та популяції клонів є клональний відбір, мережева супресія або негативний відбір. При цьому оператор клонального відбору функціонує аналогічно оператору первинного відбору антитіл. Різниця між цими операторами полягає тільки в типі об’єктів, що підлягають редагуванню. Оператор первинного відбору редагує популяцію вихідних антитіл, відбираючи об’єкти, які характеризуються найкращими аффінностями до цільових об’єктів чи об’єктів навчальної вибірки, для їх подальшого клонування. Оператор відбору клонів редагує популяцію клонів, змінених після мутації, для їх подальшого поєднання із популяцією антитіл та підвищення міжпопуляційної аффінності. В даний час існує кілька видів оператора відбору клонів [16]:

– статичний відбір;

– критеріальний відбір.

При використанні оператора статичного відбору клонів кількість об’єктів, які виживають в результаті редагування популяції, обмежується деяким наперед заданим значенням, яке є вхідним аргументом імунного алгоритму [15-17]. При цьому після визначення аффінності між цільовими об’єктами та клонами в результаті роботи оператора клонального відбору в популяцію антитіл додаються тільки клони з максимальними аффінностями до цільових об’єктів. Використання такого підходу до організації відбору

клонів підвищує швидкість обробки популяції антитіл під час імунного навчання або процесу саморегуляції імкнної мережі, а також спрощує реалізацію алгоритму. Головним недоліком цього підходу є підвищення імовірності видалення клонів з високими аффінностями до цільових об’єктів, тому що відбір проходить лише невелика кількість клонів. Наслідком цього є збільшення кількості популяцій антитіл, які необхідно сформувати під час імунного навчання алгоритму.

Критеріальний відбір клонів відрізняється від статичного тим, що в ньому замість заданої кількості об’єктів використовується деяке граничне значення аффінності, яке і є критерієм відбору [15-17]. При цьому клони, у яких аффінності до цільових об’єктів перевищують значення граничної афінності, проходять відбір та надходять у популяцію антитіл. В іншому випадку, якщо аффінності клонів до цільових об’єктів є меншими за порогову аффінність – відбувається видалення даних клонів. Слід зазначити, що застосування критеріального відбору призводить до прискорення процесу імунного навчання, оскільки при кожному відборі виживають усі об’єкти, які характеризуються максимальними аффінностями, а не їх частина, як це відбувається при використанні оператору статичного відбору клонів.

Використання оператора супресії притаманно методам, які функціонують на основі моделі штучної імунної мережі [15-17]. При цьому, на відміну від оператора відбору клонів під час супресії, клони можуть взаємодіяти з антитілами та між собою. Окрім того дії даного оператора на різних етапах навчання та саморегуляції мережі піддаються як клони, так і антитіла. Відповідно до цього оператори супресії поділяють на дві основні групи: внутрішньопопуляційні та міжпопуляційні. При роботі внутрішньопопуляційних операторів супресії відбувається аналіз взаємодії об’єктів між собою у межах однієї популяції. Наприклад, при використанні внутрішньопопуляційного оператора супресії для редагування клонів, всі клони визначають аффінності між собою та на основі деякого критерію відбору, відбувається скорочення цієї популяції клонів. При використанні

міжпопуляційного оператора супресії редагування популяції клонів відбувається на основі аффінності з набором цільових об’єктів, який може формуватися як з антигенів, так і з антитіл. Така організація аналогічна роботі оператора клонального відбору. Таким чином, використання оператора супресії для редагування популяцій імунних об’єктів дає більше можливостей, ніж застосування оператору клонального відбору.

У методі aiNET [15-17] для редагування популяції антитіл та клонів оператор супресії використовується тричі: для виділення клонів із найкращими афінностями до антигенів; для відбору клонів; для редагування антитіл імунної мережі Окрім того, в методі aiNET використовується оператор додаткового розкиду, який формує антитіла випадковим чином замість антитіл, видалених під час супресії. При цьому імовірність створення антитіла специфічного будь-якому цільовому об’єкту вкрай низька. Наслідком цього є зниження швидкості імунного навчання та падіння точності групування об’єктів. Отже у запропонованому методі кластеризації оператора додаткового розкиду антитіл немає.

2.2 Загальна схема роботи імунного алгоритму кластеризації даних

Процес кластеризації набору вхідних об’єктів можна як послідовності наступних етапів:

1) встановлення параметрів кластеризації, та отримання початкового набору об’єктів;
2) визначення шкал ознак досліджуваних об’єктів;
3) формування K-зв’язної деревовидної імунної мережі;
4) виділення центрів кластерів;
5) визначення авідностей кластерів;
6) імунна саморегуляція мережі;
7) висновок про кластеризацію об’єктів.

Оскільки найпоширенішою моделлю функціонування штучної імунної

мережі є aiNET, цей метод є основним для створення імунних моделей, що використовують принципи роботи штучної імунної мережі. Незважаючи на свою універсальність і можливість саморегуляції та навчання, більшість імунних методів не можуть використовуватися без додаткових модифікацій при вирішенні завдань класифікації та кластеризації об’єктів. Задля підвищення швидкості імунного навчання в запропонованому імунному методі модифікації зазнали основні імунні оператори: оператор первинного відбору, оператор мутації, оператори супресії клонів та супресії антитіл. Для підвищення швидкості класифікації даних, у роботу імунного методу aiNET додані оператори шкалювання даних та формування K-зв’язної деревовидної мережі антитіл, оператор визначення початкових центрів кластерів, а також оператор визначення авідностей, що використовується на завершальному етапі групування об’єктів при кластеризації антитіл, що не формують центри сильного згущення кластерів. Слід зазначити, що оператор шкалювання не є необхідним і кластеризація даних може відбуватися і без його використання, проте застосування даного оператора призводить до підвищення швидкості мережевої саморегуляції та, як наслідок, призвести до підвищення швидкості кластеризації. Відповідно до цього, на рівні імунних операторів модифікований метод Dendric-aiNet представляється наступним у виразі 2.8.

У цьому виразі *Dendric-aiNet(AB,K,C)* є умовним позначенням методу кластеризації вхідних об’єктів *AB* при використанні деревоподібної *K*-зв’язної імунної мережі антитіл та критерію *c*, що використовується для вказівки кількості кластерів, що формуються методом набору об’єктів.

Етап підготовки до кластеризації позначається скороченням *PRP* і містить кілька операторів: оператор *Scaling(AB)*, що використовується для проведення шкалювання об’єктів; оператор *Presentation(AB)*, використовуваний визначення аффінностей між антитілами формованої імунної мережі; оператор *NATCalculation(AB)*, використовуваний визначення значення порогової аффінності *NAT* в популяції антитіл.

$$Dendric-aiNet\left(AB,K,c\right)=\overset{PRP}{\left[\begin{array}{l} Scaling\left(AB\right)\rightarrow \\ Presentation\left(AB\right)\rightarrow \\ NATCalculation\left(AB\right) \end{array}\right]}\Rightarrow$$

$$\Rightarrow\overset{DKN}{\left[\begin{array}{l} DKnetCreation\left(AB,K\right)\rightarrow \\ CalcStimulation\left(AB\right)\rightarrow \\ CentersSelection\left(AB,c,NAT\right)\rightarrow \\ DendricClustering\left(AB'\right) \end{array}\right]}\Rightarrow \qquad (2.8)$$

$$\Rightarrow\overset{NET}{\left[\begin{array}{l} Cloning\left(AB'',CL\right)\rightarrow \\ Mutation\left(CL\right)\rightarrow \\ Presentation\left(CL,AB',AB''\right)\rightarrow \\ CLSupression\left(CL,AB',AB''\right)\rightarrow \\ NetSupression\left(CL,AB''\right)\rightarrow \\ AvCalculation\left(AB''\right)\rightarrow \\ ClusterSelection\left(AB''\right) \end{array}\right]},$$

Етап роботи, спрямований на формування *K*-зв'язної імунної мережі має умовне позначення *DKN* і містить такі оператори: оператор *DKnetCreation(AB, K)*, застосовуваний формування *K*-зв'язної деревовидної мережі антитіл; оператор *CalcStimulation(AB)*, застосовуваний визначення рівня стимуляції антитіл; оператор *CentersSelection(AB, c, NAT)*, використовуваний виділення центрів кластерів; оператор $DendricClustering\left(AB'\right)$, що використовується для формування кластерів сильного згущення.

Етап мережевої взаємодії має позначення *NET* і містить наступні оператори: оператор клонування $Cloning\left(AB'',CL\right)$, що використовується для поширення популяції антитіл, не пов'язаних з жодним формованим кластером; оператор мутації *Mutation(CL)*, використовуваний зміни ознак клонів; оператор подання цільових об'єктів $Presentation\left(CL,AB',AB''\right)$, що

використовується для визначення аффінностей між клонами та об’єктами, що формують кластери сильного згущення; оператор супресії клонів $CLSupression(CL, AB', AB'')$, що використовується для редагування популяції клонів; оператор супресії мережі антитіл $NetSupression(CL, AB'')$, що використовується задля скорочення кількості об’єктів, що оперуються; оператор визначення авідностей між антитілами та кластерами сильного згущення $AvCalculation(AB'')$, що використовується для розподілу не кластеризованих об’єктів між кластерами, що формуються, та оператор $ClusterSelection(AB'')$, який використовується задля визначення належності антитіл кластерам.

Сформований метод класифікації Dendric-aiNet, відповідно до формального представлення на рівні імунних операторів (формула 2.8) складається з наступних основних кроків.

Крок 1. Підготовчий етап:

– проведення шкалювання для популяції імунних об’єктів;
– презентація антитіл один одному;
– визначення порога афінності NAT для популяції антитіл.

Крок 2. Побудова деревовидної *K*-зв’язної імунної мережі:

– виділення *K*-зв’язків із максимальними аффінностями у мережі;
– визначення рівнів стимуляції антитіл;
– виділення центрів кластерів на основі рівнів стимуляції та NAT;
– формування кластерів сильного згущення.

Крок 3. Саморегуляція та кластеризація антитіл, цикл операторів:

– клонування антитіл, які не визначили належність до кластеру;
– мутація сформованих клонів;
– презентація кластерів сильного згущення популяції клонів;
– супресія клонів;
– супрсія клонованих антитіл;
– визначення авідностей для антитіл з не визначеним кластером;

– визначення приналежності кластерам на основі авідностей.

Внаслідок роботи Dendric-aiNet відбувається кластеризація початкової популяції антитіл. При цьому дані об’єкти належать до одного з виділених у процесі імунної взаємодії деревовидної K-зв’язної мережі антитіл. Після цього проведення етапу визначення авідностей дозволяє провести кластеризацію антитіл та кластерами сильного згущення.

2.3 Особливості роботи імунних операторів

Оскільки використання методу aiNET без проведення змін у роботі його операторів при вирішенні завдання кластеризації важко, для усунення цієї проблеми, організація функціонування методу aiNET була піддана істотній модифікації. В результаті модифікації даного методу відповідно до формули 2.8, був сформований метод Dendric-aiNet, який може використовуватися для вирішення задач кластеризації. У цьому перетворенні зазнали як імунні оператори, а й деякі принципи функціонування aiNET.

Оператор презентації антитіл *Presentation(AB)* використовується на початковому етапі роботи алгоритму кластеризації. При цьому з набору вихідних об’єктів, для яких проводиться кластеризація, формується початкова популяція антитіл. Після формування популяції антитіл кожне антитіло визначає аффінності до всіх інших антитіл деревовидної імунної мережі, що формується в процесі роботи алгоритму. Використання аффінності дозволяє формувати імунну мережу антитіл для подальшої кластеризації.

Оператор *NATCalculation(AB)* застосовується для обчислення значення порогової афінності *NAT*, яка використовується на різних етапах формування *K*-зв’язної деревоподібної імунної мережі та визначення початкових центрів кластерів, на основі яких формуватимуться кластери сильного згущення. Значення *NAT* визначається як середнє значення аффінності для всієї популяції імунних об’єктів.

Оператор *DKnetCreation(AB, K)* дозволяє формувати мережу антитіл з обмеженою кількістю зв'язків між імунними об'єктами. Основним показником міцності зв'язку між антитілами є значення аффінності між імунними об'єктами. Однак, якщо кожне антитіло мережі матиме дуже велику кількість зв'язків – робота імунної мережі буде здійснюватися дуже повільно, тому оператор передбачає використання зовнішнього параметра *K*, що обмежує кількість зв'язків між антитілами. Слід зазначити, що при цьому допускаються ситуації, за яких у одного антитіла кількість зв'язків може перевищувати значення *K*. Так відбувається через те, що дане антитіло виділяє *K* зв'язків з найближчими до нього імунними об'єктами, але в мережі можуть існувати антитіла, що не входять до цієї групи об'єктів, які, тим не менш, зберігають зв'язок з цим антитілом. Внаслідок роботи даного оператора відбувається формування *K*-зв'язної мережі антитіл для подальшого визначення початкових центрів кластерів, а також формування кластерів сильного згущення.

Оператор *CalcStimlation(AB)* застосовується для визначення рівня стимуляції антитіл *K*-зв'язної деревовидної імунної мережі на основі виразу 2.2. Використання рівнів стимуляції дозволяє виділити з початкової популяції антитіл імунні об'єкти, які характеризуються великою кількістю сильних аффінності зв'язків з антитілами мережі. Таким чином, антитіла імунної мережі стимулюють ці імунні об'єкти та групуються навколо них. З цих груп антитіл згодом формуються кластери сильного згущення.

Оператор *CentersSelection(AB, c, NAT)* використовується формування кластерів сильного згущення. При цьому спочатку відбувається пошук початкових центрів кластерів, після чого формуються кластери з центром, потім всі імунні об'єкти, пов'язані з даним центром *K*-зв'язної деревовидної імунної мережі, формують кластери з центром сильного згущення. Визначення центрів кластерів починається із пошуку центру першого кластера. При цьому центрами кластерів можуть бути лише імунні об'єкти, які характеризуються максимальною кількістю зв'язків, що перевищує

кількість *K*. Ці антитіла сортуються за значенням рівня стимуляції за зростанням. Центром першого кластера є антитіло, яке має максимальний рівень стимуляції та кількість зв’язків з іншими антитілами у *K*-зв’язній деревоподібній імунній мережі, що перевищують значення *K*, яке є вхідним параметром алгоритму кластеризації. Після визначення центру відбувається пошук центрів інших кластерів. Слід зазначити, що метод Dendric-aiNet передбачає, що кількість кластерів, між якими відбувається розподіл вхідного набору об’єктів, що формують *K*-зв’язкову імунну мережу, є вхідним параметром алгоритму. Інші центри кластерів визначаються ітеративним способом. На першій ітерації відбувається пошук центру першого кластеру. На другій ітерації – центр другого кластера, у своїй він має характеризуватись високим рівнем стимуляції і навіть мати аффінність з центром першого кластера, менше, ніж значення *NAT*, що було визначено раніше. На наступній ітерації з антитіл, що залишилися, з високим рівнем стимуляції вибирається антитіло, що має аффінності з певними раніше центрами кластерів меншими, ніж значення *NAT*. Якщо жодне антитіло немає великої кількості зв'язків, але відповідає вимогам за значенням *NAT*, може бути обрано центром нового кластера.

Оператор $DendricClustering(AB')$ використовується задля формування кластера з центром сильного згущення кластерів з виділеним центром, яким є один з антитіл *K*-зв’язної деревоподібної імунної мережі. Цей етап роботи Dendric-aiNet здійснюється виключно завдяки афінним зв'язкам деревоподібної мережі без запуску мережевого механізму саморегуляції, тобто. без використання операторів клонування, мутації та редагування популяції клонів та антитіл мережі. Формування кластерів з центрами сильного згущення відбувається шляхом приєднання до кластерів антитіл, пов’язаних з їх центром безпосередньо.

Оператор клонування $Cloning(AB'', CL)$ використовується для формування популяції клонів, ідентичних клонованих антитіл для їх подальшої мутації. Слід зазначити, що підвищення кількості клонів, що

створюються, призводить до зниження швидкості імунного навчання внаслідок того, що кожен сформований клон піддається мутації і представляється цільовим об’єктам, що збільшує час імунного навчання алгоритму. Однак формування невеликої кількості клонів при клонуванні антитіл також може стати причинною збільшення часу роботи імунного алгоритму, оскільки процес досягнення специфічності антитіл цільовим об’єктам залежить від кількості формованих клонів і способу їх мутації. Таким чином, в результаті роботи оператора клонування населення формованих клонів для кожного антитіла не повинна дуже великою або дуже маленькою, т.к. це негативно вплине на час навчання алгоритму. Тому Dendric-aiNet використовується пропорційне клонування з посиленням аффінності, що дозволяє сформувати популяцію клонів оптимальним чином. Відповідно до цього кількість клонів, що формуються антитілами при клонуванні, визначається наступним чином:

$$Nc_i = M \cdot Cm \cdot \max aff\left(ab_i, ag_j\right), \tag{2.9}$$

де Nc_i – кількість клонів *i*-го антитіла;

M – загальна кількість антитіл;

Cm – коефіцієнт підсилення аффінності;

$\max aff\left(ab_i, ag_j\right)$ – максимальна аффінність з одним із цільових об’єктів.

При цьому слід зазначити, що коефіцієнт посилення аффінності є цілим значенням, використовується в Dendric-aiNet як вхідний аргумент і набуває значення в діапазоні [1;10].

Оператор мутації клонів *Mutation(CL)* використовується для внесення змін до ознак клонів, що мутують. Слід зазначити, що в методі класифікації, що моделюється, Dendric-aiNet використовується зворотно-пропорційна мутація з обмеженням нижнього порога визначення коефіцієнта мутації. На сьогоднішній день в області імунних операторів існує одна вада, характерна

для більшості існуючих операторів мутації. Цей недолік полягає у способі визначення коефіцієнта мутації. При цьому нижня межа цього коефіцієнта встановлюється рівною нулю. В результаті цього підвищується ймовірність того, що коефіцієнт мутації буде визначений як зневажливо мале значення, що призведе до збільшення кількості антитіл популяцій, необхідних для досягнення стану специфічності цільовим об'єктам. При мутації з обмеженням нижнього порога коефіцієнта мутації відбувається зміна нижньої межі діапазону можливих значень, що використовується для визначення даного коефіцієнта. Відповідно до цього відбувається встановлення залежності мінімального значення коефіцієнта мутації від аффінності батьківського антитіла клону, що змінюється, з цільовими об'єктами. Таким чином, визначення коефіцієнта мутації проводиться відповідно до наступного виразу:

$$\mu = rand\left[\frac{1}{2}\left(1 - aff\left(ab, AB\right)\right); 1 - aff\left(ab, AB\right)\right], \qquad (2.10)$$

де $aff\left(ab, AB\right)$ – аффінність між батьківським антитілом та набором цільових об’єктів або антигенів.

Завдяки даній модифікації оператора мутації досягається максимальна зміна ознак клонів залежно від зміни значення аффінності за мінімізації ймовірності втрати специфічності антигенів або інших цільових об’єктів.

Оператор $Presentation\left(CL, AB', AB''\right)$ є імунним оператором, що застосовується на етапі саморегуляції деревоподібної імунної мережі. Цей оператор використовується для представлення антитіл, які не встановили приналежності до того чи іншого кластера, антитіл, що формують центри сильного згущення кластерів, що виділяються. Отже, безлічі вільних, не класифікованих антитіл, представляються в повному обсязі інші антитіла деревоподібної мережі, лише їх невелика частина – антитіла, формують

кластери сильного згущення, виділені раніше. Під поданням антитіл мається на увазі визначення аффінності між імунними об'єктами деревоподібної мережі антитіл. Це призводить до суттєвого підвищення швидкодії та підвищення швидкості кластеризації імунних об’єктів.

Оператор спресії клонів $CLSupression(CL, AB', AB'')$ використовується для організації редагування популяції клонів, що мутували. В Dendric-aiNet використовується внутрішньопопуляційна супресія клонів. У цьому редагування популяції клонів складає основі їх аффінностей з цільовими об’єктами. Слід зазначити, що цільовими об'єктами при внутрішньопопуляційній супресії є антитіла, що формують центри кластерів сильного згущення. При цьому для кожного клону на початку супресії відбувається визначення рівнів стимуляції цільовими об’єктами на підставі аффінності, визначених на етапі представлення цільових об’єктів після мутації. Відбір клонів здійснюється шляхом зіставлення їх рівнів стимуляції, причому в результаті роботи оператора супресії клонів, у популяції залишається один об’єкт, що характеризується найкращим рівнем стимуляції. Завдяки такому способу організації супресії підвищується швидкість імунного навчання алгоритму без втрати точності класифікації антитіл.

Після формування імунної мережі та додавання до неї клонів, відібраних у результаті супресії, відбувається її редагування шляхом використання оператора мережевої супресії $NetSupression(CL, AB'')$. При цьому відбувається видалення клонованих антитіл внаслідок зіставлення їх рівнів стимуляції цільовими об'єктами. Відповідно до цього в модельованому імунному методі кластеризації Dendric-aiNet для редагування імунної мережі використовується критеріальний оператор супресії. При цьому критеріями, що регулюють процес мережевої супресії є рівні стимуляції клонованих антитіл. Якщо рівень стимуляції антитіла тим чи іншим центром кластера сильного згущення менший за відповідний рівень клону, сформований від даного антитіла в процесі клонування – дане антитіло видаляється і замінюється клоном. В іншому випадку відбувається видалення клону, а

антитіло залишається в популяції і знову піддається дії оператора клонування. Таким чином, після завершення роботи оператора мережевої супресії, в мережі залишаються лише антитіла з максимальними рівнями стимуляції до центрів згущення кластерів.

Після завершення роботи оператора мережевої супресії відбувається визначення авідностей між антитілами та виділеними центрами кластерів сильного згущення при використанні виразу 2.3. Для цього викликається імунний оператор $AvCalculation(AB'')$, робота якого завершується визначенням авідностей з усіма антитілами, які формують кластери сильного згущення.

Ітерація циклу саморегуляції деревовидної *K*-зв'язної імунної мережі завершується роботою оператора $ClusterSelection(AB'')$. У процесі роботи даного оператора кожне антитіло визначається приналежність до того чи іншого кластера на основі значення авідностей з антитілами, що формують центри сильного згущення кластерів і авідностей, визначених всередині даних кластерів. Висновок про належність антитіла тому чи іншому кластеру робиться у разі, якщо авідність між цим антитілом і кластером є не меншою, ніж авідність між антитілами центру сильного згущення кластера. Якщо антитіло визначає приналежність до одного з формованих кластерів, на них не поширюватиметься дія імунних операторів, що функціонують у рамках етапу саморегуляції імунної мережі. Якщо ж значення авідностей між антитілом і кластерами є меншим за звідність між антитілами всередині кластерів, дане антитіло залишається некласифікованим і знову піддаватиметься дії операторів клонування, супресії тощо.

Основною умовою завершення навчання є досягнення повної специфічності антитіл сформованим кластерам. Для досягнення цього стану імунного алгоритму потрібна велика кількість популяцій антитіл. При цьому визначення кількості популяцій, необхідних для досягнення повної специфічності математично неможливо, тому Dendric-aiNet використовується статичний критерій зупинки, що визначає максимальну кількість популяцій

антитіл, що формуються в процесі імунного навчання. В Dendric-aiNet використовується кілька умов припинення обробки антитіл. Першою умовою є вибір даного антитіла як центр нового кластера. У такому випадку це антитіло не піддається дії операторів клонування, мутації та супресії мережі, і залишається без змін до завершення роботи алгоритму кластеризації. Другою умовою є кластеризація антитіла завдяки використанню сильної афінності зв'язку з центром одного з кластерів, виділеної в процесі формування *K*-зв'язної деревовидної імунної мережі. Третьою умовою є специфічність антитіла кластеру. У такому випадку клонування антитіла припиняється. При цьому авідність між цим антитілом і антитілами, що формують центр сильного згущення даного кластеру, повинна бути не меншою за авідність антитіл кластеру.

3 ПРОГРАМНЕ СЕРЕДОВИЩЕ КЛАСТЕРИЗАЦІЇ ДАНИХ НА ОСНОВІ ДЕРЕВОВИДНОЇ ШТУЧНОЇ ІМУННОЇ МЕРЕЖІ

3.1 Формат даних при кластеризації

Спосіб представлення даних, які обробляються та моделюється алгоритмом, грає велику роль в організації процесу кластеризації. Визначення формату даних – один з важливих етапів у процесі розробки системи моделювання імунного алгоритму. Слід зазначити, що кількість ознак, що описують об’єкти в багатовимірному просторі, дуже впливає на час кластеризації і точність групування. Це обумовлюється тим, що збільшення кількості ознак об’єктів, що досліджуються, призводить до зниження швидкості процесу мережевої саморегуляції за рахунок збільшення кількості обчислювальних операцій визначення аффінності між імунними об’єктами. Крім того, за результатами проведення експериментальних випробувань різних імунних та методів кластеризації, можна зробити висновок про те, що точність угруповання об’єктів при використанні імунного підходу обернено-пропорційна кількості ознак, що характеризують безліч об’єктів.

На сьогоднішній день серед найбільш поширених форм представлення об’єктів при кластеризації є:

– вектор ознак (масив);
– матриця ознак (таблиця);
– зважений масив векторів ознак.

При використанні векторного формату представлення даних всі ознаки, що характеризують об’єкт, поміщаються в один вектор (масив). Слід зазначити, що такий формат даних використовується у разі однорідності ознак, тобто. ознаки, що описують об’єкт, мають однаковий тип даних та загальний діапазон допустимих значень. У разі недотримання цієї вимоги внаслідок визначення відстані між об’єктами виникають неоднозначності чи помилки. З цього випливає, що якщо об’єкт характеризується різними

типами ознак, наприклад, координатами та кольором, то ці ознаки не можуть поміщатися в один вектор. У разі якась група ознак (координати чи колір) нічого очікувати враховуватися щодо відстані між об'єктами. Таким чином, кластеризація об’єктів, описаних таким способом, може відбуватися лише за однією групою ознак. Тому векторний формат представлення даних використовується лише у випадках, коли ознаки однорідні між собою.

При використанні матричного формату представлення даних, всі ознаки, що характеризують об’єкт, поділяються на однорідні групи, що мають однаковий діапазон допустимих значень. Після цього з цих груп формується матриця ознак, що описує цей об’єкт. Використання матричного формату даних щодо відстані між об'єктами дозволяє вирішити проблему, що виникає під час роботи з вектором ознак. При цьому між кожним рядком матриць, що описують взаємодіючі об'єкти, визначається аффінність на підставі ознак, що входять до групи. Визначення загальної аффінності між двома матрицями імунних об'єктів полягає у додаванні та нормуванні окремих аффінностей між відповідними рядками ознак, що містяться в даних матрицях. Тому найбільш матричний формат даних набув широкого використання при моделюванні різних алгоритмів. Слід зазначити, що матричний формат має низку особливостей:

– нормування груп ознак для формування матриці;
– відсутність ваг у рядків матриці.

Нормування груп ознак полягає у визначенні кількості стовпців матриці, що формується, за максимальною кількістю ознак, що містяться в одному з її рядків. Таким чином, якщо матриця ознак формується з трьох груп, при цьому в першій групі міститься дві ознаки, в другій - 4, а в третій - 10, то довжини всіх груп ознак, що є рядками формується матриці, будуть рівні максимальної кількості ознак однієї з використовуваних груп, тобто, 10. Таким чином, використання матричного формату призводить до появи надлишкових даних, наслідком є збільшення кількості обчислювальних операцій. Даний недолік використання матриць ознак для опису

досліджуваної множини об'єктів набуває великого значення при використанні імунного підходу до організації процесу саморегуляції імунної мережі та угруповання антитіл. Причиною цього є використання оператора мутації, у процесі якого зміни піддаються всі ознаки об’єкта, що мутує, незалежно від того, чи використовуються вони в групі чи ні. Наслідком цього є підвищення кількості обчислювальних операцій щодо афінності між об’єктами, представленими в такий спосіб. Тому використання матричного уявлення імунних об’єктів може призвести до збільшення часу кластеризації.

Під відсутністю ваги у рядків матриці розуміється те, що всі рядки матриці мають однакову вагу при визначенні відстані між об’єктами. Відповідно, якщо матриця містить два стовпці ознак, у першому з яких розташовані координати об’єкта, а в другому – його колір, то визначення відстані між двома об’єктами буде проходити наступним чином: на першому етапі відбудеться визначення аффінностей між відповідними рядками ознак кожного об’єкта. Після цього отримані аффінності будуть нормовані. При цьому значення аффінності між двома імунними об'єктами визначається таким чином:

$$aff\,(ab_i, ab_j) = 1/k \sum_{g=1}^{k} aff\,(ab_{ig}, ab_{jg}), \qquad (3.1)$$

де k – кількість стовбців ознак;

$aff\,(ab_{ig}, ag_{jg})$ – аффінність між відповідними стовбцями антитіл.

Використання завислих матриць ознак дозволяє проводити встановлення ваг стовпчикам (векторам) ознак, які згодом будуть використовуватися при визначенні відстані між об’єктами. Таким чином, використання зважених матриць ознак надає більше засобів керування даними при моделюванні алгоритмів.

При використанні зваженого масиву векторів ознак для опису імунних

об’єктів з однорідних по діапазону допустимих значень ознак формуються вектори, що характеризуються деяким ваговим коефіцієнтом. При визначенні аффінності між двома об’єктами першому етапі відбуватиметься визначення аффінностей між відповідними векторами, але в другому відбудеться нормування даних аффінностей з урахуванням вагових коефіцієнтів. Таким чином, при використанні даного способу представлення імунних об’єктів аффінність між ними визначається наступним чином:

$$aff\,(ab_i, ag_j) = 1/k \sum_{g=1}^{k} w_g aff\,(ab_{ig}, ag_{jg}), \qquad (3.2)$$

де k – кількість стовбців ознак (векторів);

w_g – вага стовбця;

$aff\,(ab_{ig}, ab_{jg})$ – аффінність між відповідними стовбцями об’єктів.

Використання даного підходу представлення імунних об'єктів усуває недоліки векторного та матричного способів представлення об’єктів, що групуються. При цьому відбувається підвищення швидкодії імунного алгоритму внаслідок скорочення кількості надлишкових обчислень, що виникають на етапах мутації та відбору клонів, характерних для матричного способу уявлення. Тому в Dendric-aiNet для опису даних використовується саме цей формат представлення об’єктів, що кластеризуються.

3.2 Розробка структури програмного забезпечення кластеризації на основі деревовидної імунної мережі

Програмна реалізація середовища моделювання імунного методу кластеризації даних Dendric-aiNet, а також еталонних методів кластеризації, що функціонують на основі класичних принципів організації обчислень та на основі імунного підходу, була створена при використанні платформи .NET,

мови програмування C# та технології розробки та прототипування інтерфейсу користувача Windows Форми. Крім того, програма реалізована на основі однієї з найбільш поширених тришарових архітектур створення додатків – MVC.

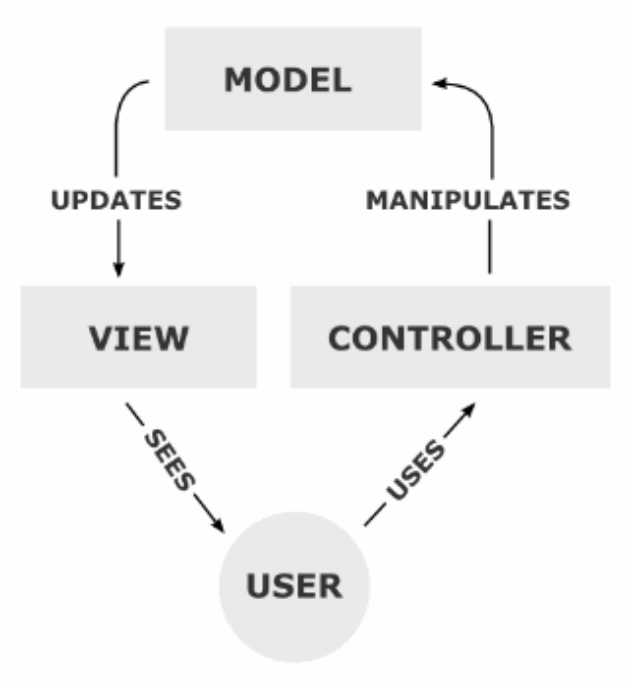

Рисунок 3.1 – Архітектура MVC

Архітектурний паттерн MVC є універсальним і не пов’язаний з якоюсь конкретною мовою програмування. Цей паттерн є першим паттерном, який отримав широке використання при розробці як web-орієнтованих додатків, так і додатків для стаціонарних комп’ютерів та мобільних пристроїв. Абревіатура MVC розшифровується як Model-View-Controller або модель-вигляд-контролер. Відповідно до цього паттерном вихідний код проекту поділяється на три основні частини, при цьому кожна частина має певний набір відповідальності. Шар Model містить набори даних, сховища об’єктів та модулі, що реалізують логіку проекту. Шар View містить набір вікон, що використовуються для організації взаємодії з користувачем програми. Шар Controller використовується для організації зв’язку між шаром Model та шаром View. Таким чином, в даному паттерні шар Controller є посередником

між двома основними шарами програми, що містить велику кількість типів даних, що реалізують бізнес-логіку проекту.

Середовище моделювання *K*-зв’язного деревовидного імунного методу кластеризації даних, що розробляється, є складним програмним засобом, та умовно поділяється на кілька основних модулів, що мають різне функціональне призначення:

– модуль інтерфейсу користувача, реалізований в шарі View;
– модуль представлення даних, реалізований у шарі Model;
– модуль алгоритмів та операторів, реалізований у шарі Model;
– модуль файлових операцій, реалізований у шарі Model;
– контролери програми, реалізовані в шарі Controller.

У модулі інтерфейсу користувача містяться типи даних, що дозволяють організовувати налаштування та запуск методу кластеризації Dendric-aiNet користувачем, проводити візуалізацію результатів, а також визначати час класифікації та точність групування даних.

У модулі даних знаходяться класи, що використовуються для роботи з імунними об'єктами, представленими населенням вихідних антитіл, навчальних антигенів, клонів та класів. Однією з основних класів, які у цьому модулі, є клас формату даних. Слід зазначити, що при цьому відповідно до основних положень теорії штучних імунних систем класи антигенів, антитіл і клонів влаштовані однаково і успадковуються від одного базового класу клітини.

Модуль представлення даних містить типи, що використовуються для представлення антитіл, клонів, кластерів та інших імунних об’єктів, що використовуються різними алгоритмами кластеризації даних.

Модуль алгоритмів та операторів містить реалізацію методу Dendric-aiNet, класи з методами обчислення манхеттенської відстані, обчислення аффінності, обчислення коефіцієнта мутації, а також реалізацію всіх імунних операторів, реалізованих у цьому методі. Крім того, у даному модулі містяться типи даних, що реалізують алгоритми кластеризації, що

функціонують на основі класичних принципів організації обчислень, а також на основі імунного підходу. При цьому кожен імунний оператор реалізований в окремому статичному класі, що спрощує можливості його використання в імунному алгоритмі.

Модуль файлових операцій містить класи, що дозволяють працювати зі стандартними типами фалів, представленими у форматах CSV та BMP. У даних форматах зберігається вся інформація про популяції імунних об’єктів, формати їх уявлення, а також результати класифікації.

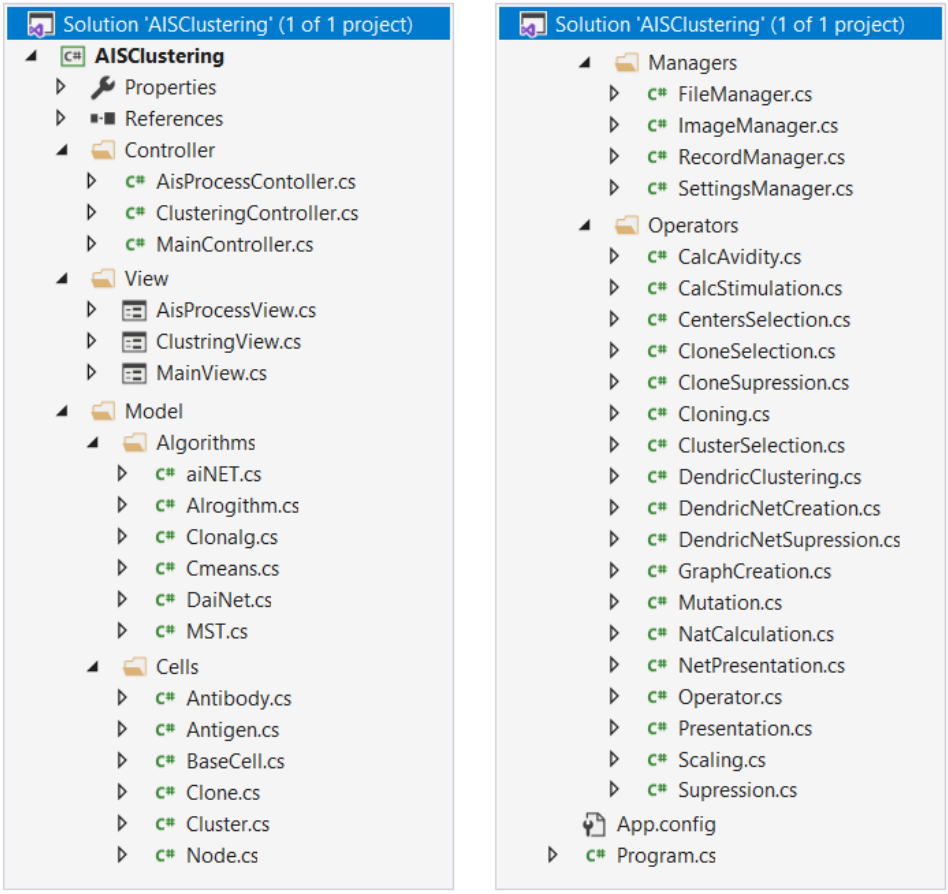

Рисунок 3.2 – Файлова структура застосунку

Модуль контролерів програми містить набір типів, що організують взаємодії між вікнами інтерфейсів користувача і різними типами і модулями шару Model. Слід зазначити, шар Controller має розподілену реалізацію, тобто. у вихідному коді програми немає одного загального класу, що реалізує

весь набір необхідного функціоналу програми. Натомість програма має набір контролерів, кожен з яких обслуговує окреме вікно програми та його функціональні зв’язки з типами, що знаходяться у шарі Model. Цей підхід повністю відповідає принципам сучасного об’єктно-орієнтованого програмування SOLID та забезпечує гнучкість архітектури та масштабованість програми.

На рисунку 3.2. наводиться файлова структура програми, розділена на три основні шари відповідно до архітектури MVC. Відповідно, папка Controllers містить три контролери, які використовуються для організації взаємодії між основними вікнами і типами, представленими в просторі Model, що реалізують основну логіку програми.

Папка View містить три типи основних вікон програми: AisProcessView – використовується для налаштування та візуалізації процесу імунного навчання та саморегуляції мережі для різних алгоритмів, що функціонують на основі імунного підходу. Тип ClusteringView представляє вікно, яке використовується для налаштування та запуску процесу кластеризації на основі різних методів та моделей групування даних. Тип MainView представляє головне вікно програми, яке дозволяє налаштовувати процес кластеризації, вибирати алгоритми групування даних, генерувати набори даних та керувати налаштуваннями процесу обробки об’єктів, що досліджуються.

Папка Algorithms містить набір алгоритмів класифікації та кластеризації даних, що функціонують як на основі класичних принципів організації обчислень, так і на основі імунного підходу. Головним типом даних, що знаходиться в цій частині проекту, є тип Algorithm, який є абстрактним класом і використовується як базовий клас для всіх інших типів даних, визначених у даному просторі імен шару Model.

Папка Cells містить набір типів даних, використовуваних представлення наборів об'єктів, які згодом у процесі функціонування алгоритмів групування даних розподіляються між класами чи кластерами,

залежно від поставленої завдання дослідження. У даному просторі імен тип BaseCell є абстрактним класом і є базовим типом для решти типів даних, оголошених в Cells. Крім того, тип Node є базовим типом даних для типів об'єктів, що використовуються в імунних методах, що функціонують на основі моделі штучної імунної мережі. Крім того, оскільки в імунних методах угруповання даних передбачається етап клонування, тип Clone є спадкоємцем типу Antibody, який, у свою чергу, є спадкоємцем типу Node. Цей ланцюжок успадкування дозволяє формувати мережу антитіл імунним методам aiNET та Dendric-aiNet.

Папка Managers містить типи даних, що дозволяють організовувати маніпуляції з наборами об'єктів та перетворення цих об'єктів у зображення (тип ImageManager), не графічні набори об'єктів даних, що групуються (тип FileManager), інформацію про результати класифікації та кластеризації (тип RecordManager), інформацію про налаштування алгоритмів угруповання даних (тип SettingsMaager). Перелічені типи даних, що знаходяться у просторі імен Managers, є статичними класами. Це зроблено для того, щоб спростити навігацію коду проекту та підвищити його масштабованість.

Папка Operators містить набір типів даних, реалізують той чи інший етап роботи алгоритму класифікації і кластеризації з урахуванням класичних принципів організації обчислень, і навіть моделей штучних імунних систем. Серед типів даних, представлених у цьому просторі імен можна назвати тип Operator, який є абстрактним базовим типом всім інших операторів, у тому числі формується алгоритм класифікації чи кластеризації даних.

На рисунку 3.3. наводиться загальний інтерфейс системи моделювання, що представлений вікном MainView. Панель Objects представляє користувачеві набір можливостей роботи з набором зважених векторів ознак, що характеризують досліджуваний набір об'єктів. Також дана панель дозволяє налаштовувати та створювати набори об'єктів для класифікації або кластеризації, або популяції антитіл та антигенів для запуску та аналізу роботи імунних методів, що використовуються для класифікації та

кластеризації даних.

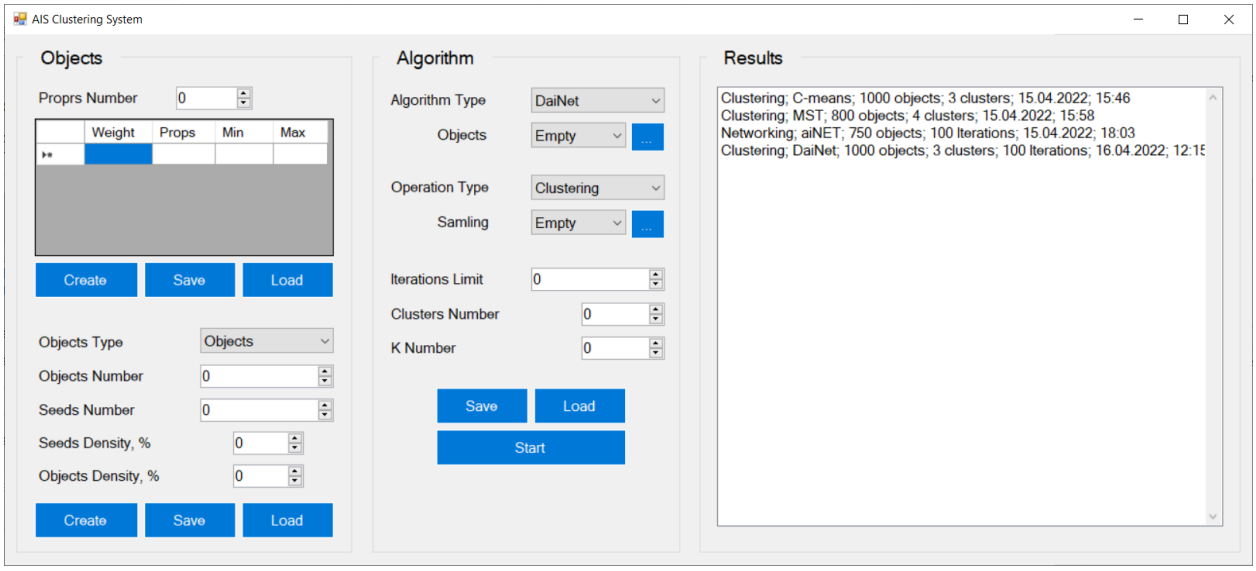

Рисунок 3.3 – Головне вікно застосунку

Панель Algorithm дозволяє користувачеві програми вибирати та настроювати роботу алгоритму класифікації або імунної моделі для візуалізації процесу роботи обраної імунної моделі. У додатку доступні такі алгоритми угруповання даних: kNN, C-Means, MST та імунні моделі Clonalg, aiNET, Dendric-aiNet. Для запуску обраного алгоритму необхідно встановити згенерувати, або завантажити набір об’єктів, що групуються. Можливості завантаження набору об'єктів для кластеризації також доступні на даній панелі, при цьому як об'єкти можуть виступати прості не імунні об'єкти – Objects, популяція антигенів – Antigens та популяція антитіл – Antibodies. Крім того, панель Algorithm дозволяє вибирати тип операції, яка виконується алгоритмом, вибраним користувачем програми. Серед таких операцій можна назвати класифікацію – Classification, кластеризацію – Clustering, моделювання роботи штучної імунної мережі – Networking. Кластеризація не має на увазі використання навчальної вибірки, але при виборі операції класифікації навчальну вибірку можна завантажити як популяцію антигенів. Крім того, кількість формованих кластерів (якщо обрана кластеризація), критерій зупинки і параметр K, використовуваний різними імунними та не

імунними методами групування даних, також можна встановити в даній панелі. Панель представляє користувачеві можливості збереження та завантаження налаштувань алгоритму групування даних та його запуску у окремому вікні.

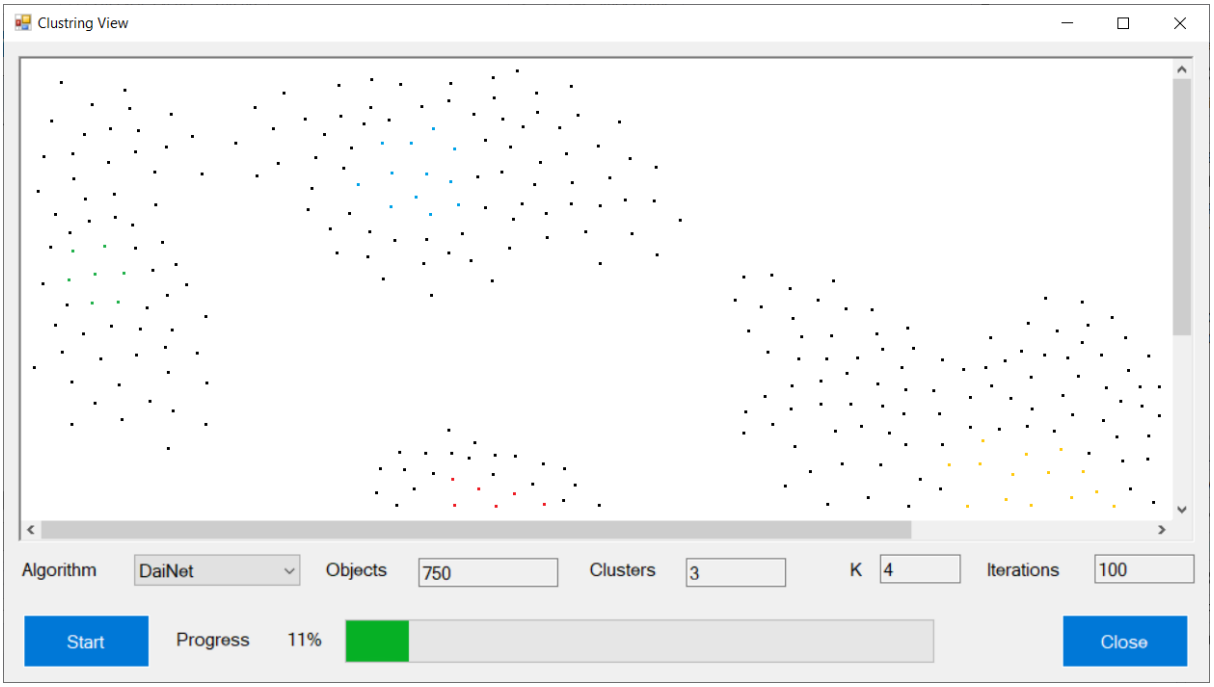

Рисунок 3.4 – Вікно запуску алгоритму кластерихації

Панель Results у вікні MainView зберігає інформацію про запуски алгоритмів угруповання даних та дозволяє користувачеві вибирати та відкривати збережену інформацію про процес роботи того чи іншого алгоритму з відповідними налаштуваннями у блокноті.

При запуску алгоритму угруповання даних або процесу імунної саморегуляції користувачеві відображається вікно ClusteringView, представлене на рисунку 3.4. Дане вікно відображає користувачеві основні налаштування алгоритму угруповання даних та тип обраної операції, а також команди запуску алгоритму та примусового завершення роботи алгоритму та перемикання на головне вікно. Крім того, вікно ClusteringView використовується для візуалізації процесу алгоритму, т.к. на кожній ітерації формується зображення, яке відображається на головній панелі. При цьому

об’єкти, що досліджуються, різняться за кольором. У разі розв’язання задачі класифікації або кластеризації всі об’єкти спочатку маркуються чорним кольором, а також в міру роботи алгоритму групування даних, приймають колір того чи іншого класу або кластера.

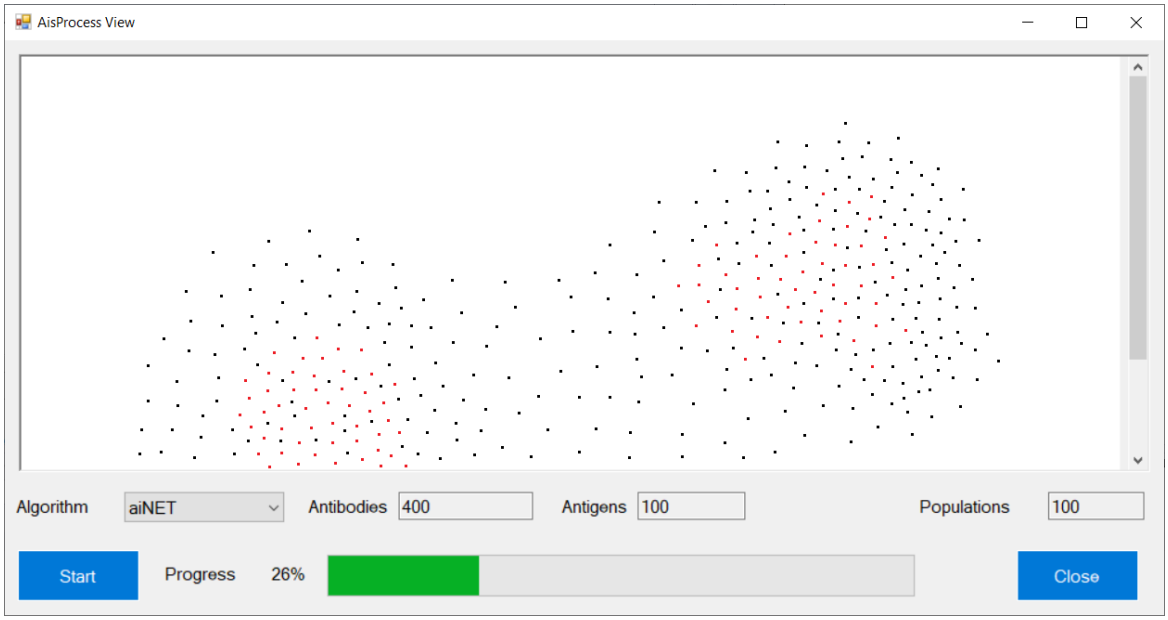

Рисунок 3.5 – Вікно запуску процесу моделювання штучної імунної системи

На рисунку 3.5. наводиться інтерфейс вікна AisProcessView, яке використовується для відображення процесу моделювання роботи штучної імунної системи, реалізованої у конкретному імунному алгоритмі. Перехід на дане вікно здійснюється з головного вікна програми у разі вибору користувачем операції типу Networking та відповідного імунного методу. Дане вікно відображає користувачеві інформацію про обраний для моделювання імунний метод, популяції антитіл і антигенів і критерії зупинки, вираженого кількістю популяцій антитіл, що формуються імунною системою за допомогою клонування вихідного набору антитіл.

3.3. Порівняння результатів кластеризації метод Dendric-aiNet та іншими методами гуртування даних

Для аналізу запропонованого методу кластеризації з іншими поширеними методами угруповання даних та імунними моделями було сформовано 3 набори об'єктів, які є обмеженою кількістю ознак. Дані набори різняться між собою кількістю об'єктів та кількістю кластерів, які мають бути отримані в результаті кластеризації. Характеристики наборів даних наведено у таблиці 3.1.

Таблиця 3.1 – Набори об'єктів для кластеризації

Ідентифікатор	Кількість	
	Об'єкти	Кластери
Набір 1	100	3
Набір 2	500	5
Набір 3	2500	10

Приведені набори об’єктів використовуються двома способами: як вихідні набори даних для алгоритмів кластеризації та як контрольні вибірки для перевірки якості класифікації. Це відбувається тому, що вказані набори даних вже класифіковані, але можуть використовуватися без вказівки початкової належності того чи іншого об’єкта деякого класу або кластеру.

Порівняння запропонованого методу з іншими імунними та не імунними методами угруповання даних наводиться у таблиці 3.2. При аналізі продуктивності запропонованого методу Dendric-aiNet використовувалися еталонні методи кластеризації: MST та C-means, а також універсальні імунні методи Clonalg та aiNET. При порівнянні продуктивності алгоритмів використовувалися дві основні метрики:

– час T, який витрачається алгоритмом кластеризації виконання групування вихідного набору об’єктів;

– точність кластеризації А, яка визначаться в результаті порівняння кластерів, сформованих під час роботи алгоритму кластеризації та належності набору об’єктів вхідним класам.

Таблиця 3.2 – Результати кластризації об'єктів

Алгоритм		Набір 1	Набір 2	Набір 3
MST	T	38%	36%	39%
	A	88%	85%	82%
C-means	T	72%	74%	72%
	A	100 %	100 %	100 %
Clonalg	T	100 %	100 %	100 %
	A	80%	83%	81%
aiNet	T	98%	95%	93%
	A	52%	50%	50%
Dendric-aiNet	T	48%	46%	47%
	A	95%	93%	96%

За підсумками кластеризації описаних наборів даних алгоритм C-means характеризується максимальною точністю угруповання об'єктів, а алгоритм Clonalg - найгіршою швидкодією. Тому дані алгортми були обрані як абсолютні максимальні значення (100%). З іншого боку, Метод MST характеризується найкращим швидкодією, тобто. на кластеризації при використанні даного методу йде найменше часу, приблизно в 4 рази менше, ніж на кластеризацію за допомогою методу Clonalg. При цьому метод MST значно поступається за точністю угруповання досліджуваної множини об'єктів іншим методам, наведеним у таблиці 3.2.

Слід зазначити, що запропонований метод Dendric-aiNet характеризується високою точністю угруповання даних. За цією характеристикою Dendric-aiNet поступається лише методом C-means на кілька відсотків. Однак, запропонований метод Dendric-aiNet перевершує метод C-means по швидкодії практично на 10% і уступає за цим показником

лише методом MST. При порівнянні методу Dendric-aiNet коїться з іншими імунними методами, які зазвичай застосовуються на вирішення завдань класифікації і кластеризації даних, слід зазначити, що Dendric-aiNet перевершує інші методи як із швидкодії, і за точності угруповання об'єктів.

Це робить даний метод найбільш адаптованим методом для вирішення задачі кластеризації даних на основі імунного підходу до організації обчислень.

ВИСНОВКИ

У кваліфікаційній розглянуто вирішення актуального завдання кластеризації даних на основі імунного підходу. При виконанні поставленого завдання було проведено аналіз найбільш популярних методів класифікації та кластеризації даних, а також принцип роботи імунних методів угруповання даних.

Проведений аналіз імунних моделей дозволив визначити основні відмінності між ними, а також досліджувати їх властивості та особливості. В результаті вивчення особливостей даних моделей було виділено їх основні переваги та недоліки. Крім того, для кожної виділеної імунної моделі було визначено універсальні алгоритми, що використовуються для вирішення різних практичних завдань. У цьому було визначено універсальні імунні оператори, які можна використовувати розробки імунного методу автоматичної класифікації. В ході дослідження особливостей імунних операторів було вивчено основні підходи, що використовуються для організації їх функціонування.

В результаті проведення низки експериментів з моделювання імунних методів, реалізованих на основі різних імунних моделей, був виділений метод Dendric-aiNet, який функціонує на основі деревоподібної k-зв'язної штучної імунної мережі, як найбільш підходящий для вирішення задачі автоматичної класифікації. Використання даного методу для вирішення поставленої задачі обумовлюється необхідністю мережевої взаємодії виділення кластерів антитіл.

При виконанні даної кваліфікаційної роботи було розроблено та реалізовано імунний метод автоматичної кластеризації Dendric-aiNet, що характеризується високою швидкістю обчислення та гарною точністю кластеризації досліджуваного набору об'єктів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Вятченин Д.А. Нечеткие методы автоматической классификации / Д.А. Вятченин. – Минск: Технопринт, 2004 – 218 с.

2. Fielding A.H. Cluster and classification techniques for the biosciences / A.H. Fielding. – Cambridge University Press, 2007. – 260 p.

3. Gordon A.G. Classification Second Edition / A.G. Gordon. – CRC Press, 1999. – 248 p.

4. Рубан А.И. Методы анализа данных / А.И. Рубан. – Красноярск: ИПЦ КГТУ, 2004. – 319 с.

5. Mirkin B.G. Clustering for Data Mining. A Data recovery Approach / B.G. Mirkin. – Taylor & Francis Group, 2005. – 278 p.

6. Han J. Data Mining Concepts and Techniques Second Edition / J. Han, M. Kamber. – Elsevier, 2006. – 772 p.

7. de Oliveira J.V. Advances in fuzzy clustering and its applications / J.V. de Oliveira, W. Pedrycz. – John Willey & Sons, 2007. – 460 p.

8. Gan G. Data clustering theory, algorithms, and applications / G. Gan, C. Ma, J. Wu. – Society Industrial and Applied Mathematics, SIAM, 2007. – 490 p.

9. Айвазян С.А. Прикладная статистика, классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М: Финансы и статистика, 1989. – 609 с.

10. Duda R.O. Pattern classification Second Edition / R.O. Duda. – Willey-Interscience, 2000. – 738 p.

11. Mittal A. Addressing the Problems of Bayesian Network classification of Video Using High Dimensional Features / A. Mittal, L.F. Cheong // IEEE Transactions on Knowledge and Data Engineering-04, Issue 2, 2004. – P. 230-244.

12. Han J. Data Mining Concepts and Techniques Second Edition / J. Han, M. Kamber. – Elsevier, 2006. – 772 p.

13. Garain U. Recognition of handwritten indic script using clonal selection algorithm / U. Garain, M.P. Chakraborty, D.Dasgupta // Springer, Lecture Notes in Computer Science, № 4163, 2006. – P. 256-266.

14. Secker A. AISEC: an artificial immune system for e-mail classification / A. Seckler, A.A. Freitas, J. Timmis // IEEE, Proc. The Congress on Evolutionary Computation, CEC-03, 2003. – P. 131-139.

15. Igawa K. Discrimination-based artificial immune system: modeling the learning mechanism of self and non-self discrimination for classification / K. Igawa, H. Ohashi // Journal of Computer Science, № 4, 2007. – P. 204-211.

16. Leung K. Generating compact classifier systems using a simple artificial immune system / K. Leung, F. Cheong, C. Cheong // IEEE, Transactions on Systems, Man, and Cybernetics, № 5, 2007. – P. 1344-1356.

17. Литвиненко В.І. Вирішення задачі класифікації з використанням механізмів ідіотипічної мережі / Литвиненко В.І. // Наукові праці: науково-методичний журнал, серія «Комп'ютерні технології» – МДТУ ім. П.Могили, Вип. 44, 2006. – С. 136-146.

18. «Охорана труда. Сборник задач» : Б. В. Дзюндзюк, , Т. Е. Стыценко – Харьков: ХНУРЭ, 2006. – 241 с

18. Кукарцев, В. В. Порівняння систем контролю версій: Git, Mercurial, CVS і SVN [Текст] / В. В. Кукарцев, С. А. Бадарчи // Синергия наук. – 2018. – №19. − С. 538-548.

19. Вирт, Н. Алгоритми та структури даних. Нова версія для Оберона [Текст] / Н. Вирт. – М.: ДМК Пресс, 2010. – 410 с.

20. Puntambekar, A. A. Software Engineering [Текст] / A. A. Puntambekar. Technical Publications, 2009. – 332 p.

21. Таненбаум Э., Остин Т. Т18 Архитектура компьютера. 6-е изд. – СПб.: Питер, 2013. – 816 с.: ил. ISBN 978-5-496-00337-7.

22. Буч Г., Рамбо Д., Якобсон І. Мова UML. Інструкція користувача. 2-е вид. [Текст]: пер. з англ. Мухін Н. –. М.: ДМК Прес. – 496 с.

23. Олифер В., Олифер Н. Компьютерные сети. Принципы, технологии,

протоколы: Учебник для вузов. 5-е изд. – СПб.: Питер, 2016. – 992 с.: ил. – (Серия «Учебник для вузов»).

24. Фаулер М. Рефакторинг: улучшение существующего кода. – Пер. с англ. – СПб: Символ-Плюс, 2003. – 432 с., ил. ISBN 5-93286-045-6.

25. Таненбаум Э., Остин Т. Т18 Архитектура компьютера. 6-е изд. – СПб.: Питер, 2013. – 816 с.: ил. ISBN 978-5-496-00337-7.

26. Мартин Р. Чистая архитектура. Искусство разработки программного обеспечения. – СПб.: Питер, 2018. – 352 с.: ил. – (Серия «Библиотека программиста»). ISBN 978-5-4461-0772-8.