МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії

ДЕРИШ Богдан Богданович

Алгоритми класифікації біомедичних зображень на основі методу k-means / Biomedical Images Classification Algorithms on the k-Means Basis

спеціальність: 123 - Комп'ютерна інженерія
магістерська програма - Комп'ютерна інженерія

Магістерська робота

Виконав студент групи КІм-21
Б. Б. Дериш
_____

Науковий керівник:
д.т.н., професор, О. М. Березький
_____

Магістерську роботу допущено
до захисту:

"31" 05 2018 р.

Завідувач кафедри
_____ **О. М. Березький**

ТЕРНОПІЛЬ - 2018

Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії
Освітній ступінь «магістр»
спеціальність: 123 - Комп'ютерна інженерія
магістерська програма - Комп'ютерна інженерія

ЗАТВЕРДЖУЮ
завідувач кафедри
д.т.н., проф. Березький О.М.
"10" 10 20 17 р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТА
Дериша Богдана Богдановича

1. Тема дипломної роботи " Алгоритми класифікації біомедичних зображень на основі методу k-means"
керівник роботи д.т.н., проф. Березький О.М.
затверджена наказом по університету від 11 жовтня 2016 р. №669.

2. Строк подання студентом роботи "15"січня 2018 року

3. Вихідні дані до магістерської роботи
Об'єкт дослідження: Біомедичні зображення.
Предмет дослідження. K-means алгоритми класифікації гістологічних та цитологічних зображень.

4. Зміст розрахунково-пояснювальнох записки (перелік питань, які потрібно вирішити):
- проаналізувати відомі алгоритми класифікації зображень;
- проаналізувати біомедичні зображення(цитологічні та гістологічні);
- розробити алгоритм вибору k-means подібних алгоритмів;
- оптимізувати алгоритми класифікації зображень на основі k-means подібних алгоритмів;
- розробити і програмно реалізувати модуль класифікації зображень;
- протестувати програмний модуль класифікації.

5. Перелік графічного матеріалу:
– тема, мета, завдання, методи досліджень, наукова новизна, практичне значення;

– актуальність;

– класифікація програмного забезпечення для діагностування біомедичних зображення;

– об'єкт дослідження;

– алгоритм сегментації методом k-середніх;

6. Консультанти розділів магістерської роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв
Антиплагіат	Мельник Г.М., доцент		
Нормо-контроль	Гураль І. В., викладач		

7.Дата видачі завдання «17» 10 2016р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1	Інтелектуальні системи діагностування	3.11.2016 – 1.01.2017	
2	Розроблення алгоритмів аналізу зображень	2.01.2017 – 31.05.2017	
3	Програмна реалізація інтелектуальної системи	1.06.2017 – 25.01.2018	
4	Нормоконтроль, попердній захист	16.01.2018 – 2.02.2018	
5	Захист	5.02.2018	

Студент _____ Дериш Б.Б.
(підпис) (прізвище та ініціали)

Керівник магістерської роботи _____ д.т.н., проф. Березький О.М.
(підпис) (прізвище та ініціали)

РЕЗЮМЕ

Дипломна робота на тему “ Алгоритми класифікації біомедичних зображень на основі методу k-means ” на здобуття освітньо-кваліфікаційного рівня “Магістр” зі спеціальності “Комп’ютерна інженерія” написана обсягом 116 сторінка і містить 39 ілюстрацій, 8 таблиць, 3 додатки та 49 джерел за переліком посилань.

Метою роботи є розроблення програмного модуля для класифікації біомедичних зображень, використовуючи алгоритми роботи із даними та метод K-means.

Методи досліджень. Для досягнення поставлених цілей в дипломній роботі використано: дерева рішень, алгоритми класифікації, теорію математичного моделювання, методи структурного програмування.

Результати дослідження: структура та модель програмного модуля класифікації, його реалізація та тестування.

Результати роботи можуть бути використані у різного роду медичних установах та діагностичних центрах, а також стануть у пригоді при побудові програмних проектів.

Орієнтовні напрямки розвитку досліджень: розроблення комп’ютерних додатків; он-лайн веб орієнтованих систем; підвищення ефективності діагностики онкологічних захворювань.

КЛЮЧОВІ СЛОВА: МОДЕЛЬ, АНАЛІЗ, КЛАСИФІКАЦІЯ, БІОМЕДИЧНІ ЗОБРАЖЕННЯ, ДІАГНОСТИКА, ОНКОЛОГІЯ, K-MEANS.

RESUME

The thesis on the topic "Algorithms for the classification of biomedical images based on the k-means method" for obtaining an educational qualification level "Master" in the specialty "Computer Engineering" is written in volume 116 pages and contains 39 illustrations, 8 tables, 3 annexes and 49 sources. by the list of links.

The aim of the work is to develop a software module for the classification of biomedical images using data-processing algorithms and the K-means method.

Research methods. To achieve the goals in the thesis used: decision trees, classification algorithms, mathematical modeling theory, methods of structural programming.

Research results: the structure and model of the program module of classification, its realization and testing.

The results of the work can be used in various medical institutions and diagnostic centers, and also will be useful in constructing program projects.

Approximate directions of research development: development of computer applications; on-line web-oriented systems; increasing the effectiveness of diagnosis of oncological diseases.

KEY WORDS: MODEL, ANALYSIS, CLASSIFICATION, BIOMEDICAL IMAGES, DIAGNOSIS, ONCOLOGY, K-MEANS.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

IRMA PX

ABBYY iOS

OCR Pro

DPI Mac

RTF HTML

ПЗ YAGF

BSD GPRSG

RWTH PHP

СНД PCA

КЛТ EVD

ХТХ SVD

ЕОФ CCA

АСІІТ CAD

УЗД MPT

КТ GAC

WNCut-CGAC

ДАБ АОИ

ISODATA DCT

ДЗЗ ПЕОМ

IA

Ks

PX

MC

ВСТУП

Актуальність теми. Рак посідає перше місце в структурі онкозахворюваності. Кожна десята людина в Україні страждає даним недугом, а згідно прогнозу канцер реєстру незабаром буде хворіти кожна дев'ята. Так, щогодини помирає одна людина від раку, а кожні 30 хвилин фіксується новий випадок. Тому, проблеми ранньої діагностики, лікування та профілактики набувають особливої актуальності. Рак молочної залози зустрічається, у порівнянні з пухлинами інших локалізацій, досить часто. У світі щорічно реєструється близько одного мільйона нових випадків захворювання на рак молочної залози. В країнах СНД ця патологія посідає перше місце у структурі смертності від онкологічних захворювань у жінок працездатного віку. Особливу актуальність проблема раку має в промислово розвинених регіонах України. На жаль, впровадження сучасних методів діагностики та синтез нових препаратів суттєво не вплинули на кількість пацієнток з уперше виявленими задавненими формами на цю патологію. На сьогодні смертність від злоякісних пухлин в світі складає 17,9 на 100 тисяч населення, а в Україні - 27,0 на 100 тисяч населення. Перед онкологом постають задачі не тільки первинної й уточнюючої діагностики та лікування раку, але й оцінки ефективності різних методів лікування цієї патології, своєчасного виявлення рецидивів після проведеного лікування. Сьогодні опрацювання біомедичних зображень є важливим напрямком застосування сучасної медичної техніки. Задачами опрацювання зображень є опис, аналіз та оброблення зображень. Проблеми аналізу біомедичних зображень включаючи класичну задачу розпізнавання фігур заданої форми, важлива також експертна оцінка, яка зараз є дорогою. Виникають проблеми, які зумовлені новими завдання опису зображення та пошуком закономірностей або наборів закономірностей, що одночасно зустрічаються в багатьох наборах. Оскільки наборів може бути велика кількість необхідно здійснювати цей пошук автоматично. Тому актуальною задачею є розробка ефективних методів розпізнавання та ранньої діагностики.

Мета і завдання дослідження. Метою дослідження є розробити програмний модуль класифікації базуючись на методі k-means.

Об’єкт дослідження. Біомедичні зображення.

Предмет дослідження. Алгоритми класифікації.

Методи досліджень. Базуються на використанні методів інтелектуального аналізу даних; алгоритмів класифікації (для класифікації мікрооб’єктів на зображеннях та побудові моделей); методу головних компонент (для роботи із обширною кількістю даних)

Наукова новизна одержаних результатів. 1.розроблено узагальнений алгоритм для побудови класифікації. 2.вперше застосовано алгоритми відповідно до наявної тестової бази біомедичних зображень та отримано відповідні класи.

Практичне значення отриманих результатів. Розроблено узагальнений алгоритм класифікації для діагностики передракових та ракових станів.

Публікації та апробація др. Основні результати дослідження «алгоритми класифікації біомедичних зображень на основі методу k-means» опубліковано на конференції ACIIT 2017 .

У першому розділі буде проведено аналіз біомедичних зображень, описано основні характеристика гістологічних та цитологічних зображень. Буде проведено аналіз методів та алгоритмів інтелектуального аналіз даних, та класифікацію стадій інтелектуального аналізу.

У другому розділі проведено аналіз модифікацій методу k-means та метод головних компонент, для формування єдиного алгоритму та структури програмного модуля .

У третьому розділі буде описано програмний засіб, тобто робочу станцію-ноутбук та програмне середовище для проведення експериментальних досліджень використовуючи наявну базу біомедичних зображень. Також буде розроблено програмний модуль класифікації. В додатках буде представлена модель дерева рішень для класифікованих мікрооб’єктів, вихідний текст програмного коду, світлокопії виданої публікації.

1АНАЛІЗ МЕТОДІВ, АЛГОРИТМІВ ОПРАЦЮВАННЯ ГІСТОЛОГІЧНИХ ЗОБРАЖЕНЬ

1.1 Біомедичні зображення їх види та особливості обробки

Біомедична інженерія — галузь науки і техніки, яка поєднує інженерно-технічні та медико-біологічні знання, засоби і методи для створення, вдосконалення і дослідження природних і штучних біологічних об’єктів, техніки, матеріалів і виробів медичного призначення, технологій і технічних систем діагностики, лікування, реабілітації і профілактики захворювань людини, а також програмного забезпечення та інформаційних технологій для вирішення прикладних і фундаментальних проблем біології і медицини. В даній галузі і розміщена робота із обробки та розпізнавання біомедичних зображень.

Метою такого дослідження є розробка ефективних засобів опрацювання діагностичної інформації та створення на її базі системи асистента, що прискорить роботу відповідних медичних спеціалістів. Хоча біомедичні зображення це лише мізерна частина такої масивної галузі, як біомедична інженерія. Вона має свій поділ на: цитологію та гістологію [1].

Для опрацювання цитологічних зображень основними операціями є їх попередня обробка контурний аналіз та виділення ядер. Найпростішим методом виділення контурів та перепадів яскравості є застосування нелінійного фільтру Робертса. Відповідно до цього методу обчислюється сума квадратів різниць між діагонально суміжними пікселями. Це може бути виконано згорткою зображення з двома ядрами: $\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}$ та $\begin{bmatrix}0 & 1\\-1 & 0\end{bmatrix}$.

Нехай $X_{i,j}$ – піксель і -го рядочка j –го стовпчика зображення, яке отримане після застосування до деякого зображення ядра перехресного оператору Робертса, тоді обчислення проводиться за однією з двох формул 1.1 та 1.2.

$$\tilde{x}_{i,j} = \sqrt{\left(x_{i,j} - x_{i+1,j+1}\right)^2 + \left(x_{i,j+1} - x_{i+1,j}\right)^2} \quad (1.1)$$

$$\tilde{x}_{i,j} = \left|x_{i,j} - x_{i+1,j+1}\right| + \left|x_{i,j+1} - x_{i+1,j}\right| \quad (1.2)$$

де $X_{i,j}$ – піксель і -го рядка j -го стовпчика початкового зображення. Перетворення кожного пікселя зображення перехресним оператором Робертса може показати похідну зображення вздовж ненульової діагоналі, комбінація цих перетворень зображень може також розглядатись як градієнт від двох верхніх пікселей до двох нижніх [2].

Позитивна риса застосування оператору Робертса – це швидкість обчислень. Він має також негативні риси: оператор дуже чутливий до наявності шумів на зображенні, отримані лінії контурів є доволі тоненькими. При використанні фільтру Робертса до зображень ядер букального епітелію результат отримали незадовільний: ліній контурів забагато, оскільки як правило зображення самого ядра є неоднорідним, самі лінії майже не розрізняються серед загального фону.

Іншою стороною даної проблеми є методи катіонних зображень Classi, які були запропоновані для установки клінічного діагнозу і працюють на принципах функціонування магнітно-резонансної або оптичної томографії.

Набір даних IRMA містить 10000 анонімних рентгенівських зображень, які були доскіпливо вибраний з клінічної практики в Аахені, технологічному університет-лікарні (RWTH) в Німеччині. Ці зображення були отримані з використанням томографії різних: методів і форми, відносних напрямків пристрою і пацієнта. Вони являють собою різні анатомічні частини тіла і біологічні системи пацієнтів різного віку, статі і патологій. Всі зображення в рівнях сірого, зберігаючи оригінальний аспект співвідношення і були класифіковані відповідно до кодів Ірми [3].

Гістологія займається вивченням мікроскопічної структури клітин і тканин організму. Знання біологічних (мікроскопічних) структур і їх функцій на субклітинному, клітинному, тканинному і рівні органів займає центральне

місце у вивченні поширення захворювання і його прогнозуванні. Крім того, щоб вивчати і аналізувати гістологічне зображення під мікроскопом, ідентифікувати патології, морфологічні характеристики тканини, яка вказує на наявність захворювання, такого як рак.

Зразок біопсії обробляється і його секції розміщуються на стеклах, щоб спостерігати за ними під мікроскопом для аналізу. Патологоанатом вивчає слайди тканини під мікроскопом і спостерігає його на різних рівнях збільшення, таких як 10X, 20X, 40X, 100X і т.д., щоб подивитися клітини, залози, ядро, і виявити схожість цих структур в нормальної та ураженої тканини. Якщо захворювання виявляє градацію виконується процес, який має справу з поширенням заражених клітини всієї тканини. Тоді для кожного пацієнта прогноз і подальше лікування планується беручи до уваги ступінь захворювання.

Цей діагноз патологом носить суб'єктивний характер. Тому кількісна оцінка цих образів є дуже важливим для об'єктивної діагностики. Крім того, через розвиток цифрових сканерів для отримання зображень ми отримуємо дані цифрового зображення для комп'ютерного аналізу з використанням допоміжної цифрової обробки. Таким чином, комп'ютерна діагностика захворювань (CAD) грає дуже важливу роль і стала одним з основних предметів дослідження гістологічної візуалізації і діагностики, де різні методи обробки зображень можуть бути використані для аналізу цих зображень для діагностики захворювань і прогнозу. Тому, гістологія дає наукову основу для клінічних досліджень, освіту і практику.

Метою даної роботи є дослідження алгоритмів надійного і точного аналізу зображень за допомогою комп'ютера, інтерпретація патоморфологічних образів. Різні методи обробки зображень будуть застосовуватися для класифікації текстури зображень , залози і ядер клітин, сегментації. Зважаючи на ідентифікацію типу клітин або класифікацію отриманих кількісних вимірювань ознаки хвороби ,по гістологічних знімках, і автоматично визначити наявність хвороби. Крім того, це дослідження допоможе вирішити різного роду

тяжкості захворювання, якщо захворювання є в зразку. Автоматизоване гістопатологічного дослідження було проведено для виявлення різних видів раку і класифікації , в тому числі простати (Найка і ін., 2007, Doyle і ін., 2007), молочної залози (Doyle і ін і ін., 2008, Найк і ін., 2008, Жан-Ромен Далла і ін., 2008), нирково-клітинний рак (С. Вахід і ін і ін., 2007) педіатричній нейробластоми пухлини (Олкей S і ін., 2009) і легень (Кайзер і ін., 2002). Використовуючи різні сегментації, виділення ознак і методів класифікації дослідники проаналізували гістологічне зображення.

Робота організована, щоб обговорити необхідність і проаналізувати процедуру комп'ютеризованої сегментації зображень гістопатології і класифікацію. Ці процедури аналізу також застосовні до всіх умов зображення в аналізі медичних зображень, як УЗД, МРТ, КТ і т.д.

Основний потік аналізу гістології зображень, а також порядок підготовки гістопатологічної гірки для мікроскопічного аналізу, включаючи збільшення зображення . Далі в розділі, методи сегментації для гістологічного зображення проглядаються. Короткий огляд про функції видобуток і вибір для різних сегментованих результатів. Огляд методів класифікації для патоморфологічних зображень з майбутнього аналізу з пов'язаної області. Далі йдуть, порівняння і висновок про дослідження. В літературі можна знайти велику кількість додатків і функції аналізу мікроскопічного зображення, видобуток і відбір поряд з багатьма методами обробки зображень для попередньої обробки, сегментації і класифікації, тут лише деякі приклади.

Після отримання цифрового гістологічного зображення через зразок біопсії, ручного обстеження це призводить до мінливості в діагностиці. Щоб подолати цю проблему, за допомогою комп'ютера системи використовуються об'єктивний аналіз захворювань. Основні кроки, необхідні для впровадження комп'ютерної системи аналізу такі, як показано на малюнку нижче. Він складається із цифрових методів обробки зображень, таких як сегментація зображення, виділення ознак, класифікація і т.д. [5].

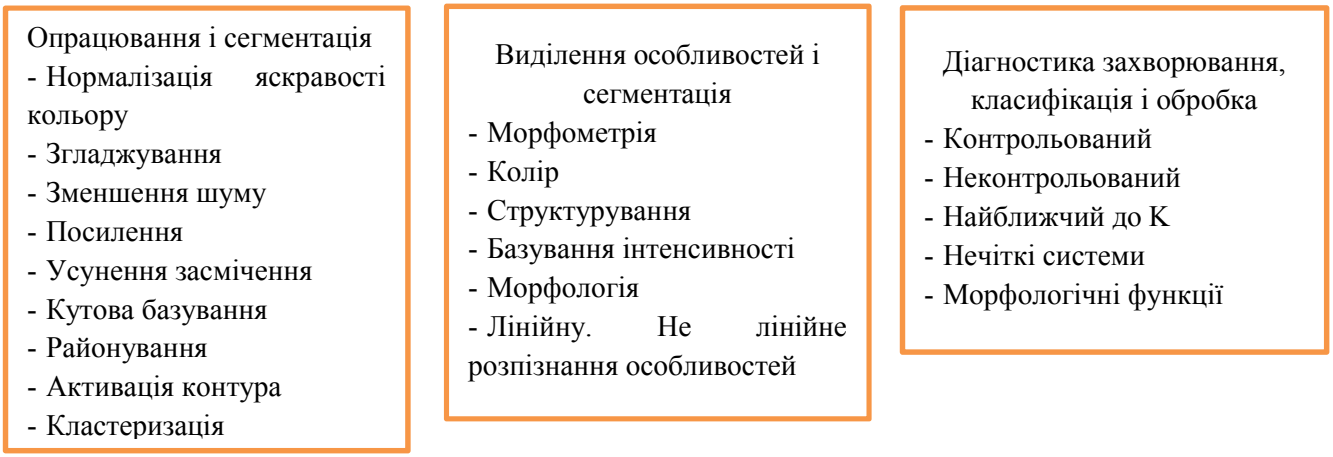

Рисунок 1.1 – Огляд гістологічного аналізу

Аналіз гістологічних зображень включає в себе обчислення, що виконуються при різних збільшеннях (× 2, × 4.5, × 10, × 20 і × 40) для багатовимірного статистичного аналізу, діагностики та класифікація. Це може бути зроблено при більш низькому збільшенні для аналізу рівня тканини. Як Сігдем Демірта ін. [6]. Обговорювалися два методів аналізу на рівні тканини і на клітинному рівні для діагностики раку. Вони проаналізували гістологічне зображення за допомогою попередньої обробки зображення, функцій витягання і методів класифікації, таких як thersholding, морфологічної обробки, і контрольованих методів класифікації. Недавнє просування "цифрової патології" потребує розвитку кількісного та автоматизованого комп'ютерного аналізу зображень, алгоритмів які допомагають патологоанатомам в інтерпретації великої кількості оцифрованих гістологічних зображень. Автоматизовані алгоритми діагностики були запропоновані для виявлення нейробластоми [7], а також комп'ютерного аналізу зображень раку нирок на основі Байесовского класифікатора, K-means алгоритм, запропонований в співавторстві із С. Вахідом [8], для поліпшення ітерацій, мінливості внутрішньо спостерігача.

Разом 98% точності класифікації з перехресної перевірки була досягнута за допомогою високого збільшення зображення, при 200X для розвитку системи. Метод аналізу нейробластоми гістологічних зображень, заснований на оцінка функції правдоподібності О. Sertel та ін. [9]. Ракові клітини були виявлені з чутливістю 81,1% з використанням двох компонентних кроків.

Аналогічним чином, в літературі можна знайти застосування таких різних методів аналізу зображень для класифікації даних. Ці методи також можуть бути застосовані до гістопатологічних зображень раку молочної залози для їх кількісного аналіз і оцінок [10].

Залежно від застосування або виду захворювання етапи обробки зображення можуть змінюватися, але загальні алгоритми обробки зображення однакові для більшості додатків. Джун Сюй та ін.[11] запропонували кольорове гістологічне зображення раку грудей на основі сегментації з використанням геодезичної активні контура (GAC) і зважене середнє зміщення нормалізованого розрізу. Вона включає в себе визначений зразок кольору із зображення і застосування нормалізованого середнього зсуву алгоритму скорочення вихідної сегментації для виявлення початкової межі об'єкта. Потім, на основі GAC виявлення краю колірного градієнта, остаточні результати сегментації були отримані.

Всього було використано 60 оцифрованих гістопатологічних слайдів і (WNCut-CGAC) результати були зіставлені з моделлю Чен Вессе. Відсоток точності і чутливість, були покращені WNCut-CGAC моделі з надійними ініціалізаторами, що має мінімальну взаємодію з користувачем. Різні алгоритми, розроблені для аналізу гістології, наведені в таблиці нижче.

Таблиця 1.1 – Різні алгоритми для аналізу гістологічних зображень

Рік публікації автор	Орган	Метод сегментації і класифікації зображення
1	2	3
Аджай Басаванхалйа 2011	Груди	ієрархічний нормал. зріз, кольоровий градієнт і активний контур
Джун Ксю 2011	Простата	Геодезичний активний контур

Продовження таблиці 1.1

1	2	3
С. Олкей 2008	Фолікулярна лімфома	Класифікація структур і нелін. колір, квантування, само організована карта
Омар Кхаді 2010	Менінгіома	Фрактальна класифікація структур
К. Демір 2010	Щитовидна залоза	Графова сегментація
Мутху Раме 2009	Ротова порожнина	Класифікатор SVM
Ху Конг 2011	Фолікулярна лімфома	Сегментація кольоровий структур клітини
М. Дундар 2011	Груди	Сегментація заснована на гаусовому розподілі
Акіф Тосум 2011	Груди	Сегментація через матрицю довжин графів

Гістологія не що інше, як мікроскопічне дослідження зразка біопсії, що закріплений на стеклах. Для того, щоб вивчити іншу архітектуру і компоненти тканини під мікроскопом, грубі секції виконані із воску, і пофарбовані з однією або декількома плямами. Фарбування використовується для патології, щоб відокремити клітинні компоненти для структурного, а також архітектурного аналізу тканини для встановлення діагнозу. Найчастіше гематоксилін -Eosin (H & E) який відокремлює клітинні ядра, цитоплазми і сполучні тканини. Гематоксилін забарвлює клітинні ядра синій, в той час як еозином плями цитоплазмі і рожевий сполучної тканини. Інші плями ДАБ, імунно-гістохимічними.

Зображення було отримано з використанням Motic B1 серії системи мікроскопа при 10-кратним збільшенням та калібруванням пікселів і 40X 10 з калібруванням 2,5 мкм / піксель. При цьому аналіз H & E забарвлених зображень будуть розглядатися для уникнення різних ефектів фарбування на кінцевий результат. Зображення, як показано на малюнках нижче, які показують низькі і високі збільшення нормальні та злоякісні зображення раку молочної залози.

Етапи підготовки слайд тканини на малюнку , і вона включає в себе:

- Виправлення: Зразки біологічної тканини / клітини є "фіксовані" з хімічної фіксації для збереження клітин / тканини.

- Обробка: тканин має важливе значення для видалення води з грубої тканини (зневоднення) і заміни його із середовищем. Це допомагає скоротити тонкість перетину зразка.

- Вбудовування зразок у віск: Результат вкладення загартованих воскових блоків містить оригінальні біологічні зразки разом з іншими речовинами в повному процесі підготовки .

- Секціонування: вбудований зразок тканини, необхідний для отримання досить тонких скибок зразка, як деталь мікроструктури клітин / тканин можна ясно спостерігати з використанням методів мікроскопії. Потім перенесять тонкий зріз зразка на чистому склі.

- Фарбування: Нарешті, змонтовані секції обробляють відповідним гістології плямою.

Фарбування біологічних тканин робиться для збільшення контрастності тканин, а також виділення деяких специфічних особливостей, що представляють інтерес - в залежності від типу тканини і використовується певна пляма.

Діагноз хвороби або її класифікації в гістологічних зображень залежить від ідентифікації гістологічних структур, таких як ядра ракових клітин, залоз, формування часточок, які відповідають раку молочної залози. Інший морфологічний вид цих структур, таких як розмір, форма, і інтенсивність

кольору, також є важливими факторами для наявності захворювання. Для того, щоб проаналізувати всі ці показники гістопатологічні зображення по-перше, повинні бути сегментованими. Для такої сегментації підходів, зображення повинні бути отримані на різних рівнях збільшення, такого як ядро сегментації 40X, ідентифікації клітин з 20X, і залози, сегментація тканини вимагає 10X / 4X збільшене зображення гістології. Ми можемо розглянути аналіз зображення для аналізу низької потужності (10X) або високого аналізу потужності (40X). Після того, гістологічні зображення справжні кольорові зображення, інтерпретуються на комп'ютері, доводиться спостерігати артефакти в зображенні через процедури фарбування.

Для усунення впливу такого шуму повинен бути препроцесором, де зашумлення і підвищення якості зображення, щоб отримати хороші сегментації і класифікації результатів. Це включає в себе нормалізацію кольору, якщо кольорове зображення обробляється для аналізу морфології клітин, щоб отримати більш детальну структуру клітин або інформацію з зображення і т.д. Крім того, важливо видалити ефект зміни в якості гістології, зображення забарвлюються з різними фарбуваннями пропорцій і умов сканування, отже, можуть бути зміна кольору в зображенні, який впливає на результати.

Після попередньої обробки, сегментації зображення є одним з найбільш важливих етапів автоматичної медичної діагностики, заснованої на аналізі мікроскопічних зображень, а також є важким завданням, щоб правильно діагностувати захворювання. Сегментація зображення відокремлює об'єкти, що представляють інтерес від фону за допомогою різних методів обробки , де значення інтенсивності використовується для поділу областей. У гістології це в основному може бути використано для виявлення ядер, строми і фону. До складного характеру патологічних образів, стандартних методів сегментації або модифікованих версій належать: приховані моделі Маркова, алгоритм вододілів, активні контури, клітинні автомати, метод Grow- Cut, а також нечіткі множини I і II типу, висівають область вирощування можуть бути використані для ідентифікації та класифікації клітин. Завдяки сегментації можна вибрати

цікаву область (АОИ) в зображенні, як клітини, ядра, пухлини і т.д., в разі патології зображень для подальшого аналізу. Сегментація на основі області виконується, якщо пухлина буде розглянута, на основі методу порогових значень, що використовуються для ідентифікації структури клітин від фонової області як об'єкт в зображенні, Юсеф Аль-Kofahi, et.al. [14], був запропонований метод сегментування клітинних ядер з переднього плану зображення за допомогою бінаризації граф-порізи на основі багато масштабного лапласіан-оф-гауссовской фільтрації дуги відстані на карті на основі вибору адаптивної шкали для виявлення ядерних насіннєві точоки. Крім того, в цитологічному зображенні аналіз, сегментація є першим кроком; перетворення Hough з алгоритмом вододіл. Для цього використовується дослідниками для автоматичної локалізації ядер [11].

1.2 Аналіз методів і алгоритмів класифікації зображень

Задача класифікації — формалізована задача, яка містить множину об'єктів (ситуацій), поділених певним чином на класи. Задана кінцева множина об'єктів, для яких відомо, до яких класів вони відносяться. Ця множина називається вибіркою. До якого класу належать інші об'єкти невідомо. Необхідно побудувати такий алгоритм, який буде здатний класифікувати довільний об'єкт з вихідної множини.

Класифікувати об'єкт — означає, вказати номер (чи назву) класу, до якого відноситься даний об'єкт.

Класифікація об'єкта — номер або найменування класу, що видається алгоритмом класифікації в результаті його застосування до даного конкретного об'єкту.

В математичній статистиці задачі класифікації називаються також задачами дискретного аналізу. В машинному навчанні завдання класифікації

вирішується, як правило, за допомогою методів штучної нейронної мережі при постановці експерименту у вигляді навчання з учителем [12].

Існують також інші способи постановки експерименту — навчання без вчителя, але вони використовуються для вирішення іншого завдання — кластеризації або таксономії. У цих завданнях поділ об'єктів навчальної вибірки на класи не задається, і потрібно класифікувати об'єкти тільки на основі їх подібності. У деяких прикладних областях, і навіть у самій математичній статистиці, через близькість завдань часто не відрізняють завдання кластеризації від завдання класифікації.

Деякі алгоритми для вирішення задач класифікації комбінують навчання з учителем і навчання без вчителя, наприклад, одна з версій нейронних мереж Кохонена — Мережі векторного квантування, яких навчають способом навчання з учителем [13]. Вся тема розділяється за певними етапами:

1 Математичне формулювання завдання

1.1 Імовірнісне формулювання завдання

1.2 Простір характеристик

2 Типи задач класифікації

2.1 Типи вхідних даних

2.2 Типи класів

Нехай X — множина описів об'єктів, Y — множина номерів класів. Існує невідома цільова залежність - відображення $y^*: X \to Y$, значення якої відомі лише на елементах кінцевої навчальної вибірки $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$. Потрібно побудувати алгоритм $a: X \to Y$, здатний класифікувати довільний об'єкт $x \in X$.

Загальнішим є імовірнісне формулювання завдання. Припускається, що множина пар «об'єкт, клас» X*Y є ймовірнісним простором з невідомою ймовірнісною мірою P. Є кінцева навчальна вибірка спостережень $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, згенеровано згідно з ймовірнісною мірою P. Необхідно побудувати алгоритм $a: X \to Y$, здатний класифікувати довільний об'єкт $x \in X$.

Характеристикою називається відображення $f:X \rightarrow D_f$, де D_f — множина допустимих значень характеристики. Якщо задані характеристики $f_1,\ldots,f_n$, то вектор $x = (f_1(x),\ldots,f_n(x))$ називається характеристичним описом об'єкта $x \in X$. Характеристики можна ототожнювати із самими об'єктами. При цьому множину $X = D_{f1} * \ldots * D_{fn}$ називають простором характеристик.

Залежно від множини D_f характеристики поділяються на такі типи:

- Бінарні характеристики: $D_f = \{0,1\}$;
- Номінальні характеристики: D_f — кінцева множина;
- Порядкові характеристики: D_f — кінцева впорядкована множина;
- Кількісні характеристики: D_f — множина дійсних чисел.

Часто зустрічаються прикладні задачі з різнотипними характеристиками, для їх вирішення підходять далеко не всі методи [14].

Характеристичний опис — найпоширеніший випадок. Кожен об'єкт описується набором своїх характеристик, які називаються ознаками. Ознаки можуть бути числовими або нечисловими.

Матриця відстаней між об'єктами. Кожен об'єкт описується відстанями до всіх інших об'єктів навчальної вибірки. З цим типом вхідних даних працюють деякі методи, зокрема, метод найближчих сусідів, метод потенційних функцій.

Часовий ряд або сигнал є послідовність вимірів у часі. Кожен вимір може представлятися числом, вектором, а в загальному випадку — характеристичним описом досліджуваного об'єкта в цей час часу.

Зустрічаються і складніші випадки, коли вхідні дані представляються у вигляді графів, текстів, результатів запитів до бази даних, і т. д. Як правило, вони приводяться до першого або другого випадку шляхом попередньої обробки даних та вилучення характеристик.

Класифікацію сигналів та зображень називають також розпізнаванням образів.

Двокласова класифікація. Найпростіший в технічному відношенні випадок, який служить основою для вирішення складніших завдань.

Коли число класів досягає багатьох тисяч (наприклад, при розпізнаванні ієрогліфів або злитого мовлення), завдання класифікації стає істотно важчим. Об'єкт може належати одночасно до декількох класів. Потрібно визначати ступінь належності об'єкта кожному з класів, звичайно це дійсне число від 0 до 1.

На сьогодні найбільш поширений і використовуваний метод дешифрування це візуальне дешифрування знімка. В цьому випадку передбачається, що дешифрування проводить експерт, який добре обізнаний з особливостями території і властивостями об'єктів, відображених на знімку [4].

Однак цей метод є трудомістким і досить тривалим, тому актуальним є дослідження способів автоматичного дешифрування (автоматичної класифікації). Автоматичною класифікацією називають процес розбиття пікселів неперервного растрового зображення на категорії на основі їх спектральних значень, в результаті чого кожному пікселю присвоюється нове значення [15].

На даний час існують два підходи у реалізації автоматичної класифікації: керована класифікація та некерована. На основі цих підходів створено багато методів, основні з яких показані на рисунку 1.2. При керованій класифікації відбувається аналіз пікселів у межах кожного еталонного полігона і створення спектральних сигнатур для кожного типу покриття. За порівнянням спектральних значень пікселів зі створеними сигнатурами виконується класифікація зображення.

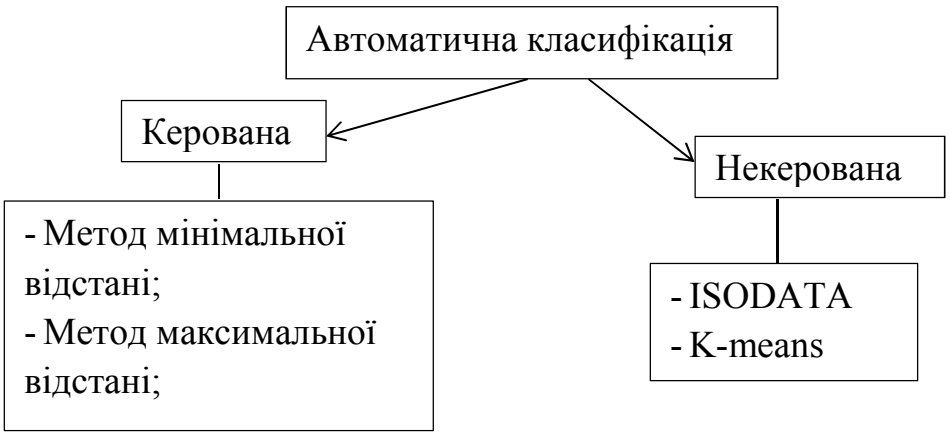

Рисунок 1.2 - Основні методи автоматичної класифікації

Класифікація за методом мінімальної відстані полягає в розрахунку евклідової відстані значень відбиття пікселя до середнього спектрального значення кожного сигнатури. Піксель призначається до класу, відстань до котрого є найменшою. Класифікація за методом максимальної вірогідності вважається однією з оптимальних, оскільки базується на імовірнісних принципах. Дисперсія значень відбиття в еталонному полігоні описується функцією імовірності щільності, яка базується на статистиці Байєса [16].

Алгоритми некерованої класифікації (їх часто називають алгоритмами кластеризації) застосовують за відсутності апріорної інформації про об'єкт зйомки. Кластерний аналіз дозволяє виділяти контури з неконтрастною по спектральній яскравості структурою, наприклад рослинність, відкриті ґрунти, воду, хмари та інші об'єкти. З використанням алгоритмів кластеризації виконується автоматичне розділення зображення на групи пікселів, подібних за спектральним характеристикам (кластери). Ці алгоритми потребують мінімум початкової інформації (число класів, кількість ітерацій) [17]. Кластеризація зображення за алгоритмом ISODATA ґрунтується на різниці між середніми значеннями кластерів (мінімальній спектральній відстані між центрами класів). [18].

Метод K-means є подібним до методу ISODATA. Головна відмінність алгоритмів ISODATA і K-means полягає в тому, що на стадії ініціалізації алгоритму ISODATA відбувається розподіл пікселів, тоді як для алгоритму K-means відбувається розподіл значень математичних очікувань. В таблиці 1.2 подані основні характеристики методів автоматичної класифікації.

Алгоритм та структурні коефіцієнти досліджено за базою із 1000 образів, згрупованих у 10 семантичних груп: леви, слони, коні, квіти, їжа, гори, автобуси, дизайн, зображення текстури, медичні образи. Кожна група містить 100 зображень у форматі JPEG із розмірами 256×384 пікселів.

Таблиця 1.2 Основні характеристики методів автоматичної класифікації

Методи	Потреба у «навчанні»	Швидкість	Переваги	Недоліки
Метод мінімальної відстані	+	Швидкий	Після класифікації немає некласифікованих пікселів.	Не враховується дисперсія між сигнатурами еталонних полігонів.
Метод максимальної відстані	+	Повільний	Вважається найбільш точним, так як працює на ймовірних принципах.	Сигнатури з великим значенням коваріації сильно підкреслюються і тому потребують нормалізації.
ISODATA	-	Швидкий	Не потребує попереднього навчання та менш залежний від людського чинника.	Невідповідність створених кластерів потрібним класам, тому вимагає подальшого об’єднання або розбиття кластерів користувачем
K-means	-	Швидкий	Досить добре працює з частково «навченими» кластерами, тобто для частини точок відомо, до якого класу вони належать.	Потрібно точно знати необхідну кількість кластерів, тому спочатку використовують інші методи кластеризації, де отримують кількість і початкове розбиття.

На алгоритми керованої класифікації доцільно застосовувати тоді, коли є додаткова інформація про об’єкти на знімку, необхідна для створення

еталонних полігонів. Вона може бути отримана з карт, планів, наземної зйомки і т.д. Якщо така інформація відсутня, то слід користуватися алгоритмами некерованої класифікації. Для підвищення точності дешифрування даних ДЗЗ часто використовують спільно і керовану і некеровану класифікації.

Приклади образів подано на рисунку. 1.3. Результати декомпозиції образів та їх сортування за кількістю регіонів ІА, коефіцієнтами структуризації 123 Ks (МС) та 123 Ks (РХ) зведені у таблицю. Також у таблиці наведені структурні властивості фрагментів образів, а саме: середній розмір фрагмента M(IA) та дисперсія розмірів фрагментів D(IA). Інтегральна класифікація образів у таблиці отримана сортуванням суми місць образів, отриманих класифікацією тільки за одним із параметрів ІА, 123 Ks (МС),123 Ks (РХ) [19].

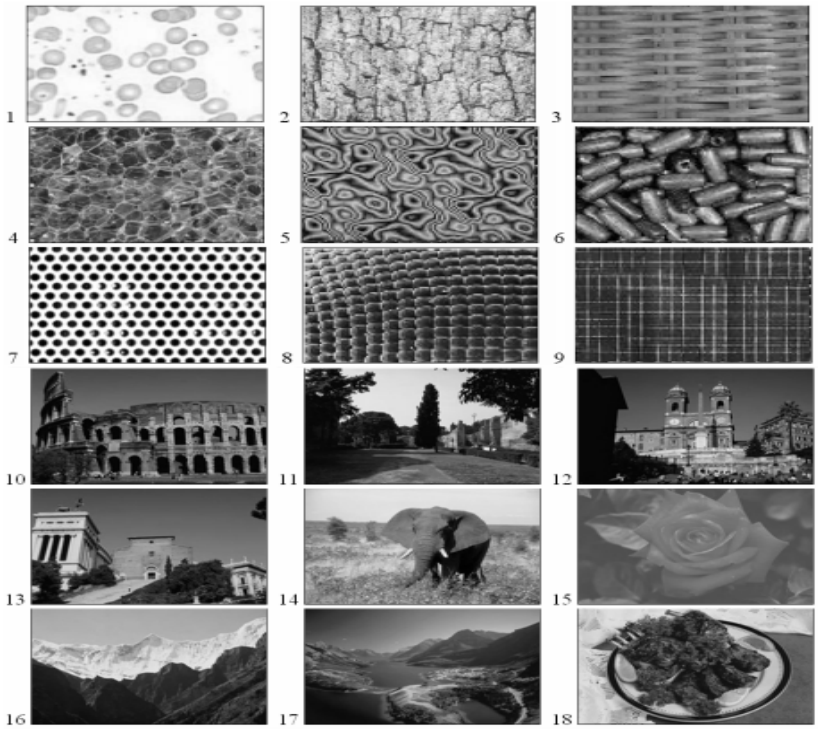

Рисунок 1.3 - Пронумеровані зразки досліджуваних образів

З таблиці отримуємо збіг класифікації деяких образів. Це означає, що ми однозначно можемо класифікувати образи текстур, медичні зразки, ландшафти або архітектурні комплекси. Комбінуючи структурні коефіцієнти та параметри об’єктів – фрагментів образів можна створити правила класифікації медичних зображень, структур матеріалів тощо. Зокрема, беручи до уваги параметри

образу M(IA) та D(IA) на рисунку 1.4, клас C можна розбити на окремий клас зображення клітин крові та зображення текстур.

Таблиця 1.2 - Класифікація образів за структурними характеристиками

Місце	LA	№ Образу	$K_5^{123}(MC$	№ Образу	$K_5^{123}(PX$	№ Образу	M(LA)	D(LA)	Інтегральна класифікація
1	178	2	0.00838	7	0.00548	2	175.43	124.91	2
2	170	4	0.00735	1	0.00523	4	191.08	166.48	4
3	148	5	0.00570	2	0.00456	5	219.49	177.61	7
4	146	7	0.00523	4	0.00449	7	119.32	244.53	5
5	143	6	0.00456	5	0.00440	6	225.80	571.13	6
6	131	14	0.00443	6	0.00403	14	247.61	376.77	14
7	130	8	0.00405	8	0.00400	8	246.79	1440.20	8
8	126	18	0.00404	14	0.00388	18	257.67	1013.51	18
9	108	9	0.00388	18	0.00332	3	300.81	194.52	3
10	108	3	0.00332	3	0.00332	9	300.78	1066.08	9
11	97	10	0.00332	9	0.00299	10	334.92	1098.05	1
12	94	13	0.00299	10	0.00289	13	345.61	717.35	10
13	91	12	0.00289	13	0.00280	12	356.74	1067.25	13
14	84	17	0.00280	12	0.00259	17	386.75	1093.51	12
15	83	11	0.00259	17	0.00255	11	388.57	1339.15	17
16	82	1	0.00257	11	0.00252	1	136.11	137.41	11
17	63	16	0.00119	16	0.00194	16	513.29	2004.53	16
18	4	15	0.00012	15	0.00012	15	8121.75	14019.37	15

Інтегральну класифікацію образів з рис. 3, наведену в таблиці, подано на рисунку 1.4.

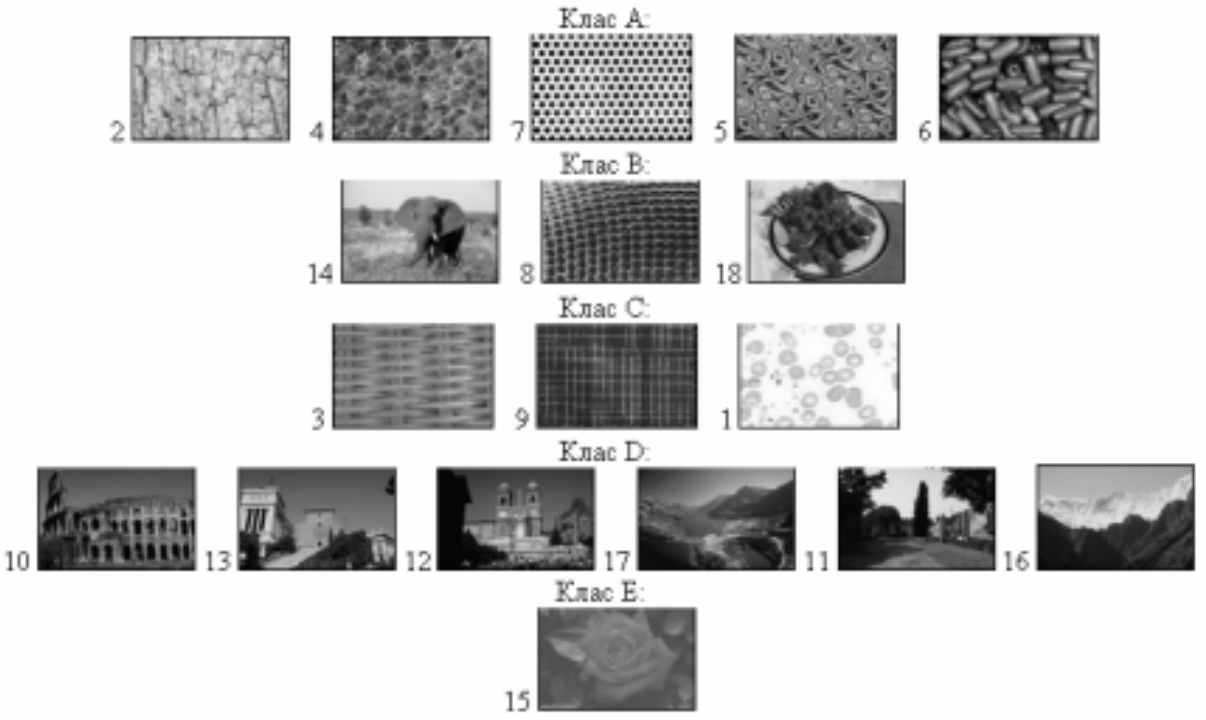

Рисунок 1.4 - Інтегральна класифікація візуальних образів

Розроблено алгоритм триступеневої кластеризації візуальних образів, який, крім виділення фрагментів різної ієрархії підпорядкованості, дає змогу отримати кількісні та якісні характеристики фрагментів та відношень між ними. Запропоновано коефіцієнти структуризації та розмитості зображення.

Експерименти підтвердили доцільність використання структурних коефіцієнтів та характеристики фрагментів для класифікації зображень. Формування ключів зображення за представленими властивостями образів рекомендуються для застосування у автоматизованих системах пошуку зображень[20].

1.3 Програмні засоби розпізнавання зображень

У процесі біологічної еволюції багато тварин за допомогою зорового й слухового апарата вирішили задачу розпізнавання образів досить добре. Створення штучних систем розпізнавання образів залишається складною

теоретичною й технічною проблемою. Необхідність у такому розпізнаванні виникає в різних областях — від військової справи й систем безпеки до оцифровування різних аналогових сигналів.

Для оптичного розпізнавання образів можна застосувати метод перебору вигляду об'єкта під різними кутами, масштабами, зсувами й т. д. Для букв потрібно перебирати шрифт, властивості шрифту й т. д. Другий підхід — знайти контур об'єкта й досліджувати його властивості (зв'язність, наявність кутів і т. д.). Ще один підхід — використовувати штучні нейронні мережі. Цей метод вимагає або великої кількості прикладів задачі розпізнавання (із правильними відповідями), або спеціальної структури нейронної мережі, що враховує специфіку даної задачі [21].

Індуктивне навчання, або навчання за прецедентами, засноване на виявленні загальних властивостей об'єктів на підставі неповної інформації, отриманих емпіричним шляхом. Дедуктивне навчання передбачає формалізацію знань експертів у вигляді баз знань (експертних систем тощо).

Ф. Розенблатт уводячи поняття про модель мозку, завдання якої полягає в тому, щоб показати, як у деякій фізичній системі, структура й функціональні властивості якої відомі, можуть виникати психологічні явища — описав найпростіші експерименти з розрізнення. Дані експерименти цілком стосуються до методів розпізнавання образів, але відрізняються тим, що алгоритм розв'язання не детермінований.

Найпростіший експеримент, на основі якого можна одержати психологічно значиму інформацію про деяку систему, зводиться до того, що моделі пред'являються два різних стимули й потрібно, щоб вона реагувала на них різним чином. Метою такого експерименту може бути дослідження можливості спонтанного розрізнення стимулів системою при відсутності втручання з боку експериментатора, або, навпаки, вивчення примусового розрізнення, при якому експериментатор прагне навчити систему здійснювати необхідну класифікацію.

У досвіді з навчанням персептрону зазвичай пред'являється деяка послідовність образів, у яку входять представники кожного із класів, що підлягають розрізненню. Відповідно до деякого правила модифікації пам'яті правильний вибір реакції підкріплюється. Потім персептрону пред'являється контрольний стимул і визначається ймовірність одержання правильної реакції для стимулів даного класу. Залежно від того, збігається чи не збігається обраний контрольний стимул з одним з образів, які використовувалися в навчальній послідовності, отримують різні результати:

Якщо контрольний стимул не збігається з жодним із навчальних стимулів, то експеримент пов'язаний не тільки з чистим розрізненням, але містить у собі й елементи узагальнення [23].

Якщо контрольний стимул збуджує деякий набір сенсорних елементів, цілком відмінних від тих елементів, які активізувалися при впливі раніше пред'явлених стимулів того ж класу, то експеримент є дослідженням чистого узагальнення.

Персептрони не мають здатності до чистого узагальнення, але вони цілком задовільно функціонують в експериментах із розрізнення, особливо якщо контрольний стимул досить близько збігається з одним з образів, щодо яких персептрон уже нагромадив певний досвід.

Але це чисто наукові теорії та експерименти які мають мало спільного із практичним застосуванням. Однак, можна навести багато сучасних програм для розпізнавання які успішно функціонують вже декілька років. Ось деякі з них: FineReader, CuneiForm, GPRSG Package.

FineReader — система оптичного розпізнавання символів, розроблена російською компанією ABBYY. Висока точність і швидкість перетворення документів.[24]

Програма швидко і з точністю розпізнає від скановані або сфотографовані документи, перетворюючи їх в електронні редаговані формати або PDF з можливістю пошуку. При розпізнаванні якісних документів швидкий режим збільшить швидкість на 40% без шкоди для точності. А для чорно-білих

документів можна використовувати також чорно-білий режим розпізнавання, який прискорить роботу ще на 30%.

Завдяки технології Adaptive Document Recognition Technology, ABBYY FineReader 12 зберігає вихідну структуру багатосторінкових документів, включаючи розташування тексту, таблиць, колонтитулів, приміток, нумерацію сторінок, змісту, та ін. Задати типи областей (Текст, Картинка, Таблиця і ін.) і вказати їх призначення, можна і вручну.

FineReader забезпечує миттєвий доступ до сторінок документа, що сканується незалежно від його розміру. Щоб почати працювати з документом, вам не потрібно чекати, поки він розізнається цілком. ABBYY FineReader 12 розпізнає документи на 190 мовах, в будь-яких комбінаціях.

ABBYY FineReader 12 вміє справлятися як з спотвореннями, характерними для цифрових фотографій (трапецієподібні спотворення, викривлення рядка, цифровий шум, і так далі), так і з дефектами зображення, пов'язаними зі станом вихідних паперових документів (пожовклий від часу папір, рукописні позначки, штампи) [25].

Програма перетворює зображення документів і PDF-файли, отримані зі сканера (без текстового шару), в формати, придатні для збереження в електронному архіві з можливістю пошуку: PDF з текстовим шаром або PDF/A.

Програма підтримує широкий набір форматів для збереження документів, необхідних вам у роботі. Можна записати результати розпізнавання в файл або відправити їх відразу в додатки Microsoft Word, Excel, PowerPoint, OpenOffice Writer та ін.

Програма підтримує збереження в найпопулярніші формати електронних книг (fb2 і ePub, також Kindle), це допоможе швидко зробити електронну копію для портативного пристрою - електронної книги, планшета, смартфона, і ін.

- Професійна версія - Версія для Microsoft Windows, призначена для індивідуальних користувачів та приватних підприємців

- Корпоративна версія - Версія для Microsoft Windows, призначена для малого і середнього бізнесу

- Pro версія для Mac - Macintosh версія для Mac OS X 10.7 і вище

- Он-лайн версія - Платна он-лайн версія FineReader.

Кожному новому зареєстрованому користувачеві надається можливість безкоштовно розпізнати 10 сторінок, та ще до 5 сторінок щомісяця. Більші кількості потребують придбання пакету відповідного об'єму [26].

ABBYY FineScanner для iOS - це додаток - "мобільний сканер", за допомогою якого можна створити електронну копію документа в форматах PDF і JPEG, а також розпізнати текст на зображенні (OCR) із збереженням форматування. Додаток обробляє одно сторінкові, так і багатосторінкові документи. За допомогою спеціальних фільтрів FineScanner поліпшує сфотографовані зображення та робить їх порівняної зі сканованими якості. Розізнається текст будь-якою з 44 підтриманих мов. Підтримано 12 вихідних форматів файлів.

CuneiForm — інструмент оптичного розпізнавання символів, розроблений російською компанією Cognitive Technologies. Програма перетворює файли зображень, отримані зі сканера або іншим шляхом на текст. Після декількох років без розвитку, 12 грудня 2007 року анонсовано відкриття джерельних текстів програми, яке відбулося 2 квітня 2008 року [27].

CuneiForm — шрифтонезалежна (OmniFont) система. Алгоритми, закладені в CuneiForm, виходять з правил написання букв, з їх топології, і не вимагають завдання яких-небудь еталонів, або навчання. Розізнаються будь-які друкарські шрифти — книги, газети, журнали, роздруківки з лазерних і матричних принтерів, тексти з друкарських машинок, і т.п. Не розізнається рукописний текст і декоративні шрифти (готичний, стилізований під рукописний). У CuneiForm існують спеціальні налаштування для розпізнавання текстів з матричного принтера і факсів 200x100 DPI.

CuneiForm зберігає форматування тексту і розпізнає складні таблиці будь-якої структури. Програма розпізнає текст англійською, болгарською, голландською, данською, естонською, іспанською, італійською, латвійською, литовською, німецькою, польською, португальською, російською, румунською,

сербською, словенською, турецькою, угорською, українською, французькою, хорватською, чеською, шведською мовами та російсько-англійський двомовний текст.

CuneiForm може зберегти розпізнаний текст у форматах RTF, HTML, або текстовому. Також можливо передати текст до текстового процесора Word або електронної таблиці Excel. Колись лідер програмного забезпечення з розпізнавання символів у Росії, CuneiForm змагався з популярною нині програмою ABBYY FineReader [28].

Cognitive Technologies розпочали програму «Розпізнавання має бути на кожному комп'ютері», щоб зробити розпізнавання доступним для всіх споживачів. Перший крок програми — випуск CuneiForm як вільно поширюваного ПЗ. Наступним кроком стане запуск вільної on-line служби розпізнавання на сайті www.cuneiform.ru до кінця січня 2008 року.

2 квітня 2008 року компанія Cognitive Technologies оголосила про відкриття джерельного коду програми. В даний час, розробники вибрали для проекту ліцензію BSD. У квітні 2009 випущена перша версія Cuneiform-Qt — графічного інтерфейсу до CuneiForm на основі бібліотеки Qt4. У червні 2009 випущена перша публічна версія YAGF — графічної оболонки до CuneiForm на основі бібліотеки Qt4.

Програмне забезпечення в рамках GPRSG пакета, який працює на будь-якому ПК, ноутбуку або Surface Pro для Microsoft Windows.

Головна версія - це програмне забезпечення початкового рівня для розпізнавання картин,. Це дає кожному шанс випробувати розпізнавання зображення. Поставляється із GPRSG Image Finder.

Основні функції професійної версії - досконале ядро розпізнавання, ручне позначення, автоматичне відзначення, пошук, об'єктний пошук, граф об’єктів є дуже корисною функцією для промислової обробки зображень, Web-Crawler, потужний органайзер, доповнена система пошуку після позначення. Інші функції не пов’язані з розпізнаванням, які прискорюють пошук. У цій версії є можливість розпізнавати лише форму об’єкта ігноруючи його колір. Крім того,

тепер є нова автоматична ітерація на Auto Tagging. Ця нова функція є фоновою і буде додавати раніше відмічувані зображення [29].

Це програмне забезпечення дозволяє користувачеві шукати дуже гарне, середнє, погане або дуже погане зображення. Тут звертається увага на такі речі, як розмиття, світло, темрява, сильні образи і колірну гаму. Зображення можуть бути копіями.

Використання GPRSG Pro дозволяє програмі автоматично отримати доступ до віртуальних дисків та хмарних ресурсів і за допомогою вмонтованого ПЗ додати наявні там зображення до бази розпізнавання. Також воно може бути використано з серверами для бізнесу.

Після того, як зображення помічені за допомогою програмного забезпечення їх можна знайти за допомогою будь-якого нормального пошуку тексту. Тому можна використовувати і створити свій власний PHP або Java скрипти які легко отримуватимуть доступ до інформації з будь-якого типу сервера або використовувати звичайні можливості пошуку, за допомогою Surface Pro Tablet [30].

1.4 Постановка завдання

За умовами поставленого завдання потрібно спроектувати гнучкий програмний модуль, який буде розпізнавати та класифікувати зображення використовуючи досягнення на попередніх рівнях комп'ютерного зору. Проектування та розробку потрібно зробити відповідно до поставлених вимог та поставлених строків.

Розпочнемо аналіз із нижчих рівнів комп'ютерного зору, оскільки класифікаціє є фінальною частиною і повинна мати теоретичний та алгоритмічний фундамент. Тому програмний модуль повинен використати найсучасніші досягнення в сегментації зображень. Опираючись на можливості використовуваного методу знайти оптимальні співвідношення для

максимальної ефективності. Також для виявлення не визначеностей та нових клітинних утворень які раніше не були класифіковані потрібно розробити обширну і фундаментальну систему навчання. В цьому згодиться наявна база біомедичних зображень та експертні оцінки спеціаліста даної галузі.

На даний час наявна маса алгоритмів як низькорівневого опрацювання зображень та і повноцінної класифікації. Однак їх результат все ще далекий від ідеалу. Якість отриманих зразків низька, що спонукає на покращення методів електронної мікроскопії та на пошук чогось кардинально нового. Це стосується не лише підготовки зразків чи сегментації зображення, адже це все дасть лише тимчасовий ефект. Ідеться про комплексний підхід в класифікації і обробці зображень. Потрібно зібрати і проаналізувати доступні найсучасніші дані їх проаналізувати виявити корисні та негативні сторони знайти шляхи покращення. Можливо доведеться кардинально переосмислити наявні дослідження.

Використавши сучасні здобутки потрібно синтезувати ефективний класифікатор, який ґрунтуватиметься на експертних оцінках та алгоритмі виявленні закономірностей. Адже знайшовши досконалий і практичний метод навчання ми зможемо мінімізувати похибки і діагностиці та прискорити час виявлення захворювання. А при таких страшних хворобах як рак кожна доля секунди є вирішальною. Інтегрований програмний модуль зможе покращити діагностику а з подальшим розвитком медичного обладнання і нових методів дослідження клітин та тканих стане тільки ефективнішим [22].

2 K-MEANS ПОДІБНІ АЛГОРИТМИ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ

2.1 Метод головних компонент

Аналіз основних компонентів (PCA) - це статистична процедура, яка використовує ортогональні перетворення для перетворення набору спостережень можливих корельованих змінних в набір значень лінійно некорельованих змінних, які називаються основними компонентами (або іноді основними варіантами). Кількість основних компонентів менше або дорівнює меншій кількості оригінальних змінних чи кількості спостережень. Це перетворення визначається таким чином, що перший основний компонент має найбільшу можливу дисперсію (тобто припускає максимальну мінливість даних), і кожен наступний компонент, у свою чергу, має найбільшу дисперсію за умови обмеження що він ортогональний до попередніх компонентів. Отримані вектори - це не корельований ортогональний базис. PCA чутливий до відносного масштабування вихідних змінних.

PCA був винайдений в 1901 році Карлом Пірсоном як аналог теореми головної осі в механіці; пізніше він був незалежним розробником і названий Гарольдом Готельлінгом у 1930-х роках [31]. Залежно від області застосування, він також називається дискретним перетворенням Кархунена-Лоєва (КЛТ) в обробці сигналів, трансформацією готелю в багатомірному контролі якості, правильному ортогональному розкладу (ПОД) в машинобудуванні, розкладі сингулярного значення (SVD) X (Golub і Van Loan, 1983), розкладання власних значень (EVD) XTX в лінійній алгебри, факторний аналіз (для обговорення відмінностей між PCA та факторним аналізом), теореми Еккарта-Юнга (Harman , 1960), теорема Шмідта-Мирського в психометрії, емпіричні ортогональні функції (ЕОФ) в метеорологічній науці, розкладання емпіричних власних функцій (Sirovich, 1987), аналіз емпіричних компонентів (Lorenz, 1956), квазігармонічні режими (Brooks et al., 1988) , спектральне розкладання шуму та вібрації, а також емпіричний модальний аналіз у структурній динаміці.

Метод в основному використовується як інструмент аналізу пошукових даних і для створення прогнозних моделей. Це часто використовується для візуалізації генетичної відстані та спорідненості між популяціями. PCA може бути здійснено шляхом декомпозиції власної величини матриці коваріації (або кореляції) даних або однозначного розкладу матриці даних, як правило, після середнього центрування (і нормалізації або використання Z-оцінок) матриці даних для кожного атрибута [32]. Результати, як правило, обговорюються в термінах компонентних балів, які іноді називаються факторними оцінками (перетворені значення змінної, що відповідають певній точці даних), і навантаження (вага, за якою кожну стандартизовану початкову змінну потрібно помножити, щоб отримати оцінку компонента) [33].

PCA - найпростіший з істинних багатомірних аналізів на основі власного вектора. Часто його функціонування можна розглядати як виявлення внутрішньої структури даних у такий спосіб, що найкраще пояснює різницю в даних. Якщо багатоваріантний набір даних візуалізується як набір координат у просторі виміру великих розмірів (1 ось на кожну змінну), PCA може надати користувачеві нижчий вимір зображення, проекцію цього об'єкта під час перегляду з його найбільш інформаційної точки зору. Це робиться за допомогою лише перших декількох основних компонентів, так що зменшується розмірність перетворених даних.

Метод головних компонент тісно пов'язана з факторним аналізом. Факторний аналіз зазвичай включає в себе більше конкретних припущень доменів про базову структуру та вирішує власні вектори дещо іншої матриці.

PCA також пов'язаний з канонічним кореляційним аналізом (CCA). CCA визначає системи координат, які оптимально описують крос-коваріацію між двома наборами даних, тоді як PCA визначає нову ортогональну систему координат, яка оптимально описує дисперсію в одному наборі даних.

Він може розглядатися як пристосування n-мірного еліпсоїда до даних, де кожна вісь еліпсоїда являє собою головний компонент. Якщо деяка вісь еліпсоїда невелика, то дисперсія уздовж цієї осі також мала, і, видаливши цю

вісь та її відповідну головну складову з нашого представлення набору даних, ми втрачаємо лише порівняно невелику кількість інформації.

Щоб знайти осі еліпсоїда, ми повинні спочатку вирахувати середнє значення кожної змінної з набору даних для центрування даних навколо походження. Потім ми обчислюємо матрицю коваріації даних і обчислюємо власне значення та відповідні власні вектори цієї матриці коваріації. Тоді ми повинні нормалізувати кожен з ортогональних власних векторів, щоб стати одиничними векторами. Як тільки це буде зроблено, кожен з взаємно ортогональних одиничних власних векторів може бути інтерпретований як ось еліпсоїда, прикріпленого до даних. Частка дисперсії, яку представляє кожний власний сектор, може бути розрахована шляхом ділення власного значення, відповідного власному вектору, на суму всіх власних значень.

Ця процедура чутлива до масштабування даних, і немає єдиної думки щодо того, як найкраще масштабувати дані для отримання оптимальних результатів.

PCA математично визначається як ортогональне лінійне перетворення, яке перетворює дані на нову систему координат так, що найбільша дисперсія за деякою проекцією даних залежить від першої координати (називається першою головною складовою частиною), друга найбільша дисперсія на друга координата тощо [34].

Розглянемо матрицю даних X, з колонкою нульового емпіричного середнього (середнє вибірки кожного стовпця було зсунуто до нуля), де кожен з n рядків являє собою інше повторення експерименту, і кожен з p-колонок дає особливий вид функції (скажімо, результати певного датчика).

Математично перетворення визначається набором p-мірних векторів ваг чи навантажень $w_{(k)} = (w_1, \dots, w_p)_{(k)}$, що карта кожного вектора рядка $\mathbf{x_{(i)}}$ з X до нового вектора основних компонентів балів $t_{(i)} = (t_1, \dots, t_m)_{(i)}$ дається $t_{k_{(i)}} = x_{(i)} * w_{(k)}$ для $i = 1, \dots, n$ $k = 1, \dots, m$ таким чином, що окремі змінні t, що розглядаються над набором даних, послідовно успадковують максимально можливу дисперсію від x, причому кожен вектор навантаження w обмежений

одиничним вектором. Щоб максимізувати дисперсію, перший вектор навантаження w (1) повинен задовольняти.

Метод головних компонент математично визначається як ортогональне лінійне перетворення, яке перетворює дані на нову систему координат так, що найбільша дисперсія за деякою проекцією даних залежить від першої координати (називається першою головною складовою частиною), друга найбільша дисперсія на друга координата тощо [35].

Розглянемо матрицю даних, X, з колонкою нульового емпіричного середнього (середнє вибірки кожного стовпця було зсунуто до нуля), де кожен з n рядків являє собою інше повторення експерименту, і кожен з p-колонів дає особливий вид функції (скажімо, результати певного датчика).

Математично перетворення визначається набором p-мірних векторів ваг чи навантажень $w_{(k)} = (w_1, \dots, w_p)_{(k)}$, що карта кожного вектора рядка x_i з X в новий вектор основних компонентів балів $t_{(i)} = (t_1, \dots, t_m)_{(i)}$ дається таким чином $t_{k_{(i)}} = x_{(i)} * w_{(k)}$ для $i = 1, \dots, n \;\; k = 1, \dots, m$, що окремі змінні t, що розглядаються над набором даних, послідовно успадковують максимально можливу дисперсію від x, причому кожен вектор навантаження w обмежений одиничним вектором.

Враховуючи набір точок в евклідовому просторі, перший основний компонент відповідає лінії, яка проходить через багатомірне середнє і мінімізує суму квадратів відстаней точок від лінії. Другий основний компонент відповідає такому ж поняттю, після того, як вся кореляція з першою основною компонентою була віднімана від точок. Одиночні значення (у Σ) - це квадратні коріння власних значень матриці XTX. Кожне власне значення пропорційно частині "дисперсії" (більш правильно від суми квадратів відстані точок від їх багатовимірного середнього), корелюючого з кожним власним вектором. Сума всіх власних значень дорівнює сумі квадратних відстаней точок від їх багатовимірного середнього. PCA по суті обертає набір точок навколо їхнього середнього, щоб узгодити його з основними компонентами. Це переміщує максимально можливу дисперсію (використовуючи ортогональне

перетворення) в перші кілька вимірів. Таким чином, значення в інших розмірах, як правило, невеликі і можуть бути зняті з мінімальною втратою інформації. PCA часто використовується таким чином для зменшення розміру. PCA має різницю між оптимальним ортогональним перетворенням для збереження підпростору, який має найбільшу "дисперсію" (як визначено вище). Проте ця перевага виникає за ціною більших обчислювальних вимог, якщо порівнювати, наприклад, і, коли це застосовується, з дискретним косинусним перетворенням, і зокрема DCT-II, який просто називається "DCT". Методи зменшення нелінійної розмірності, як правило, вимагають більшого обчислення.

PCA чутливий до масштабування змінних. Якщо у нас є лише дві змінні, і вони мають однакову дисперсію вибірки і позитивно корельовані, то PCA спричинить обертання на 45 °, а "навантаження" для обох змінних відносно основної компоненти будуть однаковими. Але якщо ми помножимо всі значення першої змінної на 100, то перший основний компонент буде майже таким же, як і ця змінна, з невеликим внеском від іншої змінної, тоді як другий компонент буде майже вирівняний з другою вихідною змінною. Це означає, що всякий раз, коли різні змінні мають різні одиниці (наприклад, температура та маса), є дещо свавільним методом аналізу. (Різні результати можна було б отримати, наприклад, замість Фаренгейта, а не за Цельсієм.) Зверніть увагу, що оригінальний документ Пірсона мав назву "На лініях та літаках найближчих до систем точок в просторі" - "в просторі" мається на увазі фізичне евклідового простору, де такі турботи не виникають. Один із способів зробити PCA менш довільним полягає в тому, щоб використовувати змінні, масштабовані таким чином, щоб мати одиничну дисперсію, шляхом стандартизації даних і, отже, використання матриці автокореляції замість матриці автоколіваріату як основи для головних компонент. Однак, це компреси (або розширює) коливання в усіх вимірах простору сигналу до дисперсії пристрою.

Середнє віднімання (a.k.a. "середнє центрування") необхідне для виконання PCA, щоб забезпечити, що перша основна складова описує напрямок максимальної дисперсії. Якщо середнє віднімання не виконується, перший

основний компонент може натомість відповідати більш-менш середнім значенням даних. Значення нуля необхідне для пошуку бази, що мінімізує середню квадратну помилку апроксимації даних [36].

Середнє центрування не є необхідним, якщо виконувати аналіз основних компонентів на матриці кореляції, оскільки дані вже центруються після обчислення кореляцій. Кореляції виходять з перехресного продукту двох стандартних балів (Z-scores) або статистичних моментів (звідси і назва: Pearson Product-Moment Correlation). Також див. Статтю Кромрі та Фостера-Джонсона (1998 р.) На тему "Середнє центрування в модерированной регресії: багато чого-небудь про ніщо".

Нейромережа з автоконектором з лінійним прихованим шаром схожа на метод головних компонент. При зближенні вагові вектори К-нейронів у прихованому шарі утворюють основу для простору, накладеного на перші основні компоненти К. На відміну від PCA, ця техніка не обов'язково буде створювати ортогональні вектори.

PCA є популярною основною технікою розпізнавання образів. Однак це не оптимізовано для сепарації класів [37]. Однак він використовувався для кількісної оцінки відстані між двома або більше класами шляхом обчислення центру маси для кожного класу в основному компонентному просторі та відображення евклідової відстані між центром маси двох або більше класів . Лінійний дискримінантний аналіз є альтернативою, оптимізованою для сепарації класів.

До недавнього часу метод головних компонент вважали різновидом факторного аналізу. Нині його відносять до групи самостійних статистико - математичних методів багатомірного аналізу. Вперше він був розроблений в 1901 р. англійським статистиком К.Пірсоном . Потім знайшов свій розвиток у працях Г.Хотеллінга, Г.Хармана, С. Рао, П.Андрукевича, С.Айвазяна і інших авторів. У нашій країні метод головних компонент одержує широке розповсюдження з появою ПЕОМ.

Як відомо, соціально-економічне явище можна характеризувати цілим рядом ознак. При великому наборі таких ознак в кореляційно -регресійному аналізі вплив зв'язків стає затрудненим, тому виникає необхідність стиснення інформації, тобто опис досліджуваного явища (об'єкта) більш укрупненими показниками, так званими " головними компонентами". Вихідним ступенем тут є кореляційна матриця, на підставі якої з використанням методу головних компонент може бути продовжено аналіз значень спостережуваних ознак.

Правильно відібрані в кореляційну модель ознаки, як правило, пов'язані між собою. Наявність таких зв'язків між ними дозволяє на основі одного фактора мати інформацію про інший. Існування тісного зв'язку між ознаками дає підставу для виключення однієї з них. Наприклад, якщо в модель урожайності включено дві змінні х¡ і х2, які характеризують грошові витрати на гектар, перша - всі види, друга - затрати на добрива. Тут практично буде зайвим при включенні в модель ознаки х¡ досліджувати також і ознаку х2, оскільки вона тісно пов'язана з першою . Ідея обліку однієї ознаки на підставі другої лежить в основі методу головних компонент. Слід відзначити, що мова не йде тільки про дві ознаки. У такому випадку метод головних компонент малоефективний. Його використовують, як правило, при десятках взаєпов'язаних ознак. При цьому ставиться мета " набрати" певну частину загальної варіації результативної ознаки мінімальною кількістю змінних. Останні підбирають до тих пір, поки сума їх дисперсій не сягатиме заданої частки у дисперсії досліджуваного явища [38].

Метод головних компонент розв'язує такі завдання:

1. Відшкодування скритих, об'єктивно існуючих закономірностей у зміні явищ.

2. Характеристика явища, що вивчається, числом ознак, значно меншим взятих, на початковому етапі. Число головних компонент, виділених в процесі дослідження, буде вміщувати (у компактній формі) більше інформації, ніж початково виміряні ознаки.

3. Виявлення ознак, найбільш тісно пов'язаних з головною компонентою. Інакше кажучи, вивчення стохастичного зв'язку між ними (зв'язок, при якому зі зміною однієї змінної змінюється закон розподілу другої).

4. Прогнозування рівнів досліджуваних явищ на підставі рівняння регресії, яке одержане по інформації головних компонент.

Переваги такого методу прогнозування на відміну від класичного регресійного аналізу можна пояснити тим, що при останньому в модель намагаються включити максимально можливу кількість факторів, які в економічних явищах часто характеризуються істотною корельованістю (мультилінеарністю). Прогноз за такими змінними, як правило, буває не точним. Тому виникає завдання про заміну вихідних взаємопов'язаних змінних сукупністю некорельованих параметрів. Це завдання вирішується математичним апаратом - методом головних компонент, який являє собою характеристики, побудовані на підставі первинно виміряних ознак.

Реалізація практичних можливостей зазначених вище завдань, які вирішуються методом головних компонент у галузі економіки, може бути представлена різними напрямами. Назвемо їх .

1. Аналіз причинно - наслідкових взаємозв'язків показників і встановлення їх стохастичного зв'язку з головними компонентами.

2. Виділення узагальнюючих економічних показників.

3. Ранжирування результатів спостережень по головних компонентах

4. Класифікація об'єктів спостереження.

5. Список вихідної інформації.

6. Побудова рівнянь регресії за узагальнюючими економічними показниками.

Як негативну сторону методу головних компонент слід назвати складність математичного апарату, зумовленого абсолютністю знань теорії ймовірностей, математичної статистики, лінійної алгебри, а також математичного забезпечення ПЕОМ. Формальне використання стандартних програм без розуміння математичної суті обчислювальних процедур може

призвести до необґрунтованих висновків. Слід також пам'ятати про професіональні знання суті досліджуваних економічних явищ. Тільки за таких умов метод головних компонент може стати могутнім математичним засобом пізнання існуючих ролей у галузі соціально - економічних явищ.

2.2 Алгоритм k-means

Кластеризація методом k-сере́дніх — популярний метод кластеризації, — впорядкування множини об'єктів в порівняно однорідні групи. Винайдений в 1950-х роках математиком Гуґо Штайнгаузом і майже одночасно Стюартом Ллойдом. Особливу популярність отримав після виходу роботи МакКвіна.

Мета методу — розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції (2.1)

$$\sum_{t=1}^{N} d(x_i, m_j(x_i))^2 \qquad (2.1)$$

де d — метрика, x_i — i-ий об'єкт даних, а $m_j(x_i)$ — центр кластера, якому на j-ій ітерації приписаний елемент x_i. Термін «k-середніх» був уперше вжитий Джеймсом МакКвіном (англ. James MacQueen) у 1967 році, хоча ідею методу вперше озвучив Гуґо Штайнгауз (англ. Hugo Steinhaus) у 1957 році[1]. Стандартний алгоритм був вперше запропонований Стюартом Лойдом (англ. Stuart Lloyd) у 1957 р.

Маємо масив спостережень (об'єктів), кожен з яких має певні значення по ряду ознак. Відповідно до цих значень об'єкт розташовується у багатовимірному просторі.

Дослідник визначає кількість кластерів, що необхідно утворити. Випадковим чином обирається k спостережень, які на цьому кроці вважаються

центрами кластерів. Кожне спостереження «приписується» до одного з n кластерів — того, відстань до якого найкоротша.

Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об'єктів, що входять у цей кластер

Відбувається така кількість ітерацій (повторюються кроки 3-4), поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об'єкти), дисперсія всередині кластера буде мінімізована, а між кластерами — максимізована.

Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 2 кластери, далі 3,4,5, порівнюючи отримані результати (Рисунок 2.1) [39].

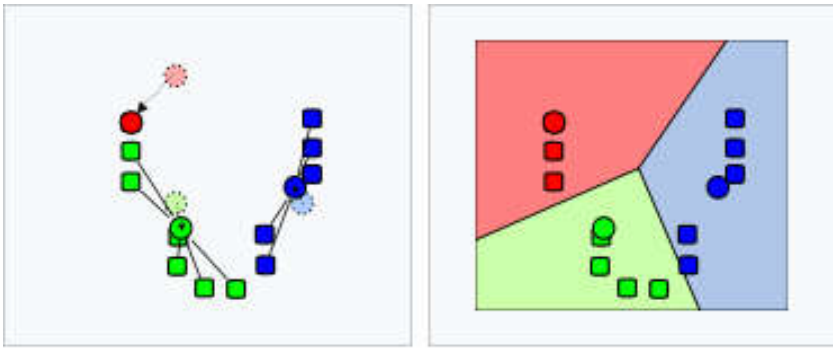

Рисунок 2.1 – Схема роботи методу

Принцип алгоритму полягає в пошуку таких центрів кластерів та наборів елементів кожного кластера при наявності деякої функції Ф(°), що виражає якість поточного розбиття множини на k кластерів, коли сумарне квадратичне відхилення елементів кластерів від центрів цих кластерів буде найменшим (2.2)

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad (2.2)$$

де k — число кластерів, S_i — отримані кластери, $i = 1,2,\dots,k$, μ_i — центри мас векторів $x_j \in S_i$.

В початковий момент роботи алгоритму довільним чином обираються центри кластерів, далі для кожного елемента множини ітеративно обраховується відстань від центрів з приєднанням кожного елемента до кластера з найближчим центром. Для кожного з отриманих кластерів обчислюються нові значення центрів, намагаючись при цьому мінімізувати функцію Ф(°), після чого повторюється процедура перерозподілу елементів між кластерами.

Алгоритм методу «Кластеризація за схемою к-середніх»:

- вибрати k інформаційних точок як центри кластерів поки не завершиться процес зміни центрів кластерів;
- зіставити кожну інформаційну точку з кластером, відстань до центра якого мінімальна;
- переконатися, що в кожному кластері міститься хоча б одна точка. Для цього кожний порожній кластер потрібно доповнити довільною точкою, що розташована «далеко» від центра кластера;
- центр кожного кластера замінити середнім від елементів кластера;
- кінець.

Головні переваги методу k-середніх — його простота та швидкість виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень, ніж метод ієрархічного кластерного аналізу (у якому дендограми стають перевантаженими і втрачають наочність).

Одним із недоліків простого методу є порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги (англ. fuzzy k-means methods), у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).

Незважаючи на очевидні переваги методу, він має суттєві недоліки:

- Результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів.
- Алгоритм чутливий до викидів, які можуть викривлювати середнє.

- Кількість кластерів повинна бути заздалегідь визначена дослідником.

2.3 Алгоритм Hard c-means

Строгий алгоритм С середніх (Hard C-Means, HCM; Jang, Sun and Mizutani, 1997.) один з найбільш широко використовуваних алгоритмів кластеризації даних. Алгоритм розбиває великі набори векторів в багатовимірному просторі за заздалегідь заданій кількості кластерів (тобто є керованим).

На вхід алгоритму подається набір векторів і кількість кластерів. Додатково може здаватися порогова величина об'єктної функції, при досягненні якої алгоритм зупиняється. Після закінчення роботи алгоритму кожен вектор буде віднесе до певного кластерному центру. Алгоритм, будучи ітераційним, залежить від початкових значень кластерних центрів, а отже може не сходитися до локального мінімуму.

На першому кроці алгоритму відбувається ініціалізація кластерним центрів, це можна зробити вибравши випадковим чином вектора з вхідного набору. На наступному кроці починається ітераційний процес, що завершується при досягненні порогового значення об'єктної функцією або при мінімальному відміну, від значення на попередній ітерації [40].

Обчислення, що виконуються на черговій ітерації алгоритму:

- Розрахунок чергової матриці (зіставлення вектора відповідного кластерному центру).

- Розрахунок об'єктної функції.

- Перерахунок кластерних центрів.

Вихідні дані: матриця $U \in R^{n*m}$, що представляє собою набір векторів $\{u_i\}_{t=1}^{n}$, $u_i \in R^m$; кількість кластерів $C \in N$. Обчислювані дані: набір векторів - кластерних центрів $\{c_i\}_{t=1}^{C}$, $c_i \in R^m$; матриця $M \in R^{n*C}$; з інформацією про

розбиття об'єктів на кластера. Елемент $m_{ij} = 1$, якщо u_i є C_i , де C_i містить всі об'єкти u_i, належать кластеру і. Інші елементи $m_{ij} = 0$.

Алгоритм:

Крок 1. Ініціалізація кластерних центрів $c_i, i = 1,2, \dots, C$. Це можна зробити вибравши випадковим чином С векторів з вхідного набору $\{u_i\}_{i=1}^{n}$.

Крок 2. Обчислення рядовий матриці М. Матриця М складається з елементів $M = ||m_{ik}||$ (2.3).

$$m_{ik} = \begin{cases} 1, ||u_k - c_i||^2 \leq \left||u_k - c_j\right|^2 \text{ для всіх } i \neq j; \ (i, j = 1,2, \dots, C; k = 1,2, \dots, n) \\ 0, \text{інші.} \end{cases} \quad (2.3)$$

Крок 3. Розрахунок об'єктної функції (2.4):

$$J = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \left(\sum_{k:\, u_k \in C_i} ||u_k - c_i||^2 \right) \quad (2.4)$$

На цьому кроці відбувається зупинка і вихід з циклу, якщо отримане значення нижче порогової величини або отримане значення не сильно відрізняється від значень, отриманих на попередніх циклах.

Крок 4. Перерахунок кластерних центрів.

Перерахунок кластерних центрів виконується у відповідності з наступним рівнянням (2.5):

$$c_i = \frac{1}{|C|_i} \sum_{k:\, u_k \in C_i} u_k \quad (2.5)$$

де $|C|_i$- кількість елементів в і-му кластері.

Крок 5. Перехід на крок 2.

Обчислювальний ядро алгоритму Hard C-Means можна скласти з множинних (всього їх nC) обчислень квадратів норми різниці векторів u_k і кластерних центрів c_i $||u_k - c_i||^2$. Як записано і в описі ядра алгоритму, основну частину методу становлять множинні (Cn) обчислення квадратів норм (2.6):

$$||u_k - c_i||^2 \quad (2.6)$$

Також присутні дані обчислювальні операції:

- Підсумовування квадратів норм векторів (2.7):

$$J = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \left(\sum_{k:\, u_k \in C_i} ||u_k - c_i||^2 \right) \quad (2.7)$$

- По елементне підсумовування векторів (2.8):

$$c_i = \frac{1}{|C|_i} \sum_{k:\, u_k \in C_i} u_k \quad (2.8)$$

Припускаємо, що гранична точність алгоритму досягає менш, ніж n ітерацій. Інакше використання алгоритму можна вважати нераціональним. Також припустимо, що розмірність m і кількість кластерних центрів C істотно менше n.

Операція обчислення квадратної норми різниці векторів вимагає виконання m-1 арифметичних операцій додавання, m операцій віднімання і m операцій множення. Кожна ітерація алгоритму вимагає виконання:

- nC операцій обчислення квадрата норми різниці векторів;
- nC операцій порівняння квадрата норми;
- 3n операцій додавання;

Таким чином, на кожній ітерації маємо:

- mnC операцій множення;
- (2mC + 3) n арифметичних операцій додавання і віднімання;

У підсумку маємо складність послідовної реалізації алгоритму O (nmC). Граф алгоритму складається з 4 видів вершин(Рисунок 2.2):

- Перша група вершин відзначена символом M. Їй відповідає обчислення мінімальної відстані до кластерних центрів, а також запис одиниці в відповідний стовпець матриці, що позначає приналежність кластерному центру:

- незалежне обчислення відстаней від заданого елемента u_i до кожного з кластерних центрів;

- послідовне порівняння отриманих значень з метою відшукання максимуму;

- операція запису 1 в відповідний стовпець матриці приналежності кластерним центрам;

- Друга вершина відзначена символом J, на цьому етапі відбувається підсумковий розрахунок об'єктної функції.

- Третя вершина позначена if, на цьому етапі відбувається перевірка умови досягнення необхідної точності.

- Четверта відповідає операції перерахунку кластерних центрів, вона відзначена як C (Рисунок 2.2).

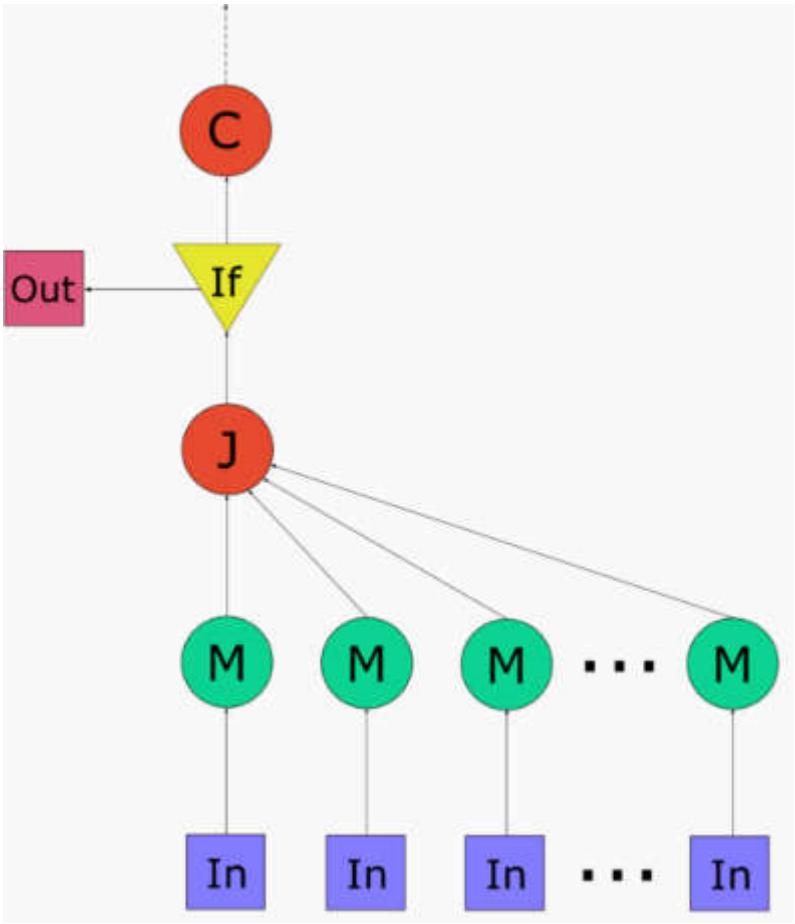

Рисунок 2.2 – Граф роботи алгоритму.

Основний ресурс паралелізму алгоритму становить ітераційний блок. Ця послідовність ітерацій може бути виконана для кожного вхідного вектора u_i незалежно. Однак, в кінці виконання кожного ітераційного блоку мають місце послідовні обчислення - складання значень об'єктної функції, отриманих незалежно на кожному процесорі, а також перерахунок кластерних центрів.

Число ітерацій алгоритму оцінимо як O(nC). Виходячи з того, що складність обчислювального блоку M дорівнює $\log_2(C)$, а складність розрахунку функціоналу J становить $\log_2(n)$, маємо, що висота ітерації складає $\log_2(nC)$. Таким чином, висота ЯПФ оцінюється як O ($nC\log_2(nC)$) [41].

Ширина ЯПФ дорівнює n. Необхідно враховувати, що при невеликому обсязі вхідної вибірки, паралельна реалізація алгоритму може лише знизити продуктивність. В першу чергу це пов'язано з пересилкою даних для обчислення підсумкових значень об'єктної функції і кластерних центрів.

Вхідні дані: матриця $U \in R^{n*m}$, що представляє собою набір векторів $\{u_i\}_{t=1}^{n}$, $u_i \in R^m$; кількість кластерів $C \in N$. Обсяг вхідних даних: nm + 1. Вихідні дані: набір векторів - кластерних центрів $\{c_i\}_{t=1}^{C}$, $c_i \in R^m$; матриця $M \in N^{n*C}$, з інформацією про розбиття об'єктів на кластера. Елемент $m_{ij} = 1$, чи $u_i \in C_i$, де C_i містить всі об'єкти u, що належать кластеру i. Інші елементи $m_{ij} = 0$.

Обсяг вихідних даних: (n + m) C. Співвідношення послідовної і паралельної складності алгоритму: лінійне. Обчислювальна потужність алгоритму, як відношення числа операцій до сумарного обсягу вхідних і вихідних даних дорівнює константі. Алгоритм є недетермінованим, тому що він є ітераційним з виходом по точності, а також використовує датчик випадкових чисел для ініціалізації кластерних центрів. Проведемо дослідження масштабованості паралельної реалізації суворого алгоритму C середніх відповідно до методики.

Дослідження проводилося на 1U сервері з наступними характеристиками:

- x Intel® Xeon® Processor E5-2697 v2, всього 24 фізичних ядра (48 логічних ядра при включеному Hyper-Threading);

- 128GB RAM;

Збірка здійснювалася з наступними параметрами:

- GCC-4.8.4;

- OpenMP-3.1;

- аргументи компілятора: -std = c ++ 11 -O3;

В результаті проведених експериментів було отримано такий діапазон ефективності реалізації алгоритму (відношення реальної продуктивності програми до пікових показників роботи обчислювальної системи):

- мінімальна ефективність реалізації 0.014;

- максимальна ефективність реалізації 0.9;

На рисунку 2.3 приведений графік продуктивності і ефективності обраної реалізації HCM в залежності від змінних параметрів запуску.

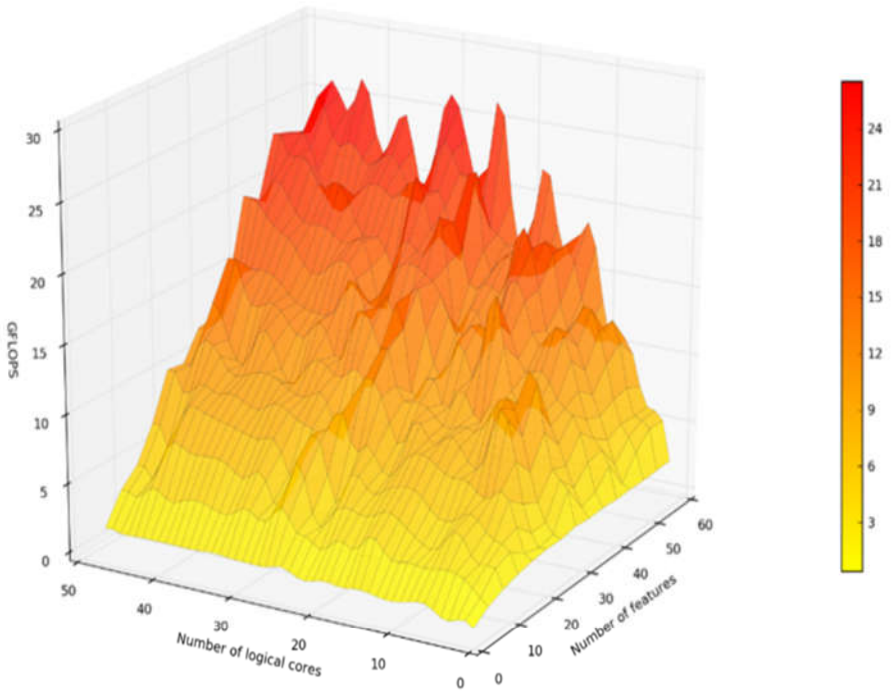

Рисунок 2.3 – Паралельна реалізація алгоритму. Зміна продуктивності в залежності від числа процесорів і розмірності векторів.

Тож як бачимо цей алгоритм є вузькоспеціалізований і призначений для роботи із великим обсягом даних, щоб нейтралізувати цей недолік в оригінального методу K-means.

Переваги:

- легкість реалізації;
- обчислювальна простота;
- може служити для кластеризації великих наборів числових даних;

Недоліки:

- завдання кількості кластерів;
- відсутність гарантії в знаходженні оптимального рішення;

2.4 Алгоритм Farthest First

У обчислювальної геометрії, найдальший перший обхід обмеженого метричного простору є послідовністю точок в просторі, де вибираються перша точка довільно і кожна наступна з набору попередньо обраних. Така концепція може бути застосована до кінцевого набору геометричних точок, шляхом їх обмеження, що належать метричному простору породженому ними ж. Для кінцевого метричного простору або кінцевої множини геометричних точок, отримана послідовність утворює перестановку, відому як жадібна перестановка.

Farthest має безліч застосувань, в тому числі апроксимація завдання комівояжера і метричної задачі к-центрі. Вони можуть бути побудовані за поліноміальний час, або (для маломірних евклідових просторів) апроксимують майже лінійний час.

Фіксоване число k, розглянута підмножина, утворене першими k точками Farthest First обходу будь-якого метричного простору. Нехай r відстань між кінцевою точкою префікса і набором попередньо вибраних точок. Тоді це підмножина має наступні дві властивості:

- всі пари вибраних точок знаходяться на відстані щонайменше, r один від одного;
- всі точки метричного простору знаходяться на відстані не більше r від підмножини.

Іншими словами, кожен префікс Farthest First обходу утворює множину Delone [42].

Перше використання Farthest First обходу було зроблено Розенкранцом, Stearns & Lewis (1977) у зв'язку з евристикою для завдання комівояжера. Farthest-insertion евристики, обговорювалися Розенкранцом та ін., Тур будується поступово, шляхом додавання однієї точки в той час, відповідно до замовлення заданого Farthest First обходом. Для того, щоб додати кожну точку в тур попередніх пунктів комівояжера, ця евристика враховує всі можливі способи злому одного кута туру, замінивши його двома кутами через нову точку, і вибирає найоптимальнішу з цих замін. Незважаючи на це, Розенкранц та ін. довести лише логарифмічний коефіцієнт апроксимації для цього методу, однак дослідження показують, що на практиці метод часто працює краще, ніж інші вставки з більш доказовими коефіцієнтами апроксимації.

Пізніше, ту ж послідовність точок популяризував Гонсалес (1985), який використовував її як частину жадібного алгоритму апроксимації для завдання знаходження K кластерів, які зводять до мінімуму максимальний діаметр кластера. Алгоритм застосовується , з такою ж якістю наближення, до метриці проблем k-середнього. Ця проблема є однією з декількох композицій, кластерного аналізу і розташування об'єкта, щоб розбити заданий набір точок на k-різних кластерів, кожному з яких відповідає центральна точка, таким чином, що максимальна відстань від будь-якої точки до центру її кластера зведиться до мінімуму. Наприклад, ця проблема може бути використана для моделювання розміщення пожежних станцій в межах міста, для того, щоб гарантувати, що до кожної адреси в межах міста можна швидко дістатися. Гонсалес описав кластерні евристики, які вибирають в якості центрів перші k-точок на Farthest First обході, а потім привласнює кожні з вхідних точок до найближчого центру. Якщо r відстань від безлічі k-обраних центрів до наступної точки в положенні k+1 обході, то ця кластеризація досягає відстані r (Рисунок 2.4). Проте, підмножина k-центрів разом з наступною точкою вже на відстані щонайменше, r один від одного, і будь-яка k-кластеризація поставить

дві з цих точок в один кластер, так що немає кластеризації з відстанню кращою, ніж r/2 . Таким чином, евристика Гонсалеса дає співвідношення апроксимації 2 для цього завдання. Ця евристика, і названа «Farthest First обхід», часто помилково відноситься до іншого документу з того ж самого часу, Hochbaum & Shmoys (1985). Проте, Hochbaum і Shmoys використовували теорії графів, а не Farthest First обхід, щоб отримати інший алгоритм апроксимації для метрики к-центру з тим же відношенням апроксимації як у евристики Гонсалеса. Для обох мін-макс завдання кластеризації діаметра і метричної задачі к-центрів, ці наближення є оптимальними: існування поліноміального часу евристики з будь-яким постійним відношенням апроксимації менше 2 означатиме, що P = NP.

А також для кластеризації Farthest First обхід може бути використаний в іншому типі завдання розміщення. Проблема дисперсії об'єкта макс-мін, в якій мета полягає в тому, щоб вибрати розташування K-різних об'єктів, так що вони були так далеко один від одного як тільки це можливо. Більше того, мета цього завдання полягає у виборі k-точок із заданого метричного простору або заданої множини точок-кандидатів, таким чином, щоб максимізувати мінімальну попарну відстань між ними. Знову ж таки, це може бути апроксимація шляхом вибору перших k-точок в Farthest First обході. Якщо r позначає відстань від k-тої точки всіх попередніх точок, то кожна точка метричного простору або набору кандидатів знаходиться в межах відстані r перших k-1 точок. За принципом Діріхле, деякі дві точки оптимального рішення обидві повинні бути в межах відстані r тої ж точки серед цих перших k-1 і в межах відстані 2r один від одного. Таким чином, евристичне рішення дається Farthest First обходу є в два рази оптимальнішим [43].

Інші застосування Farthest First обходу включає в себе колірне квантування (кластеризація кольору в зображенні) та прогресивну розгортку зображень (вибір порядку для відображення пікселів зображення таким чином, щоб префікси впорядкування давали хороші версії низької роздільної здатності всього зображення, а не заповнення зображення зверху вниз), вибір точки в

імовірнісному методі дорожньої карти для планування руху, спрощення хмар точок, що генерують маски для напівтонових зображень, ієрархічна кластеризація, знаходження схожості між полігональними сітками подібних поверхонь, виявлення несправностей в сенсорних мережах, моделювання філогенетичної різноманітності, рівномірний розподіл геодезичних обсерваторій на Землі або інших типів мережі датчиків.

Farthest First обхід кінцевої множини точок може бути обчислений за допомогою жадібного алгоритму, який підтримує відстань кожної точки з раніше обраними точками, виконуючи наступні кроки:

Ініціалізувати послідовність обраних точок на порожню послідовність і відстань кожної точки до обраної точки до нескінченності.

Переглянемо список поки що не обраних точок, щоб знайти точку p, яка має максимальну відстань від обраної точки. Видалити p із поки-що не обраних точок і додати її в кінець послідовності вибраних точок. Для кожних поки ще не обраних точок q, замініть відстаней збережених для q мінімумом його старого значення і відстанню від p до q.

Для набору n точок, цей алгоритм вимагає $O(n^2)$ кроків і $O(n^2)$ відстаней для обчислень. Більш швидкий алгоритм апроксимації, що задає Хар-Пелед і Мендель (2006), відноситься до будь-якої підмножини точок в метричному просторі з обмеженим подвоєнням розмірністю, класом просторів, які включаються в евклідовий простір обмеженої розмірності. Їх алгоритм знаходить послідовність точок, в яких кожна подальша точка має відстань в межах 1 - ε фактор на далекій відстані від раніше обраної точки, де ε може бути обраний будь-яким позитивним числом. Це вимагає часу O(nlogn).

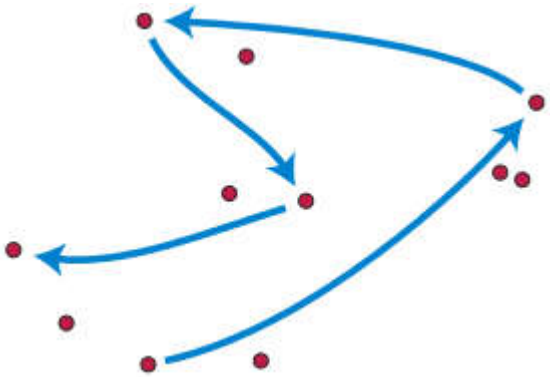

Рисунок 2.4 – Обхід множини точок.

Для вибору точок з безперервного простору, такого як евклідова площина, а не кінцева множина точок-кандидатів, ці методи не будуть працювати безпосередньо, тому що було б нескінченне число відстаней для підтримки. Замість цього кожен новий пункт повинен бути обраний в якості центру найбільшого порожнього кола, визначеного раніше. Цей центр завжди буде лежати на вершині діаграми Вороного вже обраних точок, або в точці, де край діаграми Вороного перетинає кордон домена. У цьому формулюванні метод побудови Farthest First обходів також називають інкрементними вставки Вороного. Це аналогічний алгоритму Ruppert для створення сітки кінцевих елементів, але відрізняється від вибору вершини для вставки на кожному кроці.

2.5 Аналіз K-median при роботі із спотвореннями

В даних статистики і видобутку корисних копалин, K-медіан є алгоритмом кластерного аналізу. Ця зміна K-середніх, де замість обчислення середнього значення для кожного кластера, щоб визначити його центр ваги, ми обчислюємо його медіану. Це дає ефект мінімізації помилки по всіх кластерам по відношенню до 1-норми відстані у метриці, на відміну від квадрата 2-норми відстані метрики (що робить K-середній.)

Це відноситься безпосередньо до к-медіані, яка виконує завдання знаходження K центрів таким чином, що кластери, утворені ними були

найбільш компактними. Формально задано множину точок x, К-центрам C_I повинні бути обрані таким чином, щоб мінімізувати суму відстаней від кожного x до найближчого C_I.

Функція формулює критерій таким чином, що він перевершує метод кластеризації к-середніх, в якому використовується сума квадратів відстаней. Сума відстаней широко використовується в таких додатках, як місце розташування об'єкта [44].

Пропонований алгоритм використовує Lloyd-стиль ітерації, який чергується між очікуванням (E) і максимізацією (M), що робить цей алгоритм Expectation-максимальним. На стадії E, всі об'єкти віднесені до їх найближчих медіан. На стадії M, медіани перераховуються за допомогою медіани в кожному окремому вимірі.

Медіана обчислюється в кожному окремому вимірі в композиції Manhattan-відстані при формулюванні завдання k-медіан, тому окремі атрибути будуть надходити з набору даних. Це робить алгоритм більш надійним для дискретних або навіть довічних наборів даних. На відміну від цього, використання k-середніх або евклідових медіан не обов'язково буде давати окремі атрибути з набору даних. Навіть при формулюванні Manhattan-відстані, окремі атрибути можуть надходити з різних примірників в наборі даних; Таким чином, отриманий в результаті середній показник не може бути членом вхідного набору даних(Рисунок 2.5).

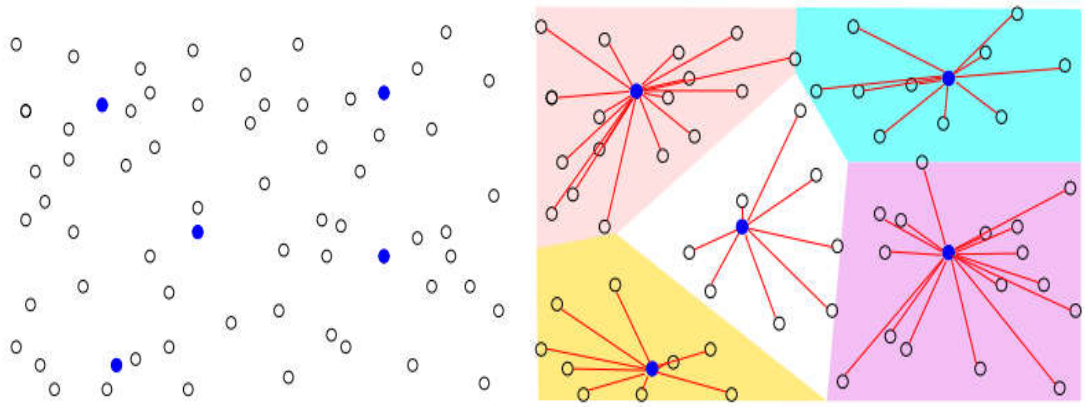

Рисунок 2.5 – Робота алгоритму.

Цей алгоритм часто плутають з алгоритмом к-medoids. Проте, медоід повинен бути фактичним екземпляром з набору даних, в той час як для багатоваріантного Manhattan-відстані медіани справедливі тільки для значень одного атрибута. Таким чином, фактична Медіана може бути комбінацією декількох екземплярів. Наприклад, якщо вектори (0,1), (1,0) і (2,2), Manhattan-відстані медіана (1,1), якої не має у вихідних даних, отже, це не медоід.

2.6 Комплексний порівняльний аналіз K-means подібних алгоритмів

Після детального аналізу та порівняння досліджуваних алгоритмів ми повинні дійти єдиного рішення. Оцінити та вибрати той алгоритм який найкраще задовольняє поставлені вимоги та підходить для реалізації програмного модуля. Адже використання всіх варіацій є трудомістким та нерентабельним процесом. Тож розглянемо короткі резюме кожного алгоритму і на базі цього побудуємо дерево рішень. В даному графі (дереві) буде відображено всі потрібні нам властивості. Використавши програмний засіб Optimizer ми отримаємо оптимальний варіант для шуканого алгоритму.

Загальна ідея алгоритмів * -means: Мінімізація відстаней між об'єктами в кластерах. Зупинка відбувається, коли мінімізувати відстані більше вже неможливо. Мінімізується функція в разі k-means така: - об'єкт кластеризації (точка) - центр кластера (центроїд). | X | = N, | C | = M

На момент старту алгоритму має бути відомо число C (кількість кластерів). Вибір числа C може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції [45].

Початковий розподіл об'єктів по кластерам. Вибираються З точок. На першому кроці ці точки вважаються центрами кластерів. Вибір початкових центроїдів може здійснюватися шляхом підбору спостережень для максимізації початкової відстані, випадковим вибором спостережень або вибором перших спостережень.

1. ітеративна перерозподіл об'єктів по кластерам. Об'єкти розподіляються по кластерам шляхом підрахунку відстані від об'єкта до центрів кластерів і вибору найменшого.

2. Коли всі об'єкти розподілені по кластерам, заново вважаються їх центри. (Можна вважати по кожній координаті окремо)

3. Якщо cj = cj - 1, то це означає, що кластерні центри стабілізувалися і відповідно розподіл закінчено. Інакше переходимо до кроку 1.

Складним є вибір числа кластерів. У разі, якщо припущень немає, зазвичай роблять кілька спроб, порівнюючи результати (скажімо, спочатку 2, потім 3 і т.д.). Перевірка якості кластеризації Після отримань результатів кластерного аналізу методом k-середніх слід перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього розраховуються середні значення для кожного кластера. При гарній кластеризації повинні бути отримані сильно відрізняються середні для всіх вимірювань або хоча б більшої їх частини. Переваги алгоритму k-середніх:

- простота використання;
- швидкість використання;
- зрозумілість і прозорість алгоритму.

Недоліки алгоритму k-середніх:

- алгоритм занадто чутливий до викидів, які можуть спотворювати середнє.
- Можливим вирішенням цієї проблеми є використання модифікації алгоритму - алгоритм k-медіани;
- алгоритм може повільно працювати на великих базах даних. можливим рішенням даної проблеми є використання вибірки даних.

Більш суворої інтерпретацією цього алгоритму є алгоритм hard c-means. Його відмінності - в мінімізується і строгості самого алгоритму:

uij = 1, якщо, і uij = 0 ,, якщо немає. Тобто мінімізується відстань від точок до центроїда, а не від центроїда до точок.

Farthest First - ще одна модифікація k-means, особливістю його є початковий вибір Центроїд - від 2 і вище вони вибираються за принципом віддаленості від інших (центроїдом вибирається точка, найбільш віддалена від інших Центроїд) [46].

Для того щоб вибрати оптимальний алгоритм для реалізації програмного модуля класифікації біомедичних зображень, було побудовано дерево рішень І-АБО. Позначкою І позначається ті елемент з яких вибірки не буде проводитись. Позначкою

АБО позначаються вибірки, які показують з яких елементі дерева рішень буде складатись підсистема після вибірки (Рисунок 2.6). Дерево рішень відображає загальну структуру програмного, однак ця структуру не враховує специфіки нашого модуля і не є оптимальною. Для того щоб спрощення нашої структури і отримання оптимальних рішень ми використовуємо спеціалізоване програмне забезпечення Optimizer [47].

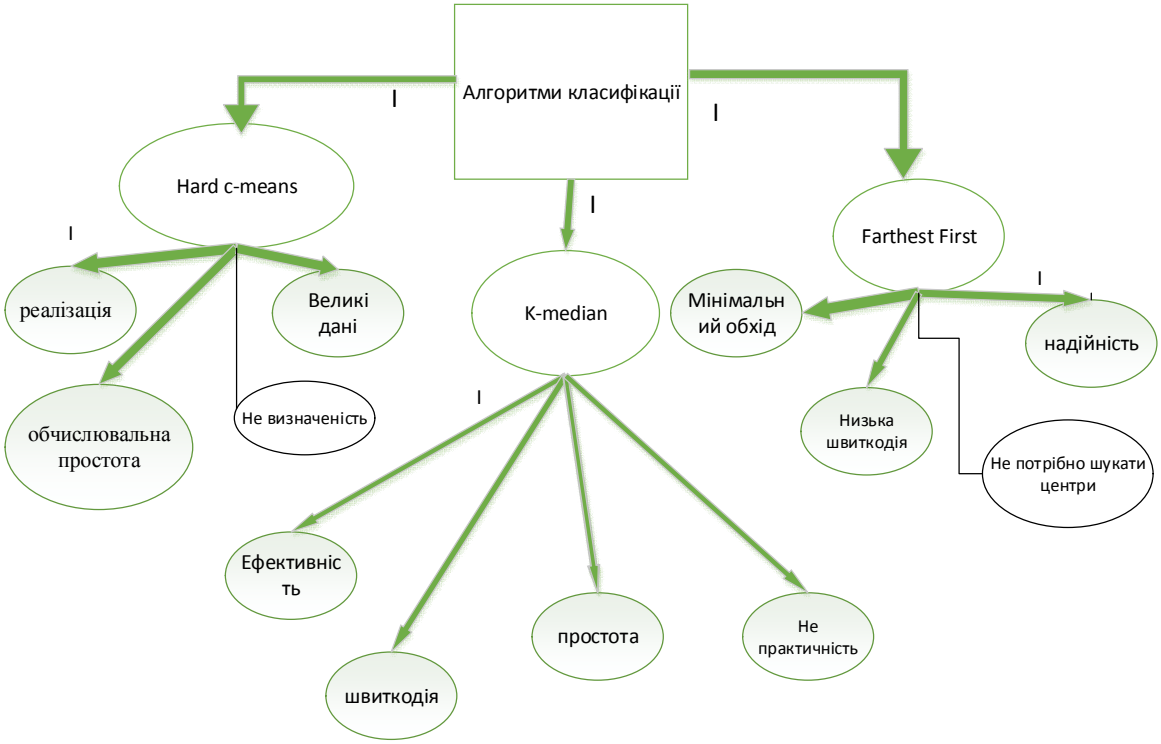

Рисунок 2.6 – Дерево рішень

Програма має чотири узагальнених критерії по яких вона проводить оптимізацію:

–адитивний критерій;

–мультиплікативний критерій;

–максимінний (мінімаксний) критерій.

Адитивний критерій. Цільову функцію будують шляхом додавання нормованих значень власних критеріїв. Власні критерії мають різну фізичну природу і тому різну розмірність. А значить просто підсумовувати їх некоректно. У зв'язку з цим критеріїв ділять на деякі дільники, що нормують і призначається в такий спосіб:

1. Як нормуючі дільники приймаються директивні значення чи параметри критеріїв, які задаються замовником (що можна отримати з технічного завдання на проектований виріб). Вважається, що значення проектних параметрів, закладені в технічному завданні, є оптимальними чи найкращими.

2. Як нормуючі дільники приймаються максимальні (мінімальні) значення критеріїв, що досягаються в області припустимих рішень.

До переваг мультиплікативного критерію можна віднести наступні:

1. Не потрібно нормування власних критеріїв.
2. Практично завжди визначається одне оптимальне рішення.

До недоліків можна віднести такі:

1. Труднощі (суб'єктивізм) у визначенні вагових коефіцієнтів.
2. Перемножування різних розмірностей.
3. Взаємна компенсація значень власних критеріїв.

Максимінний (мінімаксний) критерій. Максимінний (мінімаксний) критерії працюють за принципом компромісу, який ґрунтується на ідеї рівномірності.

Переваги:

1. Широко використовується в процесі автоматизованого проектування.
2. Дає змогу максимізувати запас по працездатності технічного вибору.

3. Шляхом зміни значень вагових коефіцієнтів є можливість досліджувати область слабо ефективних рішень (оптимальних за Слейтером).

Недоліки:

В процесі розв’язання задач багатокритеріальної оптимізації приходиться вирішувати багато однокритеріальних задач, які є складними та нелінійними і інколи зробити це практично неможливо.

Щоб оптимізувати наявне дерево рішень ми виберемо потрібні критерії, що мають ключовий вплив на функціонал підсистеми: моделі мікроконтролерів, активатори, категорії давачів. При оптимізації в програмному забезпечені потрібно водити кількість об’єктів і кількість критеріїв по яких буде робитись вибірка елементів, для оптимізації.

3. ПРОГРАМНА РЕАЛІЗАЦІЯ АЛГОРИТМІВ РОЗПІЗНАВАННЯ БІОМЕДИЧНИХ ЗОБРАЖЕНЬ

3.1. Структура програмного модуля

Маючи значне математичне та алгоритмічне підґрунтя , що в теорії вирішувало проблему пошуку та діагностування ракових захворювань на біомедичних знімках. Це пришвидшило роботу спеціалістів даної області та збільшило точність діагностики на ранніх стадіях. Однак програмний модуль не є останньою інстанцією, а лише допоміжним апаратом. Однак згодом із розвитком штучного інтелекту та розширення бази даних можна розраховувати на робо діагностів та онлайн лікарів. Зарубіжні дослідники активно розвиваються у цій області. Яскравим прикладом є онлайн лікар від технологічного гіганта IBM чи нейронна мережа Google. Тож подібні технології активно розвиваються і впроваджуються в повсякденне життя. Хоча на просторах СНГ це відбувається дещо повільніше. Тому цей програмний додаток покликаний реалізувати алгоритмічні напрацювання вчених минулих і теперішніх поколінь, провести їх переосмислення модернізацію і перетворення, щоб створити структуру програмного модуля(рисунок 3.1).

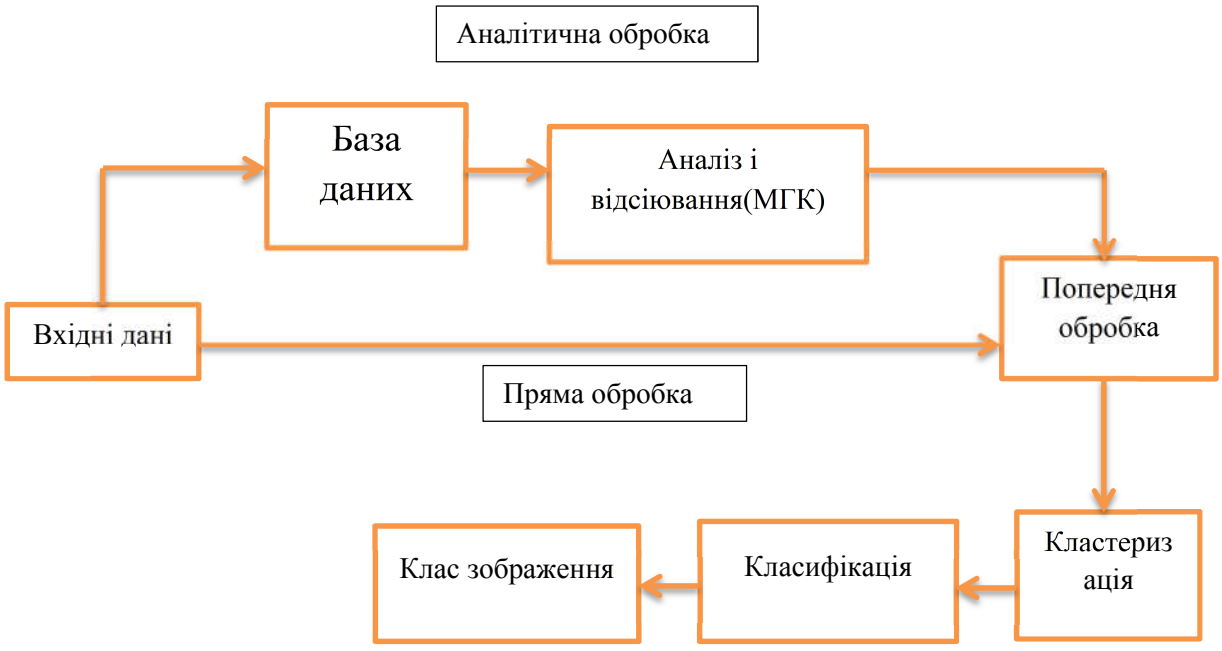

Рисунок 3.1 – Структура програмного модуля класифікації

Розглянемо описану структуру детальніше та розпишемо кожну із структурних частин детальніше. Вхідні дані це зображення отримані після аналізу клітин і тканин пацієнтів. Вони отримані завдяки електронній мікроскопії, рентгену та МРТ. Недоліком цих зображень є низька чіткість та висока зашумленість, тому важко відрізнити важливі елементи(артефакти) від фону. Виправити частину цих недоліків можна через попередню обробку. Це комплекс кроків призначений за допомогою комбінованих фільтрів відсіяти різного роду завади. Більшість зображень піддається впливу різного роду шумів в процесі передачі їх по каналах зв'язку, а також на етапі формування. Тому, першим етапом обробки зображень є фільтрація. Наявність шумів на зображенні може причинити неточності та спотворення на етапі сегментації та розпізнавання. Наприклад, система може сприйняти шуми за окремі об'єкти, що може негативно вплинути на подальші дослідження.

В результаті досліджень можна виділити такі основні види шумів: адитивний Гаусовий та імпульсний. Адитивний Гаусовий шум характеризується додаванням до кожного пікселя зображення значень з відповідного нормального розподілу з нульовим середнім значенням. Такий шум з'являється в пристроях формування цифрових зображень. Імпульсний шум характеризується заміною частини пікселів значеннями фіксованої або випадкової величини. Такий шум пов'язаний з втратами при передачі зображень по каналах зв'язку. Зазвичай, в одному зображенні можна зустріти обидва види шумів, такі шуми називають комбінованими.

Всі види фільтрів можна розділити на такі класи: частотні, лінійні, нелінійні, комбіновані. У класі частотних фільтрів обробці підлягають коефіцієнти розкладання зашумленого сигналу по базису Фур'є або інших базисах, зокрема, вейвлет-базису. Перетворення Фур'є особливо важливе для лінійних фільтрів, оскільки множення в Фур'є-області для частотних методів - це операція згортки для вихідного зображення. У цифровій обробці сигналів широко використовуються методи лінійної фільтрації. Однак, лінійна фільтрація дає хороші результати лише в разі наявності гауссова адитивного

шуму. У випадку імпульсного шуму ефективніше працюють методи нелінійної фільтрації, зокрема медіанна та рангова фільтрація. У випадку комбінованого шуму можна послідовно застосовувати лінійні і нелінійні фільтри або компонувати ці фільтри так, щоб посилити їх сильні сторони і послабити недоліки, це відбувається при використанні гібридних фільтрів. Лінійні фільтри ще називають згладжуючими або усереднюючими, тому що відповідь лінійного фільтра усереднює значення пікселів, що містяться в апертурі, і таким чином згладжує зображення.

Кластеризація займається виділенням значущих частин зображення і як ми згадували в попередніх розділах, для цього процесу буде використаний варіативний K-means. Коли на відфільтрованому зображені будуть виділені значущі частини, вони будуть порівнюватися із еталонами бази знань і на основі цього порівняння відноситися до одного із п’яти класів. Модуль варіативний і в процесі функціонування розширює базу знань та знаходить нові закономірності в отриманих даних. Вершиною розвитку такої системи є створення нових класів на основі здобутих знань та виявлених залежностей, а також розширення перетворення існуючих класів.

Однак при нагромаджені даних їхня цінність різко падає. І потрібно шукати зручні аналітичні важелі для виявлення корисної інформації сортування даних і пошуку по базі. Тому структурна схема містить розгалуження, що передбачило цей поворот подій. При незначній кількості даних, початок роботи системи або швидка діагностика, ми вхідні дані обробляємо і передаємо на кластеризацію і класифікацію. Однак на пізніх етапах це не є рентабельним і потребує звертань до бази даних. І тут для виокремлення потрібних компонентів із бази потрібен метод головних компонент, який відсіє найбільш усереднені значення і залишить лише найбільш виразні ознаки. Це спростить усі наступні етапи роботи модуля. Однак ця функція не є заскріптованою і користувач зможе її використовувати лише за бажанням. Розглянувши структуру модуля перейдемо до програмної частини його побудови та

функціонування, що буде виражено в UML-діаграмах, та основних класах і функціях.

3.2. UML-характеристика модуля

Щоб краще розуміти яким інструментом ми користуємося та які можливості він нам дає зробимо невеличку ремарку та розглянемо UML детальніше. UML (англ. Unified Modeling Language) — уніфікована мова моделювання, використовується у парадигмі об'єктно-орієнтованого програмування. Є невід'ємною частиною уніфікованого процесу розробки програмного забезпечення. UML є мовою широкого профілю, це відкритий стандарт, що використовує графічні позначення для створення абстрактної моделі системи, яка називається UML-моделлю. UML був створений для визначення, візуалізації, проектування й документування в основному програмних систем. UML не є мовою програмування, але в засобах виконання UML-моделей як інтерпретованого коду можлива кодогенерація.

Перша версія (1.0) UML вийшла 13 січня 1997, вона була створена за запитом Object Management Group (OMG) — організації, відповідальної за прийняття стандартів в галузі об'єктних технологій і баз даних. Після обговорення, у вересні 1997 року, версія 1.1 UML була представлена на голосування в OMG. Розробку UML підтримали і вже тоді використовували як стандарт такі гранди ринку інформаційних технологій, як Microsoft, IBM, Hewlett-Packard, Oracle, DEC, Sybase, Logic Works й інші.

Починаючи із середини 60-х років і донедавна , широке поширення мали структурні методології аналізу, проектування і розробки інформаційних систем, що характеризуються штучним поділом (часто неоптимальним) системи на підсистеми, а також слабким взаємозв'язком процесів і даних які присутні в системі. На відміну від них, об'єктні технології, орієнтовані на тісний взаємозв'язок процесів і даних у системах, дозволяють програмним системам

бути надійнішими, легшими для реалізації і стійкішими до змін. Крім того, така філософія моделювання найбільше відповідає загальним концепціям поведінки систем реального світу.

Незважаючи на явну перевагу об'єктно-орієнтованих технологій аналізу і проектування перед структурними, їхнє поширення було незначним, оскільки жоден з методів не давав єдиної і цілісної об'єктної моделі системи. Кожен метод добре висвітлював одну або декілька сторін реальної системи, залишаючи в тіні багато інших, не менш важливих сторін. Крім того, відсутність єдиного стандарту дуже заважало широкому поширенню об’єктно-орієнтованих методів при розробці програмного забезпечення.

Протягом 1994-96 років творці трьох найпоширеніших методологій — Граді Буч (BOOCH), Джим Рамбо (OMT — Object Modeling Technique) і Айвар Якобсон (OOSE — Object Oriented Software Engineering) об'єднали свої зусилля під егідою Rational Software Corporation для створення єдиної мови моделювання, яка б об'єднала всі істотні й успішні розробки в даній галузі і стала би стандартом мови об'єктного моделювання. Грандіозна робота, у якій поряд з Rational брали участь представники багатьох компаній, таких, як Microsoft, IBM, Hewlett-Packard, Oracle, DEC, Unisys, IntelliCorp, Platinum Technology і кількох сотень інших завершилася створенням у січні 1997 року UML 1.0, яка після бурхливого обговорення протягом 1997 року у вересні під версією 1.1 і була передана в OMG для прийняття як галузевий стандарт мови об'єктного моделювання.

UML може бути застосовано на всіх етапах життєвого циклу аналізу бізнес-систем і розробки прикладних програм. Різні види діаграм які підтримуються UML, і найбагатший набір можливостей представлення певних аспектів системи робить UML універсальним засобом опису як програмних, так і ділових систем.[48]

Діаграми дають можливість представити систему (як ділову, так і програмну) у такому вигляді, щоб її можна було легко перевести в програмний код.

Основною причиною використання мови UML є спілкування розробників між собою.[1]

Крім того, UML спеціально створювалася для оптимізації процесу розробки програмних систем, що дозволяє збільшити ефективність їх реалізації у кілька разів і помітно поліпшити якість кінцевого продукту.

UML прекрасно зарекомендувала себе в багатьох успішних програмних проектах. Засоби автоматичної генерації кодів дозволяють перетворювати моделі мовою UML у вихідний код об’єктно-орієнтованих мов програмування, що ще більш прискорює процес розробки.

Практично усі CASE-засоби (програми автоматизації процесу аналізу і проектування) мають підтримку UML. Моделі розроблені в UML, дозволяють значно спростити процес кодування і направити зусилля програмістів безпосередньо на реалізацію системи.

Діаграми підвищують супроводжуваність проекту і полегшують розробку документації.

UML необхідний:

- керівникам проектів, які керують розподілом завдань і контролем за проектом
- проектувальникам інформаційних систем які розробляють технічні завдання для програмістів;
- бізнес-аналітикам, які досліджують реальну систему і здійснюють інжиніринг і реінжиніринг бізнесу компанії;
- програмістам які реалізовують модулі інформаційної системи.

При модифікації системи об'єктний підхід дозволяє легко включати в систему нові об'єкти і виключати застарілі без істотної зміни її життєздатності. Використання побудованої моделі при модифікаціях системи дає можливість усунути небажані наслідки змін, оскільки вони не ламають структури системи, а тільки змінюють поведінку об'єктів.

3.3. Тестові послідовності

Перевірочна вибірка використовується для визначення перенавчання мережі, при якому помилка для навчальної послідовності прагне до нуля, а для перевірочної - зростає. Тестова вибірка застосовується для перевірки якості функціонування навченої мережі. На перевірочної вибірці B2 визначається середня квадратичні помилка A (B2) відхилень значень, обчислених за моделлю.

Ці параметри використовуються для завдання зупинки процесу навчання у випадках, коли помилка для перевірочної вибірки не зменшується або починає зростати. Похибка розрахунку по даному рівнянню досягає 5% як на вихідних даних, так і для перевірочної вибірки.

Похибка розрахунку по даному рівнянню досягає 5%, як на вихідних даних, так і для перевірочної вибірки. Після того як в режимі рекурентного оцінювання на навчальній вибірці BI визначені коефіцієнти моделі і оптимальні значення констант моделі Gb G2 і побудована таким чином модель довела свою придатність на перевірочної вибірці B2 подальший вибір режиму проводиться за значенням неузгодженості y[k l]- Ilpj ([k l]Cj[k]) KO. Якщо K0%, тобто немає великих збурень і змін в технології процесу, то вирішується завдання корекції. В цьому випадку при кожному черговому надходженні значень вхідних впливів за формулами (254) коригується вектор коефіцієнтів, і розраховані прогнозні значення керуючих змінних використовуються для управління процесом.

Опції Test FIS в правому нижньому кутку вікна дозволяють провести перевірку і тестування створеної і навченої системи з висновком результатів у вигляді графіків (відповідні графіки для навчальної вибірки - Training data, яка тестує вибірки - Testing data і перевірочної вибірки - Checking data. Кнопка Test Now дозволяє запустити зазначені процеси.

Особливості методу дозволяють перевірити прогнозують властивості моделі. Для цього сукупність вихідних даних ділиться на дві рівні за обсягом

вибірки. Перша вибірка умовно називається навчальною, друга - перевірочної. Пошук структури і коефіцієнтів моделі здійснюється за наступною схемою. на першому етапі розрахунків модель будується на навчальній вибірці даних. За отриманою моделі визначають розрахункові величини вихідної змінної (нафтовіддачі) для всієї сукупності точок. на другому етапі розрахунків вибірки міняються місцями, тобто колишня перевірочна вибірка стає навчальною, а колишня навчальна - перевірочної, і розрахунки повторюються . Якщо отримані таким чином моделі відрізняються незначно і характеризуються високою дисперсійною мірою ідентичності, то очевидно [49].

ВИСНОВКИ

В результаті виконання магістерського проекту було досліджено:

1) Досліджено біомедичні зображення їх різновиди та особливості обробки. Проаналізовано різні методи класифікації цих зображень та пройдено всі етапи розпізнавання зображень.

2) Досліджено базовий алгоритм кластеризації K-means.

3) Досліджено модифікацію даного алгоритми для роботи із обширним обсягом даних(Hard C-means).

4) Досліджено алгоритм усунення зовнішніх завад при визначені центрів кластера(K-median).

5) Досліджено алгоритм альтернативного пошуку центрів (Farthest-first).

6) Проаналізовано алгоритми знайдено їх переваги та недоліки та вибрано оптимальний варіант для реалізації поставленого завдання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Березький О.М. Синтез альтернативних рішень при структурному проектуванні систем автоматизованої мікроскопії / О.М. Березький, К.М. Березька, Ю.М. Батько, Г.М. Мельник // Науковий вісник Українського державного лісотехнічного університету. – 2009. – Т. 19 (5), № 23. – С. 258–268.

2. Ovalle A. KIDS: A Distributed Expert System for Biomedical Image Interpretation. / Arturo Ovalle, Catherine Garbay // Information Processing in Medical Imaging – 1991. – С. 419– 433.

3. Обработка изображений: технология, методы, применение: [учебное пособие] /С. В. Абламейко, Д. М. Лагуновский. – Минск: Амалфея, 2000. – 304 с.

4. Інформаційні технології та моделювання бізнес-процесів: [навч. посіб.] / О. М. Томашевський, Г.Г. Цегелик, М. Б. Вітер, В. І. Дудук. – К.: "Видавництво "Центр учбової літератури", 2012. – 296 с.

5. Berezsky O. Biomedical Image Search and Retrieval Algorithms / O. Berezsky, G. Melnyk, Yu. Batko// Computing – 2008. – № 7. – С. 108– 114.

6. Березький О.М. Текстурна сегментація біомедичних зображень на основі просторових моментів /О.М. Березький, Г. М. Мельник, Ю. М. Батько // Матеріали 4-ї Міжнародної "Комп’ютерні науки та інформаційні технології 2009" науково-технічної конференції. – Львів, 2009. – С. 42– 45.

7. Мельник Г.М. Метод і алгоритми аналізу симетричних зображень / Г.М. Мельник // Штучний інтелект. – 2010. – № 4. – С. 253– 261.

8. Березький О. М. Порівняння алгоритмів синтезу біомедичних зображень / О. М. Березький, Г. М. Мельник // Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту: Матеріали міжнародної наукової конференції. – Євпаторія, 2011. – Т. 2. – С. 189– 193.

9. Миронов Д. Ф. Компьютерная графика в дизайне / Миронов Д. Ф. – СПб.: БХВ-Петербург, 2008.– 560 с.

10. Мельник Г.М. Метод знаходження відповідних точок на контурах мікрооб’єктів біомедичної природи // Вісник Національного університету "Львівська політехніка" "Комп’ютерні науки та інформаційні технології". – 2012. – № 720. – С. 275– 283.

11. Дацко Т.В. Морфометричні особливості епітелію шийки матки при цитологічному дослідженні дисплазій / Т. В. Дацко, О.М. Березький, Ю.М. Батько та ін // Науково-практичний журнал "Здобутки клінічної і експериментальної медицини" – 2008. – № 2 (9). – С. 112.

12. Дэвид А. Computer Vision: A Modern Approach /А. Дэвид, М. Форсайт, Д. Понс // М. : «Вильямс», 2004. — 928 с.

13. Стокман Д, Computer Vision, Компьютерное зрение / Л. Шапиро, Д. Стокман // М. : Бином. Лаборатория знаний, 2006. — 752 с.

14. Горелик А. Л. Методы распознавания / А. Л. Горелик, В. А. Скрипкин // М.: Высшая школа, 1989.

15. Чэн Ш.-К. Принципы проектирования систем визуальной информации: М.: Мир, 1994.

16. Вапник В. Н. Теория распознавания образов / В. Н. Вапник А. Я. Червоненкис // М.: Наука, 1974. — 416 с.

17. Ackerknecht E. H. Doctor, statesman, anthropologist / E. H. Ackerknecht, Rudolf Virchow // Anthropologist, 1953. – 10c.

18. Beck A. H. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival / A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O.Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, D. Koller. // Science translational medicine, - 2011. - №5. - P.108–113

19. Cires D. C. Mitosis detection in breast cancer histology images with deep neural networks. In Medical Image Computing and Computer-Assisted Intervention / D. C. Cires, A. Giusti, L. M. Gambardella, J. Schmidhuber // MICCAI. Springer – 2013 - №5. – P.411–418.

20. Cotran R. S. Robbins pathologic basis of disease / R. S. Cotran, V. Kumar, T. Collins, and S. L. Robbins // - 1999 – P.13.

21. Czerniecki B. J. Immunohistochemistry with pancytokeratins improves the sensitivity of sentinel lymph node biopsy in patients with breast carcinoma / B. J. Czerniecki, A. M. Scheff, L. S. Callans, F. R. Spitz, I. Bedrosian, E. F. Conant, S. G. Orel, J. Berlin, C. Helsabeck, D. L. Fraker, et al. // Cancer. – 1999 - № 85. – P.1098–1103.

22. Edge S. B. The American joint committee on cancer / S. B. Edge C. C. Compton // the 7th edition of the ajcc cancer staging manual and the future of tnm. Annals of surgical oncology. – 2010 – Vol.17, №6. – P.1471–1474.

23. Elmore J. G. Diagnostic concordance among pathologists interpreting breast biopsy specimens / J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, et al. // Jama. - 2015 - P.1122–1132 .

24. Ghaznavi F. Digital imaging in pathology / F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman // whole-slide imaging and beyond. Annual Review of Pathology: Mechanisms of Disease. - 2013 - №8. - P.331–359.

25. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: a review. Biomedical Engineering, IEEE Reviews in, 2:147–171, 2009. 1, 5

26. Irshad H. Methods for nuclei detection, segmentation, and classification in digital histopathology / H. Irshad, A. Veillard, L. Roux, and D. Racoceanu // A review current status and future potential. Biomedical Engineering. IEEE. – 2014 - №5 – P.114.

26. Jaffer S. Evolution of sentinel lymph node biopsy in breast cancer, in and out of vogue? / S. Jaffer and I. J. Bleiweiss // Advances in anatomic pathology. – 2014 – Vol.21 - №6 – P.433–442.

27. Krizhevsky A. Imagenet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, and G. E. Hinton, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger // Advances in Neural Information Processing – 2012 - Systems 25 – P.1097–1105.

28. Lyman G. H. American society of clinical oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer / G. H. Lyman, A. E. Giuliano, M. R. Somerfield, A. B. Benson, D. C. Bodurka, H. J. Burstein, A. J. Cochran, H. S. Cody, S. B. Edge, S. Galper, et al. // Journal of Clinical Oncology. – 2005 – Vol.23 - №30 – P.7703–7720.

29. Lyman G. H. Sentinel lymph node biopsy for patients with early-stage breast cancer / G. H. Lyman, S. Temin, S. B. Edge, L. A. Newman, R. R. Turner, D. L. Weaver, A. B. Benson, L. D. Bosserman, H. J. Burstein, H. Cody, et al. // American society of clinical oncology clinical practice guideline update. Journal of Clinical Oncology. - 2014 – Vol.32 - №13 – P.1365–1383.

30. Nakhleh R. E. Error reduction in surgical pathology: Archives of pathology & laboratory medicine. - 2006 - P.630–632.

31. Otsu N. A Threshold Selection Method from Gray-level Histograms: IEEE Transactions on Systems. Man and Cybernetics. – 1979 - P.62–66.

32. Raab S. S. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses / S. S. Raab, D. M. Grzybicki, J. E. Janosky, R. J. Zarbo, F. A. Meier, C. Jensen, and S. J. Geyer // Cancer. - 2005 –Vol.104 - №10 – P.2205–2213.

33. Russakovsky O. Imagenet large scale visual recognition challenge / O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. // International Journal of Computer Vision. - 2015 - P.211–252.

34. Simonyan K. Very deep convolutional networks for large-scale image recognition / K. Simonyan and A. Zisserman // CoRR . Abs. - 2014 - P.1409-1556.

35. Szegedy C. Going deeper with convolutions / C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich // CVPR – 2015 – P.203.

36. Wang D. Face Search at Scale / D. Wang, C. Otto, and A. K. Jain // 80 Million Gallery. - 2015.

37. Weaver D. L. Comparison of pathologist detected and automated computer-assisted image analysis detected sentinel lymph node micro metastases in

breast cancer / D. L. Weaver, D. N. Krag, E. A. Manna, T. Ashikaga, S. P. Harlow, and K. D. Bauer // Modern pathology. - 2003 – Vol.16 - №11 - P.1159–1163.

38. Сайт general |Електронний ресурс| - режим доступу до ресурсу http://www.generalpicturerecognition.com/

39. Steinhaus H. Sur la division des corps materiels en parties: Bull. Acad. Polon. - 1956 Vol. IV – P.801 - 804.

40. Lloyd S. Least square quantization in PCM's: Bell Telephone Laboratories Paper - 1957.

41. MacQueen J. Some methods for classification and analysis of multivariate observations: In Proc. 5th Berkeley Symp. on Math. Statistics and Probability. - 1967 P. 281—297.

42. Flury B. Principal points: Biometrika. -1990 - № 77 – P. 33-41.

43. Gorban A.N. Principal Graphs and Manifolds, Ch. 2 in: Handbook of Research on Machine Learning Applications and Trends / Gorban A.N., Zinovyev A.Y Algorithms // Methods and Techniques: IGI Global. Hershey. PA. USA. – 2009 - pp. 28-59.

44. Arthur D. "How Slow is the k-means Method?" / David Arthur, Sergei Vassilvitskii // Proceedings of the 2006 Symposium on Computational Geometry (SoCG).

45. E.M. Mirkes K-means and K-medoids applet: University of Leicester. - 2011.

46. Coates A. Learning Feature Representations with K-means / Adam Coates, Andrew Y. Ng. // Stanford University. - 2012.

47. Фаулер М. UML. Основы / Фаулер М., Скотт К. / Пер. с англ. — СПб: Символ-Плюс, 2002. — 192 с.

48. Рост Р.Д. OpenGL. Трехмерная графика и язык программирования шейдеров / Р.Д. Рост/ Пер. с англ. – СПб.: Питер, 2005. – 428 с.

49. Kuncheva L. Combining pattern classifiers: methods and algorithms / L. Kuncheva. – John Wiley & Sons, 2004.