

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ БІОМЕДИЧНОЇ ІНЖЕНЕРІЇ

(повна назва інституту/факультету)

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

(повна назва кафедри)

«На правах рукопису»
УДК 616.7

«До захисту допущено»:
Завідувач кафедри БМК

(підпис) Євген НАСТЕНКО
(ім'я, ПРІЗВИЩЕ)

«___» _____ грудня _____ 2021р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Комп'ютерні технології в біології та
медицині»
зі спеціальності 122 «Комп'ютерні науки»

на тему: Дослідження ефективності критеріїв відбору у алгоритмі первин-
ної селекції ознак в задачі класифікації патології печінки

Виконала: студентка II курсу, групи БС-01мп

КОЖАРА КАТЕРИНА МИКОЛАЇВНА

(прізвище, ім'я, по батькові)

(підпис)

Науковий керівник:

доц. каф. БМК, к.т.н., доц. Павлов Володимир Анатолійович

(посада, науковий ступінь, вчене звання, прізвище, ім'я, по-батькові)

(підпис)

Консультант з розділів магістерської дисертації:

(посада, науковий ступінь, вчене звання, прізвище, ім'я, по-батькові)

(підпис)

Рецензент:

(посада, науковий ступінь, вчене звання, прізвище, ім'я, по-батькові)

(підпис)

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних посилань.

Студентка _____
(підпис)

Київ – 2021 року

Нормо контролер

Галина КОРНІЄНКО

Кожара, К. М. Дослідження ефективності критеріїв відбору у алгоритмі первинної селекції ознак в задачі класифікації патології печінки: магістерська дис. : 122 Комп'ютерні науки / Кожара Катерина Миколаївна. – Київ, 2021. – 95с.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет біомедичної інженерії
Кафедра біомедичної кібернетики

Рівень вищої освіти – другий (магістерський)

Спеціальність – 122 «Комп'ютерні науки»

Освітньо-професійна програма «Комп'ютерні технології в біології та медицині»

ЗАТВЕРДЖУЮ

Завідувач кафедри БМК

(підпис) Євген НАСТЕНКО
(ім'я, ПРІЗВИЩЕ)

«31» серпня 2021р.

ЗАВДАННЯ
на магістерську дисертацію студентці

Кожарі Катерині Миколаївні

(прізвище, ім'я, по батькові)

1. Тема дисертації **Дослідження ефективності критеріїв відбору у алгоритмі Первинної селекції ознак в задачі класифікації патології печінки**
науковий керівник дисертації

доц.,каф БМК Павлов В.А.

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «09» листопада 2021 р. №3703-с

2. Термін подання студентом дисертації: **01-04 грудня 2021 року**

3. Об'єкт дослідження: **ефективність критеріїв відбору.**

4. Вихідні дані: **результати дипломної роботи бакалавра; зображення УЗД печінки,; результати колег по цеху з кафедри біомедичної кібернетики.**

5. Перелік завдань, які потрібно розробити: **1. Проаналізувати джерела (вітчизняні та/або іноземні), які стосуються проблематики. 2. Описати задачу наукового дослідження. 3. Описати критерії для знаходження оптимального варіанту селекції ознак. 4. Використання алгоритму первинної селекції ознак 5. Розробити концептуальні висновки за результатами магістерської дисертації.**

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: **35 рисунків, 8 таблиць, презентація на 25 слайдів.**

7. Орієнтовний перелік публікацій: **1 публікація у фахових виданнях, 1 конференція.**

8. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

9. Дата видачі завдання: 31 серпня 2021 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Отримання завдання на МД	31 серпня 2021р.	
2	Практика	з 01.09.2021р. по 26.10.2021р.	
3	Вставлення результатів практики в розділи МД	4 листопада 2021р.	
4	Аналіз предметної області	4 листопада 2021р.	
5	Методи селекції ознак алгоритми класифікації	5-11 листопада 2021р.	
6	Алгоритми класифікації	12-18 листопада 2021р.	
7	Дослідження ефективності критеріїв відбору	19-25 листопада 2021р.	
8	Стартап аналіз проекту магістерської дисертації	26-29 листопада 2021р.	
9	Складання висновків до результатів виконання магістерської дисертації	30 листопада 2021р.	
10	Подання в електронному вигляді МД та анотації до неї на перевірку нормоконтролера та UNICHECK .	1-3 грудня 2021р.	
11	Передзахист МД та допуск до захисту дисертації	4-6 грудня 2021р.	
12	Подання МД рецензенту. Отримання рецензії.	7-15 грудня 2021р.	
13	Подання пакету документів по МД до захисту в ЕК ¹	15-17 грудня 2021р.	
14	Захист МД в ЕК	20-22 грудня 2021р.	

Студент

_____ (підпис)

Катерина КОЖАРА

_____ (ім'я, ПРІЗВИЩЕ)

Науковий керівник

_____ (підпис)

Володимир ПАВЛОВ

_____ (ім'я, ПРІЗВИЩЕ)

* Якщо визначені консультанти. Консультантом не може бути зазначено наукового керівника магістерської дисертації.

¹ не пізніше ніж за один тиждень до затвердженої дати захисту МД в ЕК

РЕФЕРАТ

Магістерська дисертація за темою «Дослідження ефективності критеріїв відбору у алгоритмі первинної селекції ознак в задачі класифікації патології печінки» виконана студентом кафедри біомедичної кібернетики ФСП Кожарою Катериною Миколаївною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині» та складається зі: вступу; 4 розділів (аналіз предметної області, методи реконструкції зображення, постановка задачі, аналіз ефективності відбору), розділу зі стартап проєкту, висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 30 джерела. Загальний обсяг роботи 95 сторінок.

Обсяг роботи: 95 сторінок, 35 ілюстрацій, 30 джерел посилань.

Актуальність теми. Діагностика захворювань печінки на ранніх стадіях допоможе більш якісно оцінити стан пацієнта та обрати якомога кращу лікувальну стратегію..

Мета дослідження. Знаходження оптимального варіанту селекції ознак, для ефективного виконання задачі бінарної класифікації «норма - патологія» при дифузних захворюваннях печінки.

Об'єкт дослідження. Зображення УЗД печінки.

Предмет дослідження. Ефективність критеріїв відбору у алгоритмі первинної селекції ознак в задачі класифікації патології печінки.

Методи дослідження. Методи селекції за критеріями внутрішньокласової дисперсії, міжкласової дисперсії, відношення внутрішньокласової і міжкласової дисперсій, кореляційний відбір ознак.

Інструменти дослідження. Python, Anaconda, Jupyter Notebook.

Публікації. За результатами виконаної роботи було опубліковано наукову статтю:

1. Кожара К.М., Корнієнко Г.А., Дослідження ефективності критеріїв селекції ознак в задачі класифікації норма-патологія печінки. //

Мультидисциплінарний Міжнародний науковий журнал "Інтернаука":.-Київ.-
грудень 2021. -6с.

Конференції. Доповідь на конференції:

Кожара К. Ефективність критеріїв відбору. Наука, освіта, інновації та
технології: стан, проблеми та перспективи розвитку 2021, листопад 2021.-2с.

Ключові слова: алгоритми класифікації, методи селекції, ознаки,
генетичний алгоритм, ультразвукова діагностика, Python.

Abstract

Master's dissertation on " Analysis of enrollment criteria efficiency in algorithm of the primary selection of indicators in the problematics of liver pathology classification" performed by *Kozhara Kateryna*, a student of the Department of Biomedical Cybernetics FBMI by specialty 122 "Computer Science" in the educational and professional program "Computer Technology in Biology and Medicine" and consists of an introduction, 4 chapters (subject area analysis, image reconstruction methods, problem statement, selection efficiency analysis), conclusions to each chapter, general conclusions and list of references that includes 30 points. The paper amounts to 95 pages.

Paper size: 95 pages, 35 illustrations, 30 references.

Relevance of the topic. Diagnosis of liver disease in the early stages will help to better assess the patient's condition and choose the best possible treatment strategy.

Objective of the study. Знаходження оптимального варіанту селекції ознак, щоб можна було ефективно виконати задачу бінарної класифікації «Норма:Патологія».

Object of study. Ultrasound image of the liver..

Subject of study. The effectiveness of selection criteria in the algorithm of primary selection of traits in the problem of classification of liver pathology.

Research methods. Intraclass variance, interclass variance, the ratio of intraclass and interclass variance, correlation selection of features..

Research tools. Python, Anaconda, Jupyter Notebook.

Publications. As a result of the work done 3 scientific articles were published:

Kozhara K., Kornienko H. Research of efficiency of criteria of selection of signs in a problem of classification norm-pathology of a liver. Kozhara K/// International scientific journal "Internauka".:-Kiev.- december 2021.-6p.

Conferences. One conferences were attended:

Kozhara K. Effectiveness of criterion criteria. Science, education, innovation and technology: state, problems and prospects of development 2021, November 2021.- 2p.

Key words: classification algorithms, selection methods, traits, genetic algorithm, ultrasound diagnostics, Python.

Зміст

ВСТУП.....	11
РОЗДІЛ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ	14
1.1 УЛЬТРАЗВУКОВЕ ДОСЛІДЖЕННЯ ПЕЧІНКИ.....	14
1.2 ЗАХВОРЮВАННЯ ПЕЧІНКИ	17
1.3 ДІАГНОСТИКА ЗАХВОРЮВАНЬ ПЕЧІНКИ.....	20
1.4 ТЕКСТУРНІ ОЗНАКИ	23
1.5 Методи візуалізації.....	20
Висновки до розділу 1	27
РОЗДІЛ 2 МЕТОДИ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ	29
2.1 K-найближчих сусідів (KNN).....	30
2.2 МЕТОД ОПОРНИХ ВЕКТОРІВ (SUPPORT VECTOR MACHINE – SVM) з лінійною ФУНКЦІЄЮ ЯДРА	34
2.3 SVM з ФУНКЦІЄЮ РАДІАЛЬНОГО БАЗИСУ (RADIAL BASIS FUNCTION– RBF)	35
2.4 ГАУСІВСЬКИЙ ПРОЦЕС.....	37
2.5 ДЕРЕВО ПРИЙНЯТТЯ РІШЕНЬ (ДПР).....	40
2.6 Випадковий ліс.....	40
2.7 Багатошаровий перцептрон(БШП),тобто нейтронна мережа.....	41
2.8 AdaBoost.....	43
2.9 Наївний Баєсівський класифікатор.....	44
2.10 Квадратичний дискримінантний аналіз.....	47
2.11 Методи селекції ознак.....	49
2.11.1 Прямий відбір.....	49
2.11.2 Зворотне усунення.....	51
2.11.3 Рекурсивне видалення ознак.....	52
2.11.4 Регуляризація.....	54
Висновки до розділу 2.....	54

РОЗДІЛ 3 ПОСТАНОВКА ЗАДАЧІ	60
3.1 Категорії алгоритмів відбору.....	57
3.2 Постановка задачі.....	61
3.2 Метод семплінгу.....	62
Висновки до розділу 3.....	62
РОЗДІЛ 4 АНАЛІЗ ЕФЕКТИВНОСТІ ВІДБОРУ	70
Висновки до розділу 4	88
РОЗДІЛ 5 СТАРТАП АНАЛІЗ ПРОЕКТУ	ОШИБКА! ЗАКЛАДКА НЕ ОПРЕДЕЛЕНА.
Висновки до розділу 5	ОШИБКА! ЗАКЛАДКА НЕ ОПРЕДЕЛЕНА.
ЗАГАЛЬНІ ВИСНОВКИ.....	93
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	94

ВСТУП

Актуальність теми магістерської дисертації. Захворювання печінки становлять приблизно 2 мільйони смертей на рік у всьому світі, 1 мільйон через ускладнення цирозу і 1 мільйон через вірусний гепатит та гепатоцелюлярну карциному. В даний час цироз є 11-ю найпоширенішою причиною смерті в усьому світі, а рак печінки 16-ю провідною причиною смерті; разом, вони становлять 3,5% усіх смертей у світі. Цироз входить до 20 основних причин втрати працездатності років життя та втрачених років життя, що становить 1,6% та 2,1% загального тягара у всьому світі. Близько 2 мільярдів людей у всьому світі вживають алкоголь, і понад 75 мільйонів мають діагностику розладів, пов'язаних із вживанням алкоголю, і піддаються ризику захворювання печінки, пов'язаного з вживанням алкоголю. Приблизно 2 мільярди дорослих страждають на ожиріння або надмірну вагу, а понад 400 мільйонів – на цукровий діабет; обидва з них є факторами ризику неалкогольної жирової хвороби печінки та гепатоцелюлярної карциноми. Світова поширеність вірусного гепатиту залишається високою, в той час як медикаментозне ураження печінки продовжує зростати як основна причина гострого гепатиту. Трансплантація печінки є другою за поширеністю трансплантацією твердих органів, але за нинішніми темпами задовольняється менше 10% глобальних потреб у трансплантації. Хоча ці цифри є тверезими, вони підкреслюють важливу можливість покращити здоров'я населення, оскільки більшість причин захворювань печінки можна запобігти.

Діагностика захворювань печінки на ранніх стадіях допоможе більш якісно оцінити стан пацієнта та обрати якомога кращу лікувальну стратегію[30].

Мета і завдання дослідження.

Основною задачею даної праці є виконання оптимальної селекції ознак для успішної реалізації загальної задачі класифікації патології печінки за знімками ультразвукового дослідження (УЗД). Подібна задача виникла під

час роботи кафедри біомедичної кібернетики (БМК) КПІ над проектом, основна мета якого є розробка системи підтримки прийняття рішень під час діагностики печінки, замовником якої виступає ДУ «Інститут ядерної медицини та променевої діагностики НАМН України».

Актуальність роботи полягає у дослідженні визначення найефективнішого критерія відбору у алгоритмі первинної селекції ознак в задачі класифікації патології печінки.

Мета роботи: знаходження оптимального варіанту селекції ознак, щоб можна було ефективно виконати задачу бінарної класифікації «Норма:Патологія». Для її досягнення необхідне виконання переліку наступних задач:

Для досягнення мети були поставлені наступні задачі:

1. Проаналізувати джерела (вітчизняні та/або іноземні).
2. Опис загальної вибірки УЗД зображень.
3. Для знаходження оптимального варіанту селекції ознак запропонувати 6 критеріїв.
4. Створення алгоритму первинної селекції ознак.
5. Розробити концептуальні висновки за результатами роботи.

Наукова новизна одержаних результатів

Основною задачею даної праці є виконання оптимальної селекції ознак для успішної реалізації загальної задачі класифікації патології печінки за знімками ультразвукового дослідження (УЗД).

Практичне значення одержаних результатів

Подібна задача виникла під час роботи кафедри біомедичної кібернетики (БМК) КПІ над проектом, основна мета якого є розробка системи підтримки прийняття рішень під час діагностики печінки, замовником якої виступає ДУ «Інститут ядерної медицини та променевої діагностики НАМН України».

Особистий внесок здобувача

Створення алгоритму первинної селекції ознак.

Апробація результатів магістерської дисертації.

За результатами виконаної роботи на практиці були опубліковані наступні наукові статті:

Кожара К.М., Корнієнко Г.А., Дослідження ефективності критеріїв селекції ознак в задачі класифікації норма-патологія печінки. // Мультидисциплінарний Міжнародний науковий журнал "Інтернаука":.-Київ.- грудень 2021. -6с.

Та відвідано одну конференцію:

Кожара К. Ефективність критеріїв відбору. Наука, освіта, інновації та технології: стан, проблеми та перспективи розвитку 2021, листопад 2021.

Структура дисертації

Магістерська дисертація за темою «Дослідження ефективності критеріїв відбору у алгоритмі первинної селекції ознак в задачі класифікації патології печінки» виконана студентом Кожарою Катериною Миколаївною зі спеціальності 122 «Комп'ютерні науки» за освітньо-професійною програмою «Комп'ютерні технології в біології та медицині» побудована за класичним типом та викладена на 125 сторінках машинописного тексту. Вона складається : вступу; 4 розділів (аналіз предметної області, методи реконструкції зображення, постановка задачі, аналіз ефективності відбору), розділу зі стартап проекту, висновків до кожного з цих розділів; загальних висновків; списку використаних джерел, який налічує 30 джерел (13 – на кирилиці, 17 – на латиниці). В дисертації представлено 35 рисунків і 2 таблиць.

РОЗДІЛ 1

АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Ультразвукове дослідження печінки

Ультразвукове дослідження (УЗД) є простим і неінвазивним діагностичним інструментом в області черевної порожнини. Доплерівський режим пропонує оцінку гемодинаміки в реальному часі, а УЗД з контрастуванням є одним з найбільш часто використовуваних методів для детальної оцінки. Подальший розвиток цифрових технологій дозволяє здійснювати тривимірну (3D) візуалізацію цільових зображень з високою роздільною здатністю. У цій статті розглядається широкий спектр застосування при УЗД черевної порожнини та описується останній прогрес у діагностиці хронічних захворювань печінки.

Завдяки різному спектру етіології хронічне захворювання печінки зростає в усьому світі [1]. Цироз є найбільш запущеною стадією хронічного захворювання печінки. Це пов'язано з можливими побічними явищами, такими як варикозне розширення вен шлунка, асцит, печінкова енцефалопатія та розвиток гепатоцелюлярної карциноми (ГЦК), які потребують ретельної медичної допомоги.

Через переваги простих і менш інвазивних оцінок ультразвукове дослідження (УЗД) може бути найбільш часто використовуваним інструментом візуалізації в практичному веденні пацієнтів із хронічними захворюваннями печінки. На додаток до візуалізації в В-режимі, останній розвиток цифрових технологій представив різні режими, режим кольорового/силового доплера, гармонійне зображення як контрастний режим на основі мікропухирців і тривимірний (3D) режим. З огляду на це, ця оглядова стаття описує останні досягнення США в діагностиці хронічних захворювань печінки.

Останні цифрові розробки показали покращення просторової роздільної здатності та співвідношення сигнал/шум у В-режимі, який є єдиним зображенням, яке демонструє основні зображення тканин. Роль простої методики при хронічному захворюванні печінки полягає в ідентифікації цирозу, вимірюванні діаметра (тобто печінки, селезінки та судин) і виявлення асцити, колатеральних судин і осередкових уражень печінки [8-10].

М'язи є ще однією мішенню УЗ в В-режимі. Кількісна оцінка м'язової маси є важливим питанням догляду за пацієнтами, оскільки вона тісно пов'язана з недостатнім харчуванням, що призводить до зниження якості життя, пов'язаної зі здоров'ям [11,12]. Саркопенія визначається значним зниженням м'язової маси та/або сили та широко прийнята як поширене порушення при цирозі печінки [13,14]. Це є суттєвим фактором для поганого прогнозу, для прогнозування результатів у пацієнтів з ГЦК та для розвитку ускладнень після операції на печінці або трансплантації печінки [15].

1. Перевірка якості зображення

Ультразвукове дослідження печінки зазвичай містить серію кількох зображень, зроблених технологом, який проводить обстеження. Оскільки цей тип перевірки зображень виконується від руки, він дуже залежить від користувача. Області зображення та якість знімків значно залежать від видимості та техніки. Порушені зображення можуть негативно вплинути на те, що може побачити рентгенолог. Друга думка може допомогти підтвердити, що зображення достатньо високої якості, щоб поставити точний діагноз.

2. Діагностика та характеристика захворювань печінки Ультразвукове дослідження корисно для відрізнення кіст печінки від твердих утворень. Деякі утворення мають унікальні характеристики ультразвуку, тоді як інші мають подібні характеристики ультразвуку. Ультразвукове дослідження печінки також використовується для оцінки дифузних захворювань печінки, включаючи жирову дистрофію печінки. Рентгенолог, який надає другу

думку, може допомогти охарактеризувати важке ураження печінки, оцінити печінку на наявність жиру та визначити інші тести, які можуть бути необхідні для оцінки печінки.

3. Визначення того, чи необхідна подальша візуалізація. УЗД може бути першим скринінговим тестом, який проводиться під час оцінки печінки. Залежно від результатів лікар може порекомендувати ще одне обстеження. Наприклад, КТ, МРТ або біопсія можуть бути показані для диференціації між фіброзом печінки та жирової тканини, додаткової кількісної оцінки тяжкості жиру або оцінки інших захворювань печінки. КТ або МРТ також можуть бути рекомендовані для подальшої оцінки певних уражень печінки. Радіолог, який надасть другу думку, може допомогти вам щодо додаткового тестування.

Ультразвукове дослідження дозволяє виявити багато ракових (злоякісних) і неракових (доброякісних) утворень печінки. Утворення, як правило, легше побачити за допомогою ультразвуку, чим більше воно відрізняється за зовнішнім виглядом і, отже, виділяється з фонові тканини. Хоча це не жорстке правило, у багатьох випадках більшу масу легше виявити, ніж меншу масу.

Але виявлення будь-яких утворень можливе лише за наявності хорошого ультразвукового «вікна». На жаль, кістки, кальциноз і газ діють як бар'єри для ультразвукового променя печінки. Жирова тканина перед або всередині печінки також зменшує проникнення променя і видимість мас.

Інша річ, яку слід визнати, це те, що існує деяке збігання у проявах ракових і неракових утворень. На щастя, ультразвукове дослідження з контрастним посиленням (CEUS) порівняно недавно з'явилося в Сполучених Штатах як корисний метод для подальшої оцінки уражень печінки, які погано характеризуються звичайним ультразвуком. Він використовується в Європі та Азії протягом кількох років.

За допомогою CEUS контрастну речовину з мікропухирцями вводять у вену під час ультразвукового дослідження печінки, щоб краще візуалізувати кровотік. Аналіз структур потоку корисний для диференціації певних уражень один від одного, а також для спостереження за певними ураженнями.

Ультразвукове дослідження чудово показує нормальну анатомію та наявність відхилень у печінці. Він особливо відмінно підходить для диференціації кіст від твердих мас. Прості кісти мають тонку стінку і містять рідину, яка виявляється темнішим центром, ніж тверді маси на ультразвуку. Складні кісти можуть мати супутні утворення (вузлики), кальцифікати або множинні тканинні смуги. Тверді утворення можна навіть оцінити на предмет кровотоку за допомогою методики, яка називається ультразвуком доплера, і кістозні ділянки всередині утворень можна відрізнити від твердих частин.

Ультразвукове дослідження також може оцінити дифузні захворювання печінки, такі як жирова дистрофія печінки, гепатит та цироз. Наприклад, жирова дистрофія печінки (стеатоз) зазвичай яскравіша (більш «ехогенна» або «гіперехогенна») на УЗД печінки, ніж звичайна печінка, тоді як гепатит може бути менш яскравим («гіпоехогенний»). Печінка з цирозом часто виглядає зморщеною і грудкою. До ультразвукового дослідження можна додати спеціальний метод, який називається еластографія, щоб допомогти виміряти еластичність печінки та оцінити тяжкість фіброзу.

Розширені жовчні протоки та будь-яка рідина поблизу печінки (асцит, скупчення рідини) також зазвичай виявляються на УЗД печінки. Також часто спостерігаються інші органи, включаючи жовчний міхур, праву нирку і принаймні частину підшлункової залози.

1.2 Захворювання печінки

Гепатит А - це вірусна інфекція, яка викликає запалення та ураження

печінки. Запалення - це набряк, який виникає, коли тканини тіла травмуються або інфікуються. Запалення може пошкодити органи.

Віруси зовнішнього зв'язку НІН вторгаються в нормальні клітини вашого тіла. Багато вірусів викликають інфекції, які можуть передаватися від людини до людини. Вірус гепатиту А зазвичай поширюється при контакті з їжею або водою, зараженими випорожненнями інфікованої людини.

Гепатит А є гострою або короткочасною інфекцією, що означає, що люди зазвичай одужують без лікування через кілька тижнів. У рідкісних випадках гепатит А може бути важким і призвести до печінкової недостатності та необхідності екстреної пересадки печінки, щоб вижити. Гепатит А не призводить до довгострокових ускладнень, таких як цироз

Ви можете вжити заходів, щоб захистити себе від гепатиту А, зокрема зробити вакцину проти гепатиту А. Якщо у вас гепатит А, ви можете вжити заходів, щоб запобігти поширенню гепатиту А іншим.

Гепатит В - це вірусна інфекція, яка викликає запалення та пошкодження печінки. Запалення - це набряк, який виникає, коли тканини тіла травмуються або інфікуються. Запалення може пошкодити органи[22].

Віруси зовнішнього зв'язку НІН вторгаються в нормальні клітини вашого тіла. Багато вірусів викликають інфекції, які можуть передаватися від людини до людини. Вірус гепатиту В поширюється при контакті з кров'ю, спермою або іншими рідинами організму інфікованої людини.

Ви можете вжити заходів, щоб захистити себе від гепатиту В, включаючи вакцинацію проти гепатиту В. Якщо у вас гепатит В, ви можете вжити заходів, щоб запобігти поширенню гепатиту В серед інших.

Вірус гепатиту В може викликати гостру або хронічну інфекцію.

Гепатит С - це вірусна інфекція, яка викликає запалення та пошкодження печінки. Запалення - це набряк, який виникає, коли тканини тіла травмуються або інфікуються. Запалення може пошкодити органи.

Віруси зовнішнього зв'язку НІН вторгаються в нормальні клітини вашого тіла. Багато вірусів викликають інфекції, які можуть передаватися від людини до людини. Вірус гепатиту С поширюється при контакті з кров'ю інфікованої людини.

Гепатит С може викликати гостру або хронічну інфекцію.

Хоча вакцини проти гепатиту С немає, ви можете вжити заходів, щоб захистити себе від гепатиту С. Якщо у вас гепатит С, поговоріть зі своїм лікарем про лікування. Ліки можутьвилікувати більшість випадків гепатиту С.

Гострий гепатит С є короткочасною інфекцією. Симптоми можуть тривати до 6 місяців. Іноді ваше тіло здатне боротися з інфекцією, і вірус зникає.

Цироз — це стан, при якому ваша печінка утворена рубцями та постійно пошкоджена. Рубцева тканина замінює здорову тканину печінки і перешкоджає нормальної роботи печінки. Рубцева тканина також частково блокує потік крові через вашу печінку. У міру загострення цирозу печінка починає давати збої.

Багато людей не знають, що у них цироз, оскільки вони можуть не мати ознак або симптомів, поки їх печінка не буде сильно пошкоджена. Приклад здорової і хворої печінки зображені на рис. 1.1.



Рисунок 1.1 Здорова і хвора на цирроз печінка

Гемохроматоз - це розлад, при якому в організмі накопичується додаткова зовнішня ланка заліза НІН до шкідливих рівнів.

Ваше тіло потребує заліза, щоб залишатися здоровим, створювати червоні кров'яні тільця, будувати м'язи та клітини серця, а також виконувати щоденні завдання, необхідні вашому тілу та внутрішнім органам. Однак надлишок заліза шкідливий.

Людський організм зазвичай контролює кількість заліза, яке поглинається з дієти, збільшуючи кількість, коли потрібно залізо, і зменшуючи кількість, коли рівень заліза в організмі занадто високий. При гемохроматозі організм щодня поглинає занадто багато заліза з раціону.

Без лікування гемохроматоз може викликати перевантаження залізом, накопичення заліза, яке може пошкодити багато частин тіла, включаючи печінку, серце, підшлункову залозу, ендокринні залози та суглоби.

1.2 Діагностика захворювань печінки

Люди з цирозом печінки на ранній стадії зазвичай не мають симптомів. Часто цироз спочатку виявляється за допомогою звичайного аналізу крові або огляду. Щоб допомогти підтвердити діагноз, зазвичай проводиться комбінація лабораторних і візуалізаційних досліджень.

Ваш лікар може призначити один або кілька тестів, які можуть вказати на проблему з вашою печінкою, включаючи:

Лабораторні дослідження. Ваш лікар може призначити аналізи крові, щоб перевірити ознаки порушення функції печінки, наприклад, надлишок білірубіну, а також певні ферменти, які можуть вказувати на пошкодження печінки. Щоб оцінити функцію нирок, у вашій крові перевіряють креатинін. Вас перевіряють на наявність вірусів гепатиту. Ваше міжнародне нормалізоване співвідношення (МНО) також перевіряється на здатність вашої крові згортатися.

На основі результатів аналізу крові ваш лікар може діагностувати основну причину цирозу. Він або вона може також використовувати аналізи крові, щоб визначити, наскільки серйозний ваш цироз.

Тести візуалізації. Може бути рекомендована магнітно-резонансна еластографія (МРЕ). Цей неінвазивний удосконалений тест візуалізації виявляє затвердіння або затвердіння печінки. Також можуть бути проведені інші тести візуалізації, такі як МРТ, КТ та УЗД.

Біопсія. Зразок тканини (біопсія) не обов'язково потрібен для діагностики. Однак ваш лікар може використовувати його для визначення тяжкості, ступеня та причини ураження печінки.

Якщо у вас цироз, ваш лікар, ймовірно, порекомендує регулярні діагностичні тести для контролю ознак прогресування захворювання або ускладнень, особливо варикозного розширення вен стравоходу та раку печінки. Неінвазивні тести стають все більш доступними для моніторингу. Оцінка та спостереження за фіброзу та цирозу печінки традиційно проводяться за допомогою біопсії печінки. Проте за останні 20 років стало очевидним, що цей «золотий стандарт» недосконалий; навіть за словами його прихильників, це лише «найкращий» серед доступних методів. Спроби розкрити неінвазивні діагностичні інструменти дали численні бали, формули та методи візуалізації. Всі вони краще переносяться, безпечніші, більш прийнятні для пацієнта і можуть повторюватися, по суті, так часто, як це необхідно. Більшість з них набагато дешевше, ніж біопсія печінки. Отже, їх використання зростає, а в деяких країнах кількість біопсій, що проводяться, принаймні для рутинної оцінки гепатиту В і С, різко скоротилася. Однак точність і діагностична цінність більшості, якщо не всіх, цих методів залишається спірною. У цьому огляді для практикуючого лікаря ми аналізуємо усталені та нові біомаркери та фізичні методи. В останні роки ми можемо спостерігати початок кінця першої фази розробки неінвазивних маркерів. Ранні дані свідчать про те, що вони можуть бути принаймні такими ж хорошими, як і біопсія печінки. Нові експериментальні маркери та методи

візуалізації можуть спричинити різкі зміни в діагностиці в найближчому майбутньому.

До середини 20 століття хронічні захворювання печінки можна було з упевненістю діагностувати *ante mortem* лише на дуже запущеній стадії, як правило, після початку цирозу. Першими кількісно визначеними неінвазивними маркерами захворювання печінки були рівні печінкових ферментів у сироватці крові. Лужна фосфатаза і трансамінази стали доступними в 1930 і 1955-1956 роках відповідно. Спектр хронічних захворювань печінки розширився, і стала відомою занурена частина хронічної хвороби печінки «айсберг». Майже одночасно (1958) Менгіні ввів свою техніку і голку «біопсія печінки за одну секунду». Дослідження тканин печінки «*Intra Vitam*» стало можливим і сприяло оголенню додаткових прихованих частин айсберга як якісно, так і кількісно. Пізніше з'явилися різні техніки зображення, які внесли свою частку.

Впровадження все більш ефективних терапевтичних засобів у 1980-х перетворило гепатологію з переважно описової дисципліни в активну, здатну вилікувати багатьох пацієнтів. Більш точне кількісне визначення ступеня ураження печінки стало необхідним для прийняття рішень «Лікувати чи не лікувати», але ні біопсія печінки (LB) [3], ні окремі параметри, такі як аланінамінотрансфераза (ALT), аспартатамінотрансфераза (AST) або тромбоцити цифри були достатні. Стоячи на плечах гігантів, Чайлда і Туркотта та Меддрі та інших, дослідники об'єднали силу окремих параметрів, вставляючи їх у різні оцінки та формули, таким чином значно покращивши їх прогностичну силу.

Багато сироваткових маркерів є ферментами, які вимірюють у звичайних лабораторних дослідженнях, але не є специфічними для печінки і можуть вивільнятися при запаленні інших тканин.

Інші секретуються молекули, такі як білірубін, альфа-фетопротейн, альфа-2-макроглобулін, гаптоглобін та аполіпопротеїн A1.

Альбумін специфічно виділяється з печінки, і його рівні знижуються переважно при тяжких захворюваннях печінки, а також при інших клінічно значущих захворюваннях (запальні захворювання, захворювання нирок зі значною протеїнурією, гіпотрофія, ентеропатія з втратою білка). Тому, хоча альбумін є хорошим показником поганого самопочуття, йому бракує специфічності щодо захворювань печінки.

Жоден з цих маркерів сам по собі не приносить особливої користі, але вони корисні при поєднанні в маркерних панелях[11].

В останні роки для клінічного використання були створені комбінації біомаркерів або маркерних панелей. Усі вони засновані на непрямих біомаркерах (див. нижче), за винятком гіалуронової кислоти або гіалуронану (НА) і панелей, які містять її (Fibrometer і Hepascore).

1.3 Текстурні ознаки

Текстура є однією з важливих характеристик, що використовуються для ідентифікації об'єктів або регіонів, що представляють інтерес на зображенні, будь то мікрофотографія, аерознімок чи супутниковий знімок. У цій статті описуються деякі легко обчислювані текстурні характеристики на основі просторових залежностей сірих тонів і ілюструється їх застосування в завданнях ідентифікації категорій трьох різних типів даних зображення: мікрофотографії п'яти видів пісковиків, 1:20 000 панхроматичних аерофотознімків восьми земель. -категорії використання та мультиспеціальні зображення супутника технології земних ресурсів (ERTS), що містять сім категорій землекористування. Ми використовуємо два типи правил прийняття рішень: одне, для якого області рішення є опуклими багатогранниками (правило кусково-лінійного рішення), і одне, для якого області рішення є прямокутними паралелпіпедами (правило прийняття рішень мінімального і максимального). У кожному експерименті набір даних був поділений на дві частини, навчальний набір і тестовий набір. Точність

ідентифікації тестового набору становить 89 відсотків для мікрофотографій, 82 відсотки для аерофотознімків і 83 відсотки для супутникових знімків. Ці результати вказують на те, що легко обчислювані текстурні характеристики, ймовірно, мають загальне застосування для широкого спектру додатків для класифікації зображень.

1.4 Методи візуалізації

Магнітно-резонансна томографія (МРТ) зазвичай використовується для оцінки цирозу та його ускладнень. Однак виявити менш запущені стадії фіброзу є більш складним завданням, і для цієї мети було використано кілька нових методів МРТ [88]. Встановлені модальності.

Звичайна МРТ: морфологічні зміни, пов'язані з цирозом печінки, можна оцінити за допомогою звичайної МРТ. Макроструктурні зміни включають вузлуватість поверхні, розширення тріщин, розширення ямки жовчного міхура, виїмку правої частки та збільшення бічних сегментів лівої частки та хвостатої частки. Зміни паренхіми включають фіброзні перегородки та містки, регенеративні вузлики та сидеротичні вузлики або стеатотичні вузлики. Інші помітні зміни в деяких випадках пов'язані з портальною гіпертензією, включаючи спленомегалію, портосистемні варикозні розширення вен, асцит і потовщення стінки кишечника. Введення внутрішньовенного (iv) контрастного матеріалу покращує видимість фіброзу та пов'язаних з цирозом змін та ускладнень цирозу. Фіброз має специфічний характер посилення з піковим посиленням на пізніх фазах (фази венозної/рівноважної). Цей характерний малюнок посилення та ретикулярний вигляд дозволяють диференціювати його від інших судинних уражень, пов'язаних з цирозом (наприклад, артеріо-портальних шунтів, ГЦК) [89]. Інноваційні техніки

МР-еластографія: як і сонографічна транзитрна еластографія (ТЕ), МР-еластографія заснована на тому факті, що швидкість і довжина хвилі, що

поширюється в тканині, збільшуються зі збільшенням жорсткості середовища, наприклад, фіброзної печінки. Для виконання МР-еластографії потрібне спеціальне програмне та апаратне забезпечення. Пристрій-драйвер розміщується над правою верхньою частиною живота пацієнта і створює акустичні хвилі тиску з частотою 40-120 Гц. Ці хвилі створюють зсувні хвилі в печінці. Зображення зображують механічну хвилю, яка поширюється, а певний алгоритм створює кількісну карту жорсткості[24].

У кількох дослідженнях МР-еластографія виявила прогресуючий фіброз і цироз у пацієнтів з НАЖХП і хронічним гепатитом В. Кількісна оцінка суттєво корелювала зі стадією фіброзу. Він також виявився ефективним засобом для диференціації нижчих і вищих ступенів цирозу.

У порівнянні з іншими методами МРТ, МР-еластографія є більш чутливою для оцінки фіброзу та цирозу печінки порівняно з морфологічними ознаками, виявленими за допомогою звичайної МРТ. Він також має набагато вищу узгодженість між спостерігачами порівняно з МРТ та дифузійно зваженими зображеннями (DWI). На відміну від ультразвукового дослідження, на МР-еластографію не впливає відсутність акустичного вікна, ожиріння або наявності асцитів, і вона не залежить від оператора. Huwart та співавтори [90] показали, що МР-еластографія є більш точною, ніж еластографія США, APRI або їх комбінація, і її повторюваність коефіцієнта була кращою, ніж еластографія США.

У мета-аналізі 12 досліджень, проведених Singh et al [91] з використанням LB як стандарту, МР-еластографія виявилася високоточною для діагностики розвиненого фіброзу, незалежно від віку, статі, ІМТ, запалення та етіології печінки. захворювання.

Обмеженнями МР-еластографії є її вартість і те, що вона займає багато часу. На жорсткість печінки можуть впливати також перевантаження печінки залізом, стеатоз, судинний застій, холестази та портальна гіпертензія. У цих випадках точність МР-еластографії може бути змінена.

Картування печінки T1: у цьому методі отримують зображення часу релаксації T1 і створюють карти T1 за допомогою програмного забезпечення сканера. Haimeiri та співавтори показали, що карти T1 після введення специфічного для печінки контрастного засобу (Gd-EOB-DTPA) корелювали зі стадією цирозу, але кореляції між фіброзом та зображеннями без контрасту виявлено не було. Інші дослідження Allkemper et al та Rauscher et al [103] виявили кореляцію між цирозом і часом розслаблення T1 у МР без контрасту.

У дослідженні Banerjee et al використовувалося картування T1 для оцінки фіброзу, 1H МР-спектроскопія для кількісного визначення вмісту ліпідів і послідовність T2* для оцінки перевантаження залізом. Алгоритм створив значення T1, виправлене залізом, усуваючи вплив підвищеного вмісту заліза на значення T1. Значення МР порівнювали з даними гістології. Виправлене значення T1 визначило фіброз з чутливістю 86% і специфічністю 93% і сильно корелювало з різними стадіями фіброзу, за винятком перекриття між легким і помірним фіброзом. Крім того, 1H МР спектроскопія сильно корелює зі стеатозом печінки. Вміст заліза в печінці мав сильну негативну кореляцію з T2*. У цьому дослідженні дані для всіх трьох параметрів - фіброз, стеатоз і вміст заліза - були отримані під час 23-хвилинного сканування.

Контрастні речовини, специфічні для ретикулоендотеліальної системи: було проведено кілька досліджень із специфічними для ретикулоендотеліальної системи контрастними речовинами. Суперпарамагнітний оксид заліза (SPIO) спричиняє падіння сигналу в гепатоциті, що містить паренхіму печінки, і, як наслідок, підвищує помітність виявлення фіброзної тканини, яка менше піддається впливу цього контрастного засобу.

Інші дослідження досліджували техніку МРТ з подвійним контрастом. Ця методика поєднує контрастну речовину на основі гадолінію та SPIO в одному дослідженні. Синергетичний ефект обох контрастних речовин збільшує видимість фіброзної тканини і допомагає диференціювати

запущений фіброз печінки від легкого фіброзу. Цей метод також дозволив кількісно визначити структуру печінки та її кореляцію зі стадією фіброзу. Однак ці контрастні речовини більше клінічно недоступні.

MPT, зважена на чутливість: зображення, зважене на чутливість, є градієнтною ехо-послідовністю з підвищеною чутливістю до присутності заліза, гемоглобіну та кальцифікатів. Було показано, що вимірювання співвідношення інтенсивності сигналу печінки та м'язів корелює з фіброзом печінки з високою узгодженістю між спостерігачами.

Дифузійно зважена MPT: послідовності DWI оцінюють здатність протонів дифундувати в тканині. Ця послідовність зазвичай використовується в онкологічних цілях. Карта видимого коефіцієнта дифузії (ADC) є розрахунковою картою, отриманою на основі зображень DWI і корелює зі здатністю протона до дифузії. Попередні дослідження з використанням різних апаратних засобів і різних послідовностей намагалися встановити кореляцію між зниженим значенням ADC, яке з'являється при фіброзі, і ступенем фіброзу, але результати не були узгоджені. У дослідженнях Razek et al та Lewin et al DWI корелював з фіброзом у дітей та дорослих. В іншому дослідженні DWI корелював зі стадіями фіброзу з чутливістю 75%-85% (залежить від стадії фіброзу) і специфічністю 68%-94%. У цьому дослідженні здатність ідентифікувати фіброз була значно вищою при МР-еластографії, ніж DWI.

Обмеження DWI в оцінці фіброзу пов'язані з тим, що на дифузію впливають зміни перфузії, стеатоз печінки, наявність заліза в тканині та запальні зміни. Більше того, послідовність чутлива до артефактів, пов'язаних із сприйнятливістю та рухом; і оскільки кількісний аналіз заснований на зображеннях, він також дуже обмежений.

Висновки до розділу 1

В даному розділі були описані захворювання печінки. Глобальна

поширеність цирозу за результатами розтину коливається від 4,5% до 9,5% загальної популяції 5, 6, 7. Таким чином, за нашими оцінками, понад п'ятдесят мільйонів людей у світі, враховуючи доросле населення, будуть хворіти на хронічне захворювання печінки . У всьому світі алкоголь, НАСГ та вірусний гепатит наразі є найпоширенішими причинними факторами. Поширеність цирозу, ймовірно, буде недооцінена, оскільки майже третина пацієнтів залишається безсимптомною. З використанням неінвазивних тестів, таких як транзитрна еластографія, найближчим часом може виникнути більш реалістична картина. Протягом 2001 року в усьому світі смертність від цирозу печінки становила 771 000 осіб, що займає 14-е та 10-е місце серед провідних причин смерті у світі та в розвинених країнах відповідно, 8. За оцінками, смертність від цирозу печінки збільшується, і вона посідає 12-е місце. причина смерті у 2020 році.

РОЗДІЛ 2

АЛГОРИТМИ КЛАСИФІКАЦІЇ ТА МЕТОДИ СЕЛЕКЦІЇ ОЗНАК

Класифікація – це процес розпізнавання, розуміння та групування ідей та об'єктів у попередньо встановлені категорії або «підгрупи». Використовуючи попередньо категоризовані набори навчальних даних, програми машинного навчання використовують різноманітні алгоритми для класифікації майбутніх наборів даних за категоріями.

Алгоритми класифікації в машинному навчанні використовують вхідні навчальні дані, щоб передбачити ймовірність того, що наступні дані потраплять в одну із заздалегідь визначених категорій. Одним із найпоширеніших застосувань класифікації є фільтрація електронних листів на «спам» або «не-спам».

Коротше кажучи, класифікація — це форма «розпізнавання шаблонів», з алгоритмами класифікації, які застосовуються до навчальних даних, щоб знайти той самий шаблон (подібні слова чи почуття, числові послідовності тощо) у майбутніх наборах даних.

Використовуючи алгоритми класифікації, які ми докладніше розглянемо нижче, програмне забезпечення для аналізу тексту може виконувати такі завдання, як аналіз настроїв на основі аспектів, щоб класифікувати неструктурований текст за темою та полярністю думок (позитивні, негативні, нейтральні та інші).

Спробуйте цей попередньо навчений класифікатор настроїв, щоб зрозуміти, як алгоритми класифікації працюють на практиці, а потім читайте далі, щоб дізнатися більше про різні типи алгоритмів класифікації.

Алгоритм класифікації – це методика навчання під керівництвом, яка використовується для визначення категорії нових спостережень на основі навчальних даних. У класифікації програма вивчає заданий набір даних або спостереження, а потім класифікує нове спостереження на ряд класів або

груп. Наприклад, так чи ні, 0 чи 1, спам чи не спам, кішка чи собака тощо. Класи можна називати цілями/мітками чи категоріями.

На відміну від регресії, вихідна змінна класифікації є категорією, а не значенням, як-от «Зелений чи синій», «фрукт чи тварина» тощо. Оскільки алгоритм класифікації є методом навчання з наглядом, він приймає позначені вхідні дані, які означає, що містить вхідні дані з відповідним виводом[10].

2.1 K-найближчих сусідів (KNN).

Хорошим прикладом алгоритму, який є загальним апроксиматором зв'язків між наборами даних, є алгоритм k-найближчого сусіда. Цей алгоритм не робить жодних припущень щодо взаємозв'язків між даними, але замість цього використовує лише дані для створення нових прогнозів із спостережуваними даними; модель притаманна даним, і можна навіть сказати, що «дані є моделлю». Цей алгоритм знаходить приклади серед спостережень, які найбільше схожі на нові дані, представлені алгоритму, а потім призначає новій точці даних прогнозоване значення зразка, найбільш схожого на нього.

На рис. 2.1 наведено приклад рішення класифікації за допомогою алгоритму k-найближчого сусіда. У цьому наборі даних є 2 змінних предиктора, перший і другий. Існуючі спостереження в наборі навчальних даних представлені колами і квадратами, категорія А та категорія В відповідно, у межах діаграми розсіювання. Нові спостереження представлені трикутниками. Завдання алгоритму k-найближчого сусіда полягає в тому, щоб передбачити категорію (А або В), до якої належать трикутні (нові) точки даних, представлених на рис. 2.1.

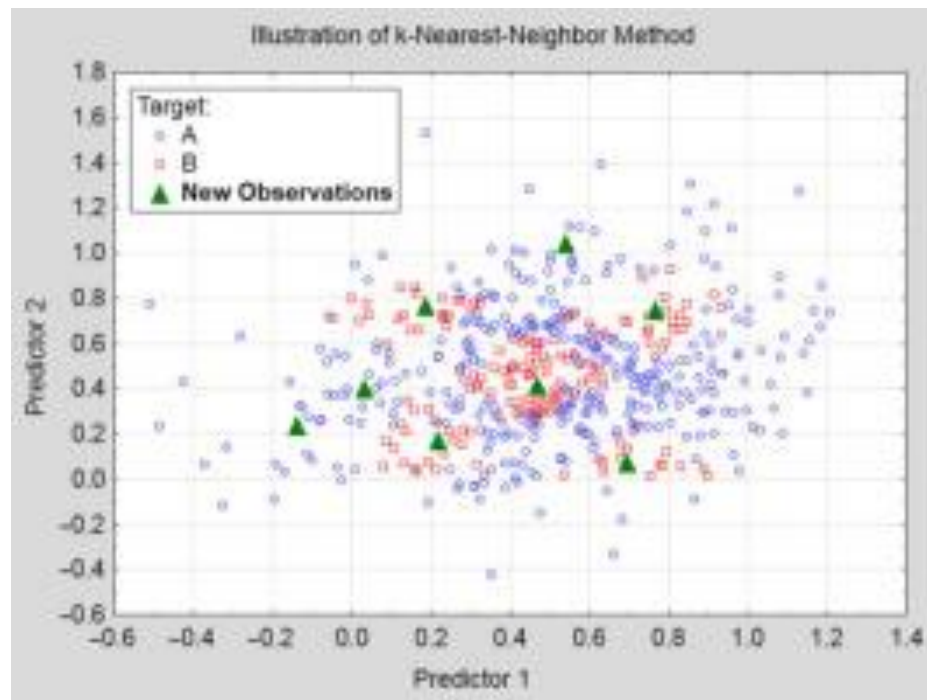


Рисунок 2.1. Класифікація нових спостережень із використанням методу k-найближчого сусіда.

Візуально оглянувши діаграму розсіювання, ви можете помітити, що положення трикутників розташовані в групах квадратів і кіл, які утворюють досить однорідні околиці навколо трикутників. Інтуїтивно легко прогнозувати правильну класифікацію нових спостережень (трикутників), ґрунтуючись на формі більшості точок даних, нанесених на графік у цих околицях. Хоча цей приклад класифікації може здатися легким, завдання пошуку лінійного (найкращої відповідності) рішення для цього прикладу або навіть криволінійного рішення було б надзвичайно складним, якщо не неможливою, що ілюструє корисність цього конкретного загального алгоритму апроксимації.

Алгоритм k-найближчого сусіда також відомий як «який навчається на основі пам'яті». Загальний підхід до моделювання аналогічний гри: перевернути флеш-карту, щоб побачити картинку, перевернути її, а потім спробувати передбачити, які флеш-карти з'являться наступного разу, ґрунтуючись на вашій пам'яті про те, де були попередні зображення.

розташований впалуба. Алгоритм запам'ятовує, де знаходяться всі дані, що спостерігаються, а потім робить прогнози для нових даних на основі пам'яті про те, де знаходяться спостерігаються дані. Для практичної реалізації цього загального апроксиматора є додаткові важливі міркування, такі як:

- Як обчислюється «відстань» в деякій метриці подібності між спостереженнями, що спостерігаються і прогнозованими.
- Скільки сусідів враховуватиме при прогнозуванні.
- Як зробити остаточний прогноз, враховуючи більше одного найближчого сусіда.

Переваги алгоритму k-найближчого сусіда. Двома основними перевагами алгоритму k-найближчого сусіда є ефективність та гнучкість. Алгоритм ефективний своєю простотою, швидкістю та масштабованістю. Як описано вище, механізм алгоритму очевидний, його легко зрозуміти та реалізувати. Через простоту він також дуже швидкий, і ця швидкість робить його дуже простим для розгортання на дуже великих і складних наборах даних, що робить його масштабованим для даних, де інші складніші алгоритми можуть відчувати труднощі. Гнучкість цього алгоритму проявляється у його здатності обробляти набори даних, у яких відносини дуже складні та незрозумілі ні лінійними, ні криволінійними відносинами.

Через здатність цього алгоритму визначати подібність між даними, він часто використовується для заміни відсутніх даних. Це вимагає відповідної попередньої обробки змінних, що використовуються в аналізі, а також встановлення відповідних параметрів для вибору, які спостереження алгоритм запам'ятовуватиме, щоб робити нові прогнози на невидимих даних. Однак, як тільки це буде зроблено, він виконує дуже хорошу роботу з обчислення значень чи прогнозів для відсутніх даних. Наприклад, методи, описані у цій книзі, можна використовувати для класифікації медичних заяв. Цього можна досягти, спочатку попросивши експертів вибрати типові «зразкові претензії», які становлять інтерес, які відносяться до певної

категорії претензій, такої як шахрайство, а потім використовувати ці зразки для автоматичного присвоєння нових претензій категоріям, що цікавлять.

Недоліки алгоритму k-найближчого сусіда. У певному сенсі цей алгоритм насправді не є алгоритмом навчання, а ґрунтується на пам'яті «пошуком». Змінні-предиктори і навчальна вибірка (примірники) повинні бути обрані апіорі. Потім алгоритм запам'ятовує ці дані і робить прогнози на основі цієї пам'яті. Алгоритм не «вивчає» закономірності даних. З іншого боку, щойно навчальну вибірку визначено, необхідно вирішити, скільки спостережень використовуватиметься найближчих сусідів під час прогнозування нових даних. Крім того, цей алгоритм страждає від прокляття розмірності (кількості предикторів), і точність прогнозування може сильно постраждати, коли використовуються сотні, якщо не тисячі змінних-предикторів, що часто має місце в багатьох наборах даних, для яких алгоритми інтелектуального аналізу даних часто використовуються. Деякі інші алгоритми інтелектуального аналізу даних дозволяють зрозуміти, які змінні є найважливішими для створення нових прогнозів. Ця можливість недоступна в алгоритмі k-найближчого сусіда і, таким чином, є суттєвим недоліком. Розширення підходу[20].

Алгоритм машини опорних векторів (SVM) є розширенням цього підходу, який розширює алгоритм k-найближчого сусіда з учня на основі пам'яті до реального алгоритму навчання. Подібно до алгоритму k-найближчого сусіда, SVM передбачає, що алгоритм визначає околиці на основі спостережуваних даних, представлених алгоритму, з яких робляться прогнози для нових невидимих даних. Однак SVM роблять додаткові кроки, розділяючи навчальні дані на розділи. Ці розділи створюються з використанням ліній або гіперплощин (що складаються з більш ніж трьох вимірювань), які обчислюються на основі даних навчання з використанням векторів кожної змінної-предиктора для складання розмірів гіперплощини. SVM схожі на інші алгоритми навчання, особливо нейронні мережі.

2.2 Метод опорних векторів (Support Vector Machine – SVM) з лінійною функцією ядра

SVM - це контрольований алгоритм навчання, який може бути корисним для цілей класифікації та регресії (Vapnik, 1998). SVM можна використовувати для аналізу даних для класифікації та регресії з використанням алгоритмів та ядер у SVM (Cortes and Vapnik, 1995). Класифікація опорних векторів (SVC) також являє собою алгоритм, який шукає оптимальну поверхню, що розділяє. SVC спочатку описується для лінійно відокремленого випадку (Burbidge and Buxton, 2001). Методи ядра SVM використовуються, коли повний поділ двох класів неможливий. Деякі широко використовувані ядра в SVM – це поліноміальні, квадратичні та радіальні базисні функції[21].

Можна розширити методологію опорних векторів (у сенсі використання карти ознак у первинному та заснованого на ядрі уявлення в дуальному) для ширшого кола завдань. Це було зроблено у контексті LS-SVM. Один замінює обмеження нерівності обмеженнями рівності та використовує функції втрат. Таким чином, «основні моделі» виходять для ядерних методів у класифікації, регресії, PCA, CCA, спектральної кластеризації, рекурентних мережах та інших.

Класифікатори SVM та LS-SVM були протестовані на багатьох проблемах порівняно з кількома традиційними підходами, наприклад Van Gestel et al . Байєвський висновок також може бути виконаний для визначення гіперпараметрів і виконання вибору моделі, а також вибору ознак за допомогою автоматичного визначення релевантності. Ці методи незмінно дуже добре справляються з широким колом завдань. Для вибору моделі часто використовуються методи перехресної перевірки для встановлення константи регуляризації γ та визначення параметрів ядра (таких як у ядрі радіальної базисної функції). Зокрема, перехресна перевірка з винятком по одному може бути ефективно виконана, як показано, наприклад, Cawley та Talbot.

працюючи з іншою компромісною схемою регуляризації. Константи регуляризації також можуть бути отримані опуклою оптимізацією.

У разі найменших квадратів одержана модель не є розрідженою. Однак розрідженість може бути досягнута за рахунок додаткової обрізки або скорочення або за рахунок роботи з методами фіксованого розміру, коли кінцева розмірна апроксимація карти ознак береться з оцінкою первинної. Крім того, оптимальна розрідженість на рівні перевірки може бути отримана під час роботи з підкладками LS-SVM. Статистична стійкість може бути досягнута шляхом додаткового зваженого методу найменших квадратів, а допомогою ітеративно зважених найменших квадратів можна вирішити для стандартної SVM.

2.3 SVM з функцією радіального базису (Radial Basis Function–RBF)

Радіальні базисні функції - це засоби апроксимації функцій багатьох змінних (також званих багатовимірними) лінійними комбінаціями членів на основі однієї функції одного параметра (радіальна базова функція). Він виконаний за радіусом, тому його можна використовувати більш ніж в одному вимірі. Зазвичай вони застосовуються до наближених функцій або даних (Пауелл, 1981, Чейні, 1966, Девіс, 1975), які відомі тільки в кінцевому числі точок (або занадто важкі для оцінки в іншому випадку), так що тоді оцінки апроксимуючої функції можуть мати місце часто та якісно.

Насамперед у обчислювальних додатках функції багатьох змінних часто необхідно апроксимувати іншими функціями, які краще зрозуміти чи легше обчислити. Це може бути зроблено для їхнього частого відображення, наприклад, на екрані комп'ютера, тому комп'ютерна графіка є областю практичного використання. Радіальні базисні функції - один із ефективних і часто використовуваних способів зробити це. Інші програми включають важливі області нейронної мережі та теорія навчання. Оскільки вони є

радіально-симетричними функціями, які зсуваються точками в багатовимірному евклідовому просторі, а потім лінійно комбінуються, вони утворюють апроксимації, що залежать від даних простору. Ця залежність від даних робить сформовані таким чином простори, що підходять для забезпечення наближень до великих класів заданих функцій. Це також відкриває двері до існування та унікальності результатів для інтерполяції розрізнених даних радіальними базовими функціями у дуже загальних налаштуваннях (зокрема, у багатьох вимірах).

Насправді, одна з найбільших переваг цього методу полягає в його застосовності практично в будь-якому вимірі (звідси і його універсальність), оскільки зазвичай існують невеликі обмеження способу завдання даних. Приклади, наведені нижче, включають ядра з позитивною визначеністю, які не мають обмежень на дані, за винятком того, що вони повинні знаходитися в різних точках. Цьому слід протиставити, наприклад, багатопараметричну поліноміальну інтерполяцію (але див. Особливо гнучкий підхід у de Boor and Ron 1990) або сплайн. Ще однією перевагою є їхня висока точність або швидка збіжність до наближеної цільової функції у багатьох випадках, коли дані стають щільними[23].

Для додатків дійсно бажано, щоб було кілька умов геометрії або напрямків, в яких точки даних повинні бути розміщені в просторі. Для алгоритмів радіальної базової функції не потрібно тріангуляції точок даних і т.п., тоді як, наприклад, методи кінцевих елементів або багатовимірні сплайнові методи зазвичай вимагають тріангуляції. Фактично, попереднє структурування даних, від якого залежать деякі інші схеми апроксимації, може бути надмірно дорогим для обчислень у додатках, особливо у більш ніж двох вимірах. Тому наші наближення тут розглядаються як безсітчні наближення, також, наприклад, для спрощення чисельного розв'язання рівнянь у приватних похідних. З іншого боку, як буде згадано нижче, розширені чисельні методи обчислення наближень радіальної базисної

функції потрібні, коли набір даних великий, тоді як потрібно більш стандартне програмне забезпечення, наприклад, для кінцевих елементів.

2.4 Гаусівський процес

Гаусівський процес - це розподіл ймовірностей за можливими функціями, які відповідають набору точок. Переваги гаусівських процесів:

1. Прогноз інтерполює спостереження (принаймні для звичайних ядер).
2. Прогноз є імовірнісним (гаусівським), так що можна обчислити емпіричні довірчі інтервали і на їх основі вирішити, чи слід коригувати (онлайн-підбір, адаптивне припасування) прогноз в деякій області, що цікавить.
3. Універсальність: можна вказати різні ядра. Надаються загальні ядра, але можна вказати власні ядра.

До недоліків гаусівських процесів можна віднести: вони не є розрідженими, тобто використовують всю інформацію про ознаки по характеристиках для виконання прогнозу. Вони втрачають ефективність у просторах великих розмірів, саме, коли кількість функцій перевищує кілька десятків.

Гаусівські процеси узагальнюють багатовимірні гаусівські розподіли за кінцевими векторами до нескінченної розмірності. Зокрема, гаусівський процес – це випадковий процес, який має гаусівські розподілені кінцеві маргінальні розподіли, звідси й назва. У цьому він визначає розподіл за функціям, т. е. кожне вилучення з гаусовського процесу є функцією. Гаусівські процеси забезпечують принциповий, практичний та ймовірнісний підхід до логічних висновків та навчання в машинах з ядром.

Байєсовські ймовірні підходи мають багато переваг, у тому числі їх здатність включати попередні знання і їх здатність пов'язувати пов'язані джерела інформації. Як правило, нам дається набір точок даних, взятих з базового, але невідомого розподілу, кожна з яких включає вхід x і вихід y ,

наприклад, показані на рис. 1а. Завдання - дізнатися функціональний зв'язок між x та y . Традиційно, у параметричному підході, припущення про математичну форму взаємозв'язку, таку як лінійна, поліноміальна, експоненційна або їх комбінація, має бути обране апіорі. Згодом ваги (або параметри) поміщаються в кожен з обраних форм, а потім визначається попередній розподіл за параметрами. Таким чином, завдання навчання тепер зводиться до байєсівської оцінки параметрів, порівн. Мал. 1а-с. Однак цей підхід не завжди може бути практичним, як показано на рис. 1d. Щоб виявити прихований взаємозв'язок між входом та виходом на рис. 1d, нам може знадобитися безліч функціональних форм, а це означає нескінченну кількість параметрів. Замість роботи з простором параметрів, гауссівські процеси поміщують апіор прямо в простір функцій без параметризації функції, отже, непараметрично. Як показано нижче, обчислювальна складність виведення тепер масштабується як кількість точок даних, а чи не кількість параметрів на рис. 2.2.

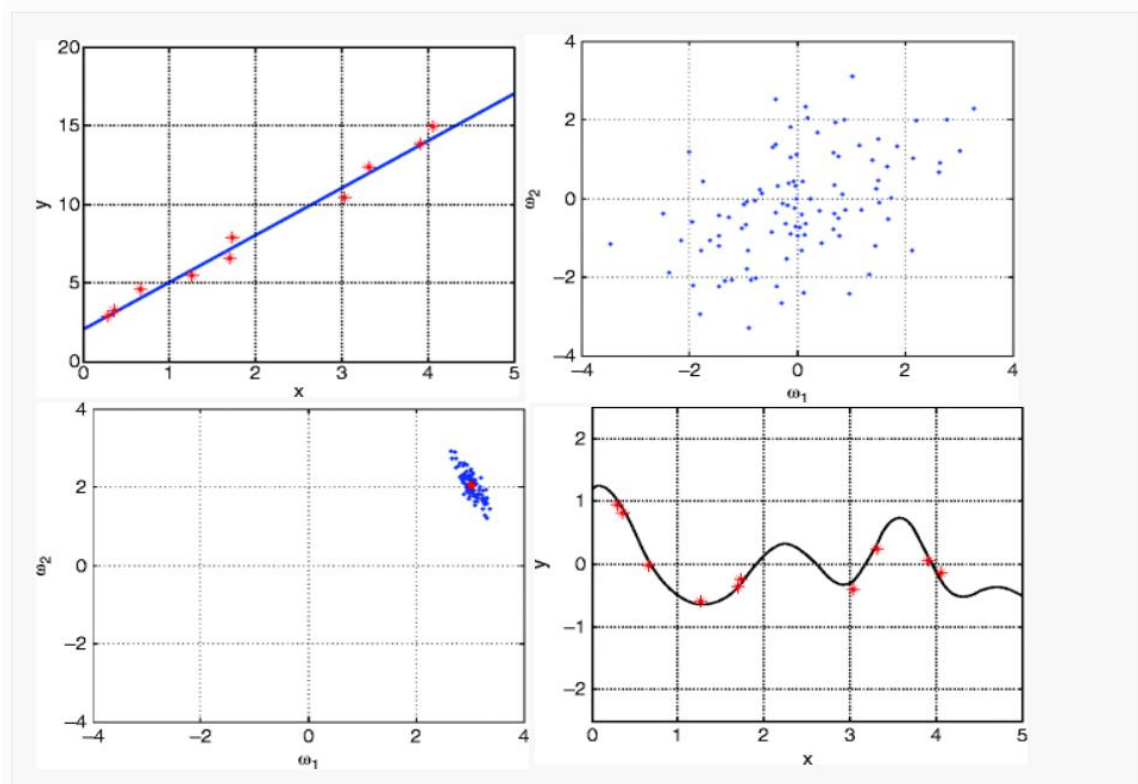


Рисунок 2.2 Гаусівський процес

Десять спостережень (одномірні вхідні змінні x та вихідні y), згенеровані з моделі лінійної регресії $y = 3x + 2 + \varepsilon$ з гаусовим шумом ε . Завдання полягає в тому, щоб дізнатися функціональний зв'язок між x та y . Припускаючи, що параметрична модель $y = \omega_1 x + \omega_2 + \varepsilon$, тобто, апостеріорний розподіл по ω можна оцінити, як показано на (c). Його середнє значення (3,0287, 2,0364) близьке до дійсних параметрів (3, 2). Висновок, однак, був зроблений в ідеальній ситуації, коли співвідношення між x та y дійсно було лінійним. Якщо справжній взаємозв'язок невідомий заздалегідь та/або не може бути легко описаний з використанням кінцевого набору параметрів, цей підхід може зазнати невдачі. Наприклад, (d) може знадобитися нескінченна кількість параметрів для відновлення функціонального взаємозв'язку[24].

Було розроблено кілька непараметричних байєсовських моделей для різних завдань, таких як оцінка щільності, регресія, класифікація, аналіз часу виживання, тематичне моделювання тощо. Як і гаусівський процес, процес Діріхле має кінцеві маргінальні розподіли Діріхле, звідси і назва.

Формально гаусівський процес - це випадковий процес (тобто. набір випадкових величин), у якому все кінцеві розподіли є багатовимірними гауссовськими розподілами для будь-якого кінцевого вибору змінних.

Гаусівський розподіл повністю визначається його середнім значенням та матрицею коваріації.

Коваріаційна функція відіграє у моделі гаусовського процесу, оскільки її властивості безперервності визначають властивості безперервності вибірок (функцій) з гаусовського процесу. У літературі коваріаційні функції також відомі як позитивні (напів) певні ядра або ядра Мерселя.

Зазвичай існує два типи коваріаційних функцій: стаціонарні та нестаціонарні.

2.5 Дерево прийняття рішень (ДПР)

Дерева рішень – це моделі підтримки прийняття рішень, які класифікують шаблони з використанням послідовності чітко визначених правил. Це деревоподібні графи, в яких кожен вузол гілки є варіантом між кількома альтернативами, а кожен кінцевий вузол є результатом сукупного вибору. Щоб застосувати дерева рішень для прогнозування пептидів зв'язування МНС, позиційно-специфічні мотиви зв'язування спочатку перетворюються на серію правил. Потім кожне правило впроваджується у вузли дерева. Отримана деревоподібна структура вказує на властивості амінокислот, які сильно корелюють з фізико-хімічними властивостями пептидів, що зв'язують. Потім пептидні послідовності можуть бути пропущені через ряд вузлів і результати всіх переходів від вузла до вузла використовуються для виведення їх специфічності зв'язування. Приклад мережі дерева рішень показаний на рис. 2.3.

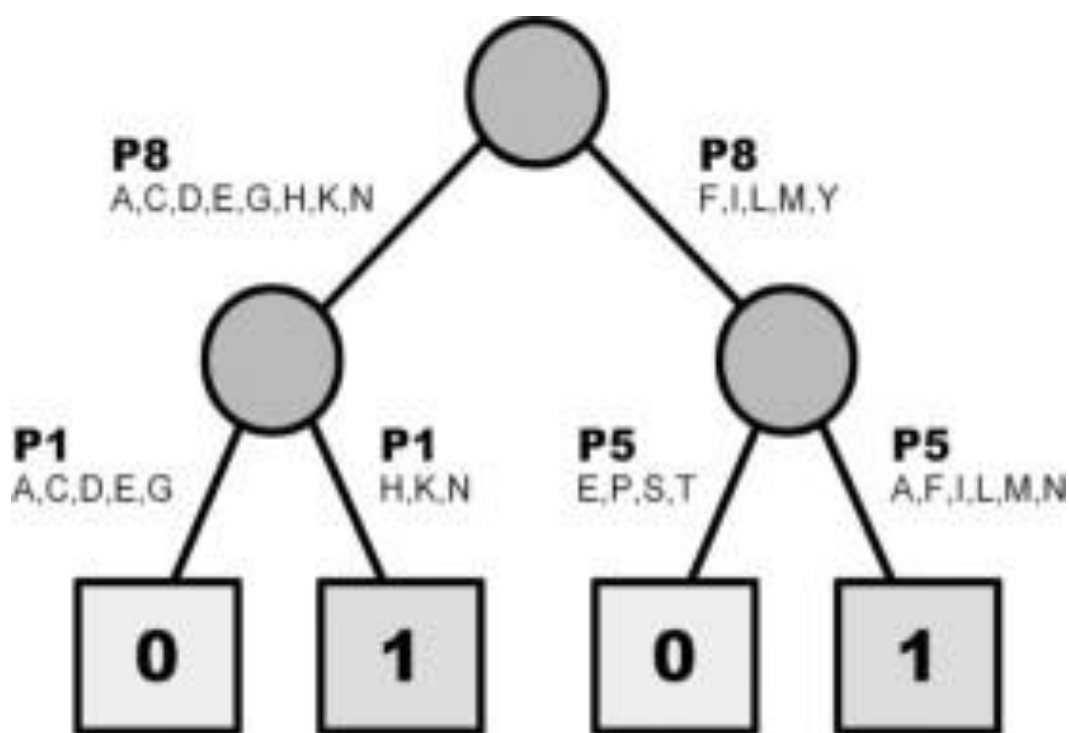


Рисунок 2.3 Дерево рішень

Дерева рішень - важливий і широко використовується метод класифікації даних. Нещодавно вони були адаптовані для використання у обчислювальній біології, особливо для аналізу даних послідовностей ДНК. У цьому розділі описується, як дерева рішень використовуються у системі пошуку генів MORGAN [1], і як можуть бути використані на вирішення інших завдань аналізу послідовностей. Спочатку я представляю вступний курс з дерев рішень, що описує, що вони з себе представляють і як їх можна створювати автоматично з даних. Читачі, які вже знайомі з алгоритмами дерева рішень, можуть пропустити цей навчальний розділ і одразу перейти до розділу, що описує застосування дерев рішень в аналізі послідовності ДНК. Навчальний посібник також містить вказівники на кілька місць в Інтернеті, де зацікавлений читач може знайти та завантажити програмне забезпечення для дерева рішень[26].

Обчислюючи очікувану корисність чи цінність кожного варіанта у дереві, ви можете мінімізувати ризик та максимізувати ймовірність досягнення бажаного результату.

Щоб розрахувати очікувану корисність вибору, просто відніміть вартість цього рішення від очікуваних результатів. Очікувані вигоди дорівнюють спільній цінності всіх результатів, які можуть виникнути в результаті цього вибору, причому кожне значення множиться на ймовірність того, що це станеться.

При визначенні того, який результат є найбільш бажаним, важливо враховувати переваги корисності особи, яка приймає рішення. Наприклад, деякі можуть віддати перевагу варіантам з низьким рівнем ризику, в той час як інші готові піти на ризик заради більшої вигоди.

Коли ви використовуєте своє дерево рішень з супутньою моделлю ймовірності, ви можете використовувати його для розрахунку умовної ймовірності події або ймовірності того, що воно відбудеться, за умови, що відбудеться інша подія. Для цього просто почніть з початкової події, потім

простежте шлях від цієї події до цільової події, помножуючи ймовірність кожного з цих подій разом.

Таким чином, дерево рішень можна використовувати як традиційну деревоподібну діаграму, яка відображає ймовірності певних подій, таких як двічі підкидання монети.

Дерева рішень залишаються популярними з таких причин: наскільки легко їх зрозуміти, вони можуть бути корисні з твердими даними або без них і будь-які дані вимагають мінімальної підготовки. Нові параметри можуть бути додані до існуючих дерев.

Їхня цінність у виборі кращого з кількох варіантів. Наскільки легко вони поєднуються з іншими інструментами прийняття рішень

Однак дерева рішень можуть стати надмірно складними. У таких випадках оптимальною альтернативою може бути більш компактна діаграма впливу.

Дерево рішень також можна використовувати для створення автоматизованих прогнозних моделей, які мають програми у машинному навчанні, інтелектуальному аналізі даних та статистиці. Цей метод, відомий як навчання з дерева рішень, бере до уваги спостереження за елементом для прогнозування його цінності.

У цих деревах рішень вузли представляють дані, а чи не рішення. Цей тип дерева також відомий як класифікація дерева. Кожна гілка містить набір атрибутів або правил класифікації, пов'язаних із певною міткою класу, що знаходиться в кінці гілки.

Ці правила, також відомі як правила прийняття рішень, можуть бути виражені в пропозиції if-then, при цьому кожне рішення або значення даних формують пропозицію, наприклад, «якщо умови 1, 2 та 3 виконані, то результат x буде результатом з упевненістю» .

Кожен додатковий фрагмент даних допомагає моделі точніше передбачити, якого з кінцевого набору значень належить аналізований об'єкт.

Потім цю інформацію можна використовувати як вхідні дані в більшій моделі прийняття рішень.

Іноді прогнозована змінна може бути дійсним числом, наприклад, ціною. Дерева рішень із безперервними, нескінченними можливими наслідками називаються деревами регресії.

2.6 Випадковий ліс.

Випадковий ліс - контрольований алгоритм машинного навчання, побудований на основі алгоритмів дерева рішень. Цей алгоритм застосовується у різних галузях, таких як банківська справа та електронна комерція, для прогнозування поведінки та результатів.

Випадковий ліс - це метод машинного навчання, який використовується для вирішення задач регресії та класифікації. Він використовує ансамблеве навчання, яке є методом, що поєднує безліч класифікаторів для пошуку рішень складних проблем.

Алгоритм випадкового лісу складається з безлічі дерев рішень. Ліс, згенерований алгоритмом випадкового лісу, навчається за допомогою пакетування або агрегування початкового завантаження. Bagging – це ансамблевий метаалгоритм, що підвищує точність алгоритмів машинного навчання.

Алгоритм (випадковий ліс) встановлює результат з урахуванням прогнозів дерев рішень. Він прогнозує, взявши середнє значення вихідних даних різних дерев. Збільшення кількості дерев підвищує точність результату.

Випадковий ліс усуває обмеження алгоритму дерева рішень. Це знижує можливість перенавчання наборів даних і підвищує точність. Особливості алгоритму випадкового лісу. Він точніший, ніж алгоритм дерева рішень. Він забезпечує ефективний спосіб обробки даних, що відсутні. Він може дати розумний прогноз без налаштування гіперпараметрів. Це вирішує проблему

перенавчання у деревах рішень. У кожному дереві випадкового лісу випадково вибирається підмножина функцій у точці поділу вузла. Як працює алгоритм випадкового лісу. Розуміння дерев рішень: Дерева рішень – це будівельні блоки алгоритму випадкового лісу. Дерево рішень - це спосіб підтримки прийняття рішень, який формує деревоподібну структуру. Огляд дерев рішень допоможе зрозуміти, як працюють алгоритми випадкового лісу. Дерево рішень складається з трьох компонентів: вузлів рішень, листових вузлів та кореневого вузла. Алгоритм дерева рішень ділить навчальний набір даних на гілки, які розділяються на інші гілки. Ця послідовність триває доти, доки досягне листової вузол. Листовий вузол не може бути відокремлений. Вузли у дереві рішень представляють атрибути, які використовуються для прогнозування результату. Вузли ухвалення рішень надають посилання на листя. На наступній діаграмі показано три типи вузлів у дереві рішень.

Основна відмінність між алгоритмом дерева рішень та алгоритмом випадкового лісу полягає в тому, що в останньому встановлення кореневих вузлів та поділ вузлів виконується випадковим чином. Випадковий ліс використовує метод упакування для створення необхідного прогнозу[12].

Пакетування передбачає використання різних вибірок даних (навчальних даних), а не лише однієї вибірки. Набір навчальних даних включає спостереження та функції, які використовуються для прогнозування. Дерева рішень виробляють різні вихідні дані залежно від навчальних даних, що подаються алгоритму випадкового лісу. Ці результати будуть ранжовані, і в якості остаточного буде обрано найвищий результат.

2.7 Багат шаровий перцептрон (БШП), тобто нейронна мережа.

Багат шаровий перцептрон (MLP) – це доповнення до нейронної мережі з прямим зв'язком. Він складається з трьох типів шарів: вхідного, вихідного та прихованого, як показано на рис. 3.. Вхідний шар отримує вхідний сигнал для обробки. Необхідна задача, така як прогнозування та

класифікація, виконується вихідним шаром. Довільна кількість прихованих шарів, що поміщаються між вхідним та вихідним шарами, є справжнім обчислювальним механізмом MLP. Подібно до мережі з прямим зв'язком в MLP, дані передаються в прямому напрямку від вхідного до вихідного шару. Нейрони MLP навчаються за допомогою алгоритму навчання зворотного поширення. MLP призначені для апроксимації будь-якої безперервної функції та можуть вирішувати завдання, які не можна розділити лінійно. Основні варіанти використання MLP – це класифікація, розпізнавання, прогнозування та апроксимація образів. Багатошаровий перцептрон зображений на рис. 2.4.

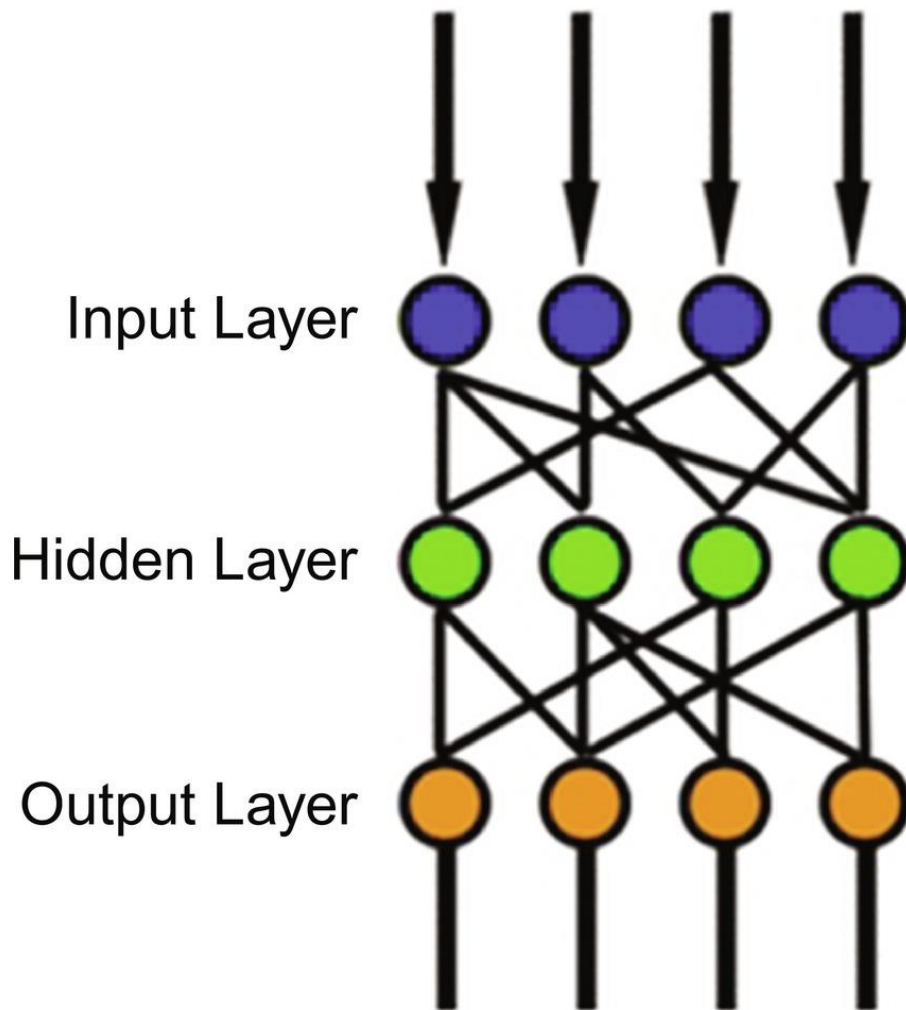


Рисунок 2.4 Багатошаровий перцептрон.

Багатошаровий перцептрон було розглянуто як забезпечення нелінійного відображення між вхідним вектором та відповідним вихідним вектором. Більшість роботи у цій галузі було присвячено отриманню цього нелінійного відображення у статичній обстановці. Багато практичних проблем можна змоделювати за допомогою статичних моделей, наприклад, розпізнавання символів. З іншого боку, багато практичних завдань, таких як прогнозування часових рядів, зір, мова та управління моторикою, потребують динамічного моделювання: поточний вихід залежить від попередніх входів та виходів.

2.8 Адаптивний бустинг (AdaBoost)

AdaBoost, також званий Adaptive Boosting, - це метод машинного навчання, що використовується як метод ансамблю. Найбільш поширений алгоритм, який використовується з AdaBoost, - це дерева рішень з одним рівнем, тобто дерева рішень з одним поділом. Ці дерева також називають пнями прийняття рішень. Що робить цей алгоритм, так це те, що він будує модель і надає рівні ваги всім точкам даних. Потім він надає більш високу вагу помилково класифікованим точкам. Тепер усім точкам з вищою вагою в наступній моделі надається більше значення. Тренувальні моделі будуть продовжуватися доти, доки не буде отримана низька помилка. Експерти говорять про AdaBoost як про одну з найбільш виважених комбінацій класифікаторів, що чутлива до шуму і сприяє певним результатам машинного навчання. Певна плутанина виникає через те, що AdaBoost може використовуватися з декількома екземплярами одного й того ж класифікатора з різними параметрами - де професіонали можуть говорити про AdaBoost, що «має лише один класифікатор», і не розуміти, як відбувається зважування.

AdaBoost – це тип алгоритму, який використовує підхід ансамблевого навчання для зважування різних вхідних даних. Він був розроблений Йоавом

Фройндом та Робертом Шапіром на початку 21 століття. В даний час він став своєрідним методом для різних видів підвищення в парадигмах машинного навчання[19].

AdaBoost також представляє особливу філософію машинного навчання - як інструмент ансамблевого навчання він виходить з фундаментальної ідеї, що багато слабких учнів можуть отримати кращі результати, ніж одна сильніша сутність, що навчається. Використовуючи AdaBoost, фахівці з машинного навчання часто створюють системи, які враховуватимуть ряд вхідних даних і комбінуватимуть їх для досягнення оптимального результату. Деякі розглядають цю ідею ще більшою мірою, говорячи про те, як AdaBoost може командувати «арміями пнів для прийняття рішень», які є істотно менш досвідченими учнями, які використовуються у великих кількостях для обробки даних, де цей підхід краще використання одного класифікатора.

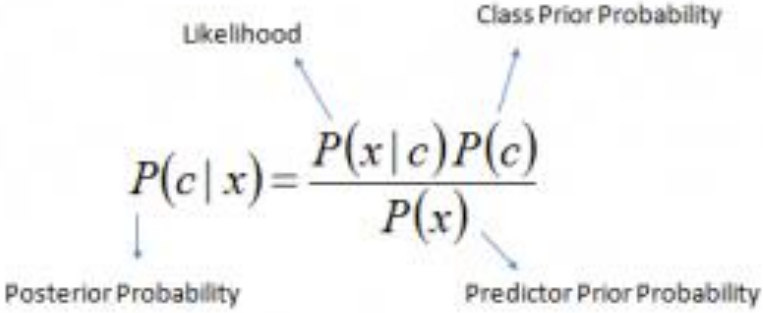
2.9 Наївний Байєсівський класифікатор

Це метод класифікації, заснований на теоремі Байєса із припущенням незалежності предикторів. Простіше кажучи, наївний класифікатор байєсовського передбачає, що наявність певної функції в класі не пов'язана з наявністю будь-якої іншої функції.

Наприклад, фрукт може вважатися яблуком, якщо він червоний, круглий та має діаметр близько 3 дюймів. Навіть якщо ці характеристики залежать один від одного або від наявності інших характеристик, всі ці властивості незалежно один від одного впливають на ймовірність того, що цей фрукт є яблуком, і тому він відомий як наївний.

Наївну байєсовську модель легко побудувати, і вона особливо корисна для великих наборів даних. Відомо, що наївний байєсовський метод не лише простий, а й перевершує навіть найскладніші методи класифікації.

Теорема Байєса надає спосіб обчислення апостеріорної ймовірності $P(c|x)$ з $P(c)$, $P(x)$ та $P(x|c)$. Подивіться на рівняння нижче яке предсталене на рис. 2.5.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Рисунок 2.5. Наївний Байєсівський класифікатор

Вище, $P(c|x)$ - апостеріорна ймовірність класу (c , target) для даного предиктора (x , атрибуту). $P(c)$ - апріорна ймовірність класу. $P(x|c)$ - це ймовірність, яка є ймовірністю предиктора даного класу. $P(x)$ - апріорна ймовірність провісника.

Плюси: легко та швидко передбачити клас набору тестових даних. Він також добре працює у прогнозуванні кількох класів. Коли виконується припущення про незалежність, наївний класифікатор Байєса працює краще в порівнянні з іншими моделями, такими як логістична регресія, і вам потрібно менше даних для навчання.

Він добре працює у разі категоріальних вхідних змінних у порівнянні з числовими змінними. Для числової змінної передбачається нормальний розподіл (крива дзвони, що є сильним припущенням).

Мінуси: якщо категоріальна змінна має категорію (у наборі тестових даних), яка не спостерігалася в наборі навчальних даних, то модель надасть можливість 0 (нуль) і не зможе зробити прогноз. Це часто називають

"нульовою частотою". Для вирішення цієї проблеми ми можемо використовувати техніку згладжування. Один із найпростіших методів згладжування називається оцінкою Лапласа.

З іншого боку, наївний байєсовський метод також відомий як поганий оцінювач, тому до ймовірносних вихідних даних `pred_proba` не варто ставитися надто серйозно.

Ще одне обмеження наївного Байєса – це припущення про незалежних провісників. У реальному житті практично неможливо отримати набір повністю незалежних предикторів.

4 застосування наївних байєсівських алгоритмів:

1.Прогнозування в реальному часі: Наївний клас класифікатора Байєса швидко навчається. Таким чином, його можна використовувати для прогнозування у реальному часі.

2.Прогнозування кількох класів: цей алгоритм також добре відомий функцією прогнозування кількох класів. Тут ми можемо передбачити можливість кількох класів цільової змінної.

3.Класифікація тексту / Фільтрація спаму / Аналіз тональності : Наївні байєсівські класифікатори, які в основному використовуються при класифікації тексту (через найкращий результат у багатокласових задачах та правила незалежності), мають більш високий відсоток успіху в порівнянні з іншими алгоритмами. В результаті він широко використовується у фільтрації спаму (визначення спаму в електронній пошті) та аналізі настроїв (в аналізі соціальних мереж для виявлення позитивних та негативних настроїв клієнтів).

Система рекомендацій: наївний класифікатор Байєса і спільна фільтрація разом створюють систему рекомендацій, яка використовує методи машинного навчання та інтелектуального аналізу даних для фільтрації невидимої інформації та прогнозування того, чи хоче користувач даний ресурс чи ні.

Як побудувати базову модель, використовуючи наївний байєсовський метод у Python та R? Знову ж таки, scikit learn (бібліотека python) допоможе тут побудувати наївну байєсівську модель на Python. У бібліотеці scikit-learn є три типи наївної байєсівської моделі:

Гаусівський процес використовують у класифікації і передбачає, що функції підпорядковуються нормальному розподілу.

Поліноміальний: використовується для дискретних підрахунків. Наприклад, припустимо, ми маємо проблему класифікації тексту. Тут ми можемо розглянути випробування Бернуллі, які йдуть на один крок далі, і замість "слова, що зустрічається в документі", у нас є "підрахувати, як часто слово зустрічається в документі", ви можете думати про це як "кількість разів, коли спостерігається" номер результату x з n судових процесів».

Бернуллі: Біноміальна модель корисна, якщо ваші вектори ознак є двійковими (тобто з нулями та одиницями). Одним із додатків може бути класифікація тексту з моделлю «мішок слів», де одиниці та нулі означають «слово зустрічається в документі» та «слово не зустрічається в документі» відповідно.

2.10 Квадратичний дискримінантний аналіз (КДА)

Дискримінантний аналіз - це метод, який використовується дослідником для аналізу даних дослідження, коли критерій чи залежна змінна є категоріальною, а предиктор чи незалежна змінна є інтервальною за своєю природою. Термін "категоріальна змінна" означає, що залежна змінна ділиться на кілька категорій. Наприклад, три марки комп'ютерів: комп'ютер А, комп'ютер В та комп'ютер С можуть бути категоріальною залежною змінною.

Метою дискримінантного аналізу є розробка дискримінантних функцій, які являють собою не що інше, як лінійну комбінацію незалежних змінних, яка ідеально розрізнятиме категорії залежної змінної. Це дозволяє

досліднику вивчити, чи існують суттєві відмінності між групами з погляду змінних-предикторів. Він також оцінює точність класифікації.

Дискримінантний аналіз описується кількістю категорій, якими має залежна змінна.

Як і в статистиці, все передбачається нескінченно, тому в цьому випадку, коли залежна змінна має дві категорії, використовується двогруповий дискримінантний аналіз. Якщо залежна змінна має три або більше трьох категорій, використовується множинний дискримінантний аналіз. Основна відмінність між типами дискримінантного аналізу полягає в тому, що для двох груп можна отримати лише одну дискримінантну функцію. З іншого боку, у разі множинного дискримінантного аналізу можна обчислити більше однієї дискримінантної функції.

Є багато прикладів, які можуть пояснити коли підходить дискримінантний аналіз. Його можна використовувати, щоб дізнатися, чи різняться важкі, середні та легкі споживачі безалкогольних напоїв з погляду споживання заморожених продуктів. У галузі психології його можна використовувати для розрізнення чутливих до ціни та нечутливих до ціни покупців продуктів з погляду їх психологічних властивостей чи характеристик. У сфері бізнесу його можна використовувати для розуміння характеристик або атрибутів покупця, який має лояльність до магазину, та покупця, який не має лояльності до магазину.

Для дослідника важливо розуміти взаємозв'язок дискримінантного аналізу з регресією та дисперсійним аналізом (ANOVA), який має багато подібностей та відмінностей. Часто ми можемо знайти подібність і розходження з людьми, з якими стикаємося. Так само є деякі подібності та відмінності з дискримінантним аналізом та двома іншими процедурами. Подібність полягає в тому, що кількість залежних змінних дорівнює одній в дискримінантному аналізі, а в двох інших процедур кількість незалежних змінних кратно в дискримінантному аналізі. У дискримінантному аналізі різниця буває категоріальною чи двійковою, а в двох інших процедурах –

метричною. Природа незалежних змінних категоріальна в дисперсійному аналізі (ANOVA), але є метрикою в регресійному та дискримінантному аналізі.

Під час проведення дискримінантного аналізу необхідно виконати такі кроки:

- Завдання формулюється перед проведенням.
- Оцінюються коефіцієнти дискримінантної функції.
- Наступним кроком є визначення значущості цих дискримінантних функцій.
- Необхідно інтерпретувати отримані результати.
- Останній та найважливіший крок – оцінити достовірність.

2.11 Методи селекції ознак

2.11.1 Прямий відбір

Прямий вибір — це тип ступінчастої регресії, яка починається з порожньої моделі та додає змінні одну за одною. На кожному наступному кроці ви додаєте одну змінну, яка забезпечує найкраще покращення вашої моделі.

Це один із двох широко використовуваних методів ступінчастої регресії; інший – зворотне усунення і майже протилежний. При цьому ви починаєте з моделі, яка включає всі можливі змінні та усуває сторонні змінні одну за одною[20].

Прямий вибір зазвичай починається лише з перехоплення. Перевіряються різні змінні, які можуть бути релевантними, і «найкраща» змінна, де «найкраща» визначається деякими заздалегідь визначеними критеріями, додається до моделі.

Оскільки модель продовжує вдосконалюватися (за тими самими критеріями), ми продовжуємо процес, додаючи по одній змінній за раз і тестуючи на кожному кроці. Як тільки модель більше не покращується з

додаванням додаткових змінних, процес зупиняється. Приклад прямого відбору зображений на рис. 2.6.

Forward stepwise selection example with 5 variables:

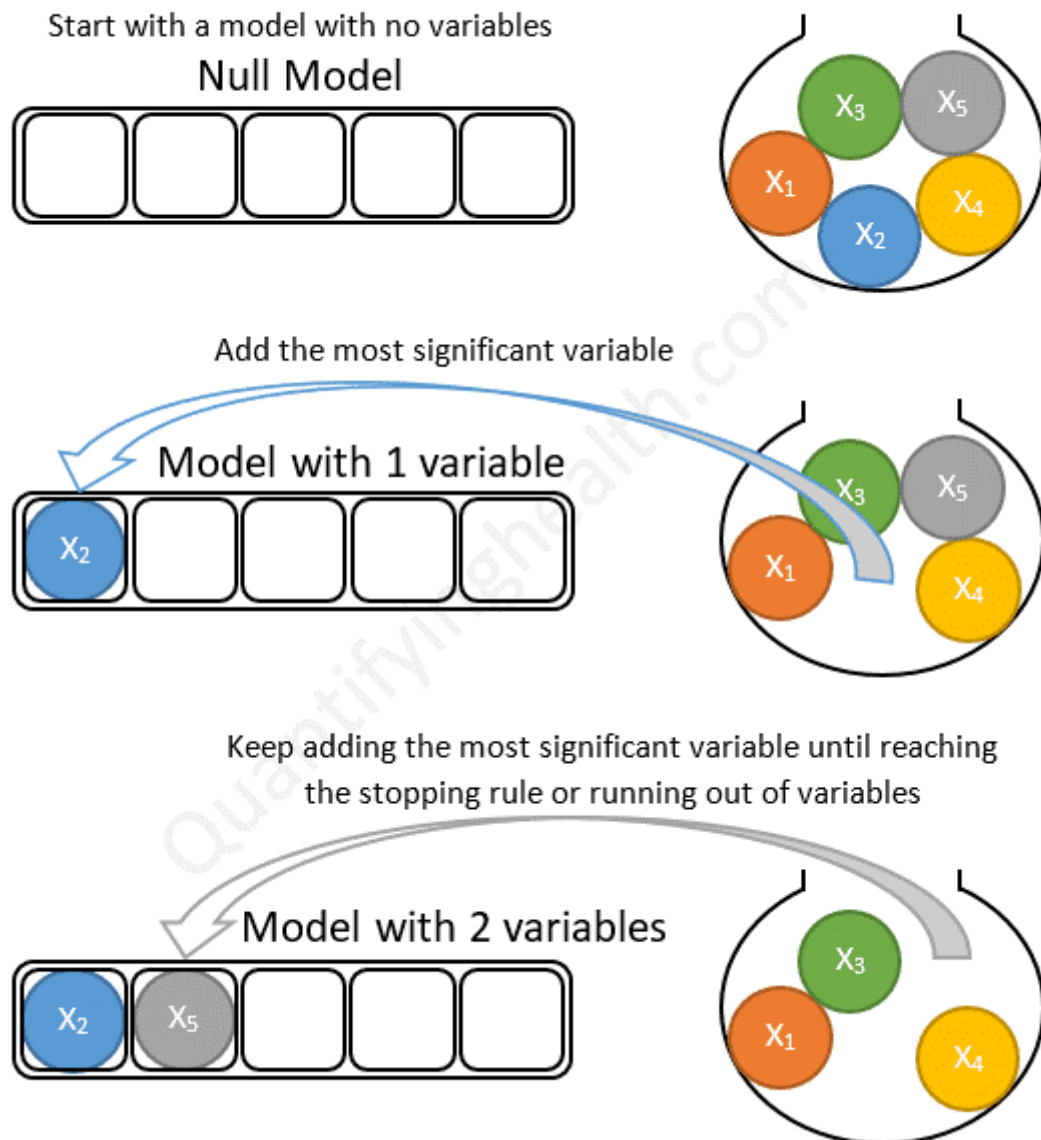


Рисунок 2.6 Приклад прямого відбору

Критерій, що використовується для визначення того, яка змінна входить, коли вони змінюються. Можливо, ви намагаєтеся знайти найнижчий бал при перехресній перевірці, найнижче значення p або будь-який з ряду інших тестів чи показників точності.

Оскільки ступінчаста регресія має тенденцію до надмірного підгонки, що відбувається, коли ми вводимо більше змінних, ніж насправді добре для моделі; вона зазвичай показує дуже близьке, акуратне підходження даних, що використовуються в регресії, але модель буде далека від додаткових точок даних і не підходить для інтерполяції. Тому зазвичай добре мати суворі критерії для додавання будь-яких змінних.

2.11.2 Зворотне усунення

Поступовий відбір назад (або зворотне усунення) — це метод змінного відбору, який: Починається з моделі, яка містить усі розглянуті змінні (називається повною моделлю). Потім починає видаляти найменш значущі змінні одну за одною. Поки не буде досягнуто заздалегідь визначене правило зупинки або поки в моделі не залишиться жодної змінної. Ось приклад зворотного вилучення з 5 змінними який представлений на рис. 2.7.

Backward stepwise selection example with 5 variables:

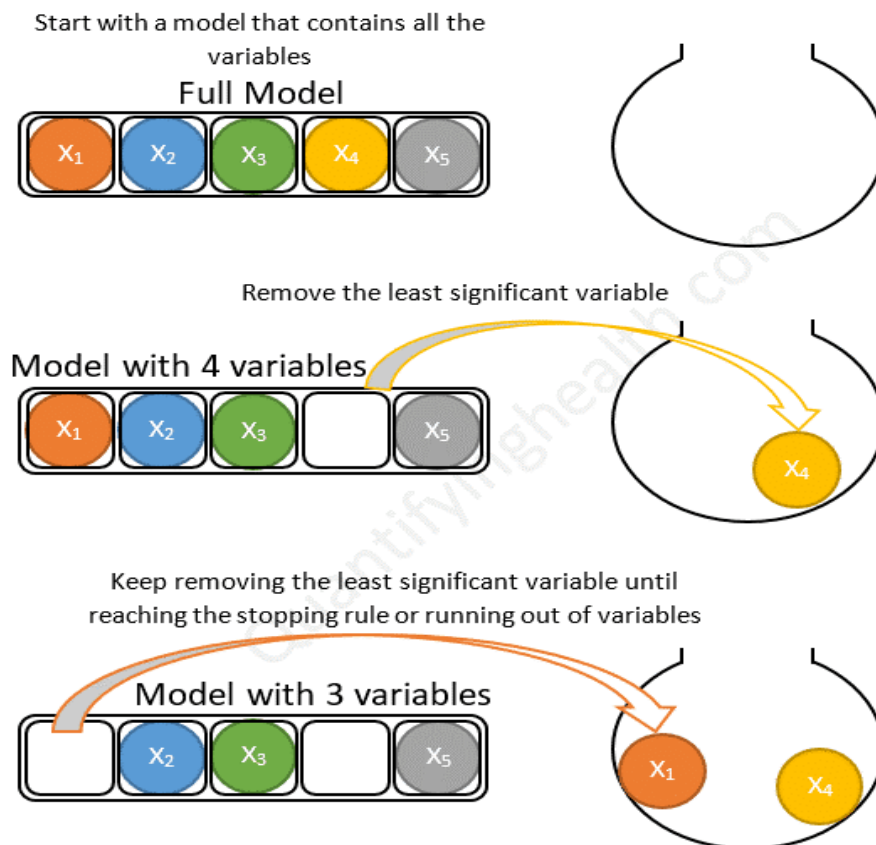


Рисунок 2.7 Приклад зворотного вилучення з 5 змінними

Як ми зробили з прямим вибором, щоб зрозуміти, як працює зворотне видалення, нам потрібно буде обговорити, як визначити: Найменш значуща змінна на кожному кроці. Правило зупинки. Визначте найменшу змінну для видалення на кожному кроці

Найменш значимою змінною є змінна, яка: Має найвище p -значення в моделі, або його виключення з моделі викликає найменше падіння R^2 , або його виключення з моделі спричиняє найменше збільшення RSS (Residuals Sum of Squares) порівняно з іншими предикторами.

Виберіть правило зупинки. Правило зупинки виконується, коли всі змінні, що залишилися в моделі, мають значення p , менше за деякий заздалегідь заданий поріг. Коли ми досягнемо цього стану, зворотне видалення завершиться і поверне модель поточного кроку. Як визначити поріг? Як і для прямого вибору, порогове значення може бути: Фіксована величина (наприклад: 0,05 або 0,2 або 0,5). Визначається AIC (інформаційний критерій Акаїке). Визначається BIC (інформаційний критерій Байєса).

2.11.3 Рекурсивне видалення ознак

Рекурсивне видалення ознак, або скорочено RFE, є популярним алгоритмом вибору ознак.

RFE користується популярністю, оскільки його легко налаштувати та використовувати, а також тому, що він ефективний для вибору тих функцій (стовпців) у наборі навчальних даних, які є більш або найбільш релевантними для прогнозування цільової змінної.

Під час використання RFE є два важливі параметри конфігурації: вибір кількості функцій для вибору та вибір алгоритму, який використовується для вибору функцій. Обидва ці гіперпараметри можна дослідити, хоча продуктивність методу не сильно залежить від того, чи правильно налаштовані ці гіперпараметри.

У цьому підручнику ви дізнаєтеся, як використовувати рекурсивне видалення ознак (RFE) для вибору функцій у Python.

RFE — це ефективний підхід для видалення функцій із навчального набору даних для вибору функцій.

Рекурсивне видалення ознак, або скорочено RFE, є популярним алгоритмом вибору ознак.

RFE користується популярністю, оскільки його легко налаштувати та використовувати, а також тому, що він ефективний для вибору тих функцій (стовпців) у наборі навчальних даних, які є більш або найбільш релевантними для прогнозування цільової змінної.

Під час використання RFE є два важливі параметри конфігурації: вибір кількості функцій для вибору та вибір алгоритму, який використовується для вибору функцій. Обидва ці гіперпараметри можна дослідити, хоча продуктивність методу не сильно залежить від того, чи правильно налаштовані ці гіперпараметри.

У цьому підручнику ви дізнаєтеся, як використовувати рекурсивне видалення ознак (RFE) для вибору функцій у Python.

Набір даних машинного навчання для класифікації або регресії складається з рядків і стовпців, як електронна таблиця Excel. Рядки часто називають зразками, а стовпці — ознаками, напр. особливості спостереження в проблемній області.

Вибір ознак відноситься до методів, які вибирають підмножину найбільш релевантних ознак (стовпців) для набору даних. Менша кількість функцій може дозволити алгоритмам машинного навчання працювати ефективніше (менша складність простору або часу) і бути ефективнішими. Деякі алгоритми машинного навчання можуть бути введені в оману невідповідними функціями введення, що призведе до погіршення прогнозової продуктивності.

RFE — це алгоритм вибору функцій обгортки. Це означає, що інший алгоритм машинного навчання надається і використовується в ядрі методу, обгортається RFE і використовується для вибору функцій. Це на відміну від

вибору об'єктів на основі фільтрів, які оцінюють кожну характеристику та вибирають об'єкти з найбільшою (або найменшою) оцінкою.

Технічно, RFE — це алгоритм вибору функцій у стилі обгортки, який також використовує внутрішній вибір на основі фільтрів.

RFE працює шляхом пошуку підмножини функцій, починаючи з усіх функцій у наборі навчальних даних і успішно видаляючи функції, доки не залишиться бажане число.

Це досягається шляхом підгонки даного алгоритму машинного навчання, який використовується в ядрі моделі, ранжирування функцій за важливістю, відкидання найменш важливих ознак і повторного налаштування моделі. Цей процес повторюється до тих пір, поки не залишиться певна кількість функцій.

2.11.4 Регуляризація

Іноді наша модель машинного навчання добре працює з навчальними даними, але погано працює з невидимими або тестовими даними. Це означає, що модель не може передбачити вихідний або цільовий стовпець для невидимих даних, вводячи шум у вихідні дані, і, отже, модель називається переобладнаною моделлю. Давайте коротко розберемося в значенні «Шум»:

Під шумом ми маємо на увазі ті точки даних у наборі даних, які насправді не представляють справжні властивості ваших даних, а лише через випадковий випадок.

Отже, для вирішення проблеми переобладнання ми використовуємо методику регуляризації.

Регуляризація - це одна з найважливіших концепцій машинного навчання. Ця техніка запобігає переобладнанню моделі, додаючи до неї додаткову інформацію.

Це форма регресії, яка зменшує оцінки коефіцієнта до нуля. Іншими словами, ця методика змушує нас не вивчати більш складну чи гнучку модель, щоб уникнути проблеми переобладнання.

Тепер давайте розберемося в «Як представлена гнучкість моделі?»

Для задач регресії збільшення гнучкості моделі представлено збільшенням її коефіцієнтів, які розраховуються з лінії регресії.

Простіше кажучи: «У техніці регуляризації ми зменшуємо величину незалежних змінних, зберігаючи однакову кількість змінних». Він підтримує точність, а також узагальнення моделі.

Як працює регуляризація? Регуляризація працює шляхом додавання штрафу або терміну складності або терміну стиснення із залишковою сумою квадратів (RSS) до комплексної моделі. Давайте розглянемо просте рівняння лінійної регресії:

Тут Y представляє залежну ознаку або відповідь, яка є вивченим відношенням. Тоді, Y наближено до $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$. Тут $X_1, X_2, \dots, X_p$ є незалежними ознаками або предикторами для Y і $\beta_0, \beta_1, \dots, \beta_p$ представляє оцінки коефіцієнтів для різних змінних або предикторів (X), що описує ваги або величину, пов'язану з ознаками, відповідно.

У простій лінійній регресії наша функція оптимізації або функція втрат відома як залишкова сума квадратів (RSS).

Ми вибираємо такий набір коефіцієнтів, щоб мінімізувати наступну функцію втрат. 1Функція вартості для простої лінійної регресії зображена рис. 2.8.

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 .$$

Рисунок 2.8 Функція вартості для простої лінійної регресії

Тепер це коригує оцінки коефіцієнтів на основі даних навчання. Якщо в даних навчання присутній шум, то оцінені коефіцієнти не будуть добре узагальнюватися і не можуть передбачити майбутні дані.

Ось тут і виникає регуляризація, яка зменшує або регулює ці засвоєні оцінки до нуля, додаючи функцію втрат з оптимізуючими параметрами, щоб створити модель, яка може передбачити точне значення Y .

Висновки до розділу 2

Було розглянуто різні алгоритми класифікації. Також були розглянуті методи відбору ознак. Класифікація – це процес розпізнавання, розуміння та групування ідей та об'єктів у попередньо встановлені категорії або «підгрупи». Використовуючи попередньо категоризовані набори навчальних даних, програми машинного навчання використовують різноманітні алгоритми для класифікації майбутніх наборів даних за категоріями.

РОЗДІЛ 3

ПОСТАНОВКА ЗАДАЧІ

Відбір (обирання) ознак (англ. Feature Selection) – це процес відбору підмножини вагомих незалежних (вхідних) ознак для їх використання в моделюванні. Слід відрізнити це від «виділення ознак», яке створює нові ознаки, як функції від оригінальних. Дана процедура необхідна для того, щоб не виникало проблем прокляття розмірності» [11] та мультиколінеарності [12]. Розв’язання подібних проблем дозволить отримати більш якісні та адекватні моделі прогнозування стану пацієнта за ультразвуковими знімками, що у свою чергу забезпечить підтримку лікарям під час прийняття рішень.

3.1 Категорії алгоритмів відбору ознак

Існують три категорії алгоритмів відбору ознак:

- «Методи фільтрів (рис. 3.1), які використовують непрямий показник замість показника помилки для оцінки підмножини ознак, і вибір підмножини проводять незалежно від моделювання. Прикладами таких методів є: нескінченний відбір ознак [13], t-тест Велча [14] та власний вектор центральності відбору ознак [15].

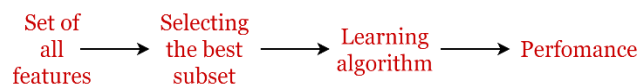


Рисунок 3.1 Принцип методів фільтру

- Обгорткові методи (рис. 3.2), які використовують модель апріорної оцінки результату для оцінки підмножин ознак, і, на відміну від підходів фільтрації, дозволяють знайти взаємозв’язок між змінними. Для подібного підходу можна використовувати різні еволюційні алгоритми, такі як: генетичний алгоритм [16], мурашиний алгоритм, метод рою часток [17] і так

далі. Серед самостійних алгоритмів існують кореляційний відбір ознак [18] та одновимірний відбір ознак.

- Вбудовані методи (рис.3.3), які є узагальнюючою групою технік, що здійснюють відбір ознак як частина процесу побудови моделі, були запропоновані як спроба комбінації переваг двох попередніх видів методів. Відомими алгоритмами з подібним підходом є метод LASSO [19] та рекурсивне виключення ознак.

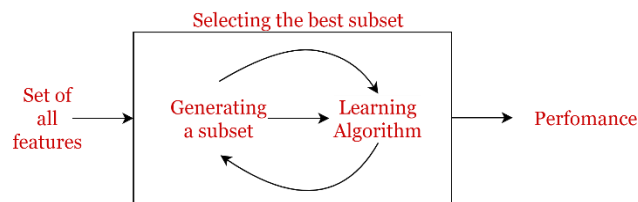


Рисунок 3.2. Принцип обгорткових методів

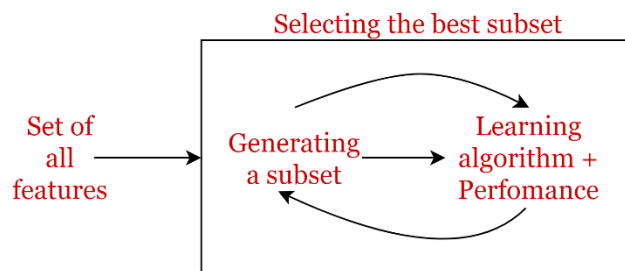


Рисунок 3.3 Принцип вбудованих методів

Із цих категорій для дослідження найбільше підходять обгорткові методи, оскільки вони не залежать від алгоритму моделювання, та дозволяють отримати найкращу підмножину із заданого k ознак.

3.2 Постановка задачі

Основною задачею даної праці є виконання оптимальної селекції ознак для успішної реалізації загальної задачі класифікації патології печінки за знімками ультразвукового дослідження (УЗД). Подібна задача виникла під час роботи кафедри біомедичної кібернетики (БМК) КПІ над проектом,

основна мета якого є розробка системи підтримки прийняття рішень під час діагностики печінки, замовником якої виступає ДУ «Інститут ядерної медицини та променевої діагностики НАМН України». Спеціалістами саме цієї державної установи були надані ультразвукові знімки для проведення досліджень.

Основна задача проекту полягає у бінарній класифікації «Норма:Патологія» областей інтересу (які були помічені спеціалістами державної установи) зображень УЗД печінки. Спеціалістами УЗД було взято трьома різними датчиками:

1. Конвексним(рис. 3.4).
2. Лінійним у стандартному режимі(рис. 3.5).
3. Лінійним у посиленому режимі(рис. 3.6).



Рисунок 3.4 Зображення УЗД печінки в нормі(конвексний датчик)[25].

Конвексний датчик характерний тим, що його верхівка вигнута.

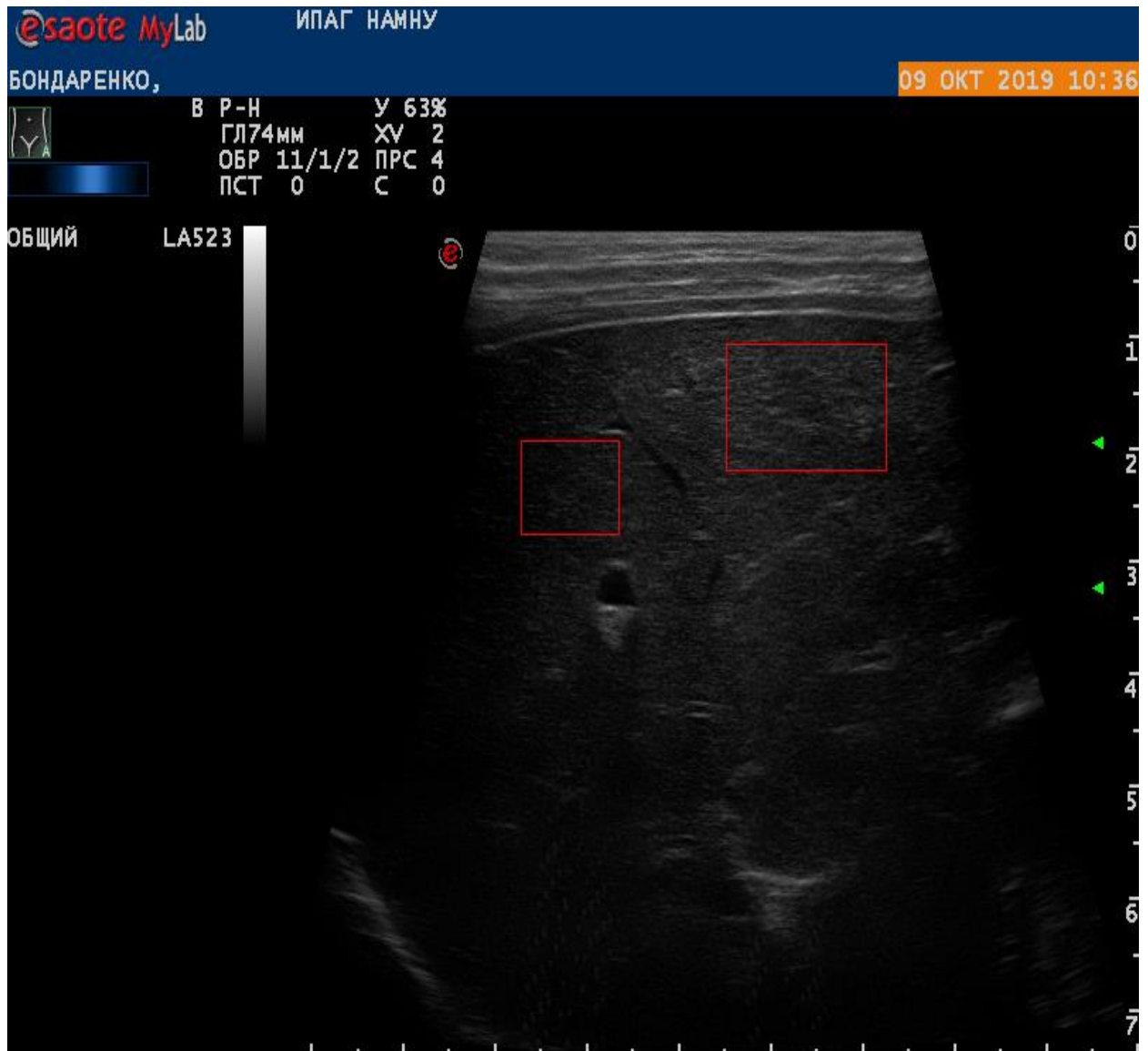


Рисунок 3.5 Зображення УЗД печінки в нормі(лінійним у стандартному режимі)

«Лінійний датчик характерний тим, що його верхівка пряма.

Посилений лінійний датчик:

Тут верхівка теж пряма, але ще даний датчик характерний тим, що його коефіцієнт масштабування більше (тобто він більш наближений до досліджуваної області печінки) » [25].

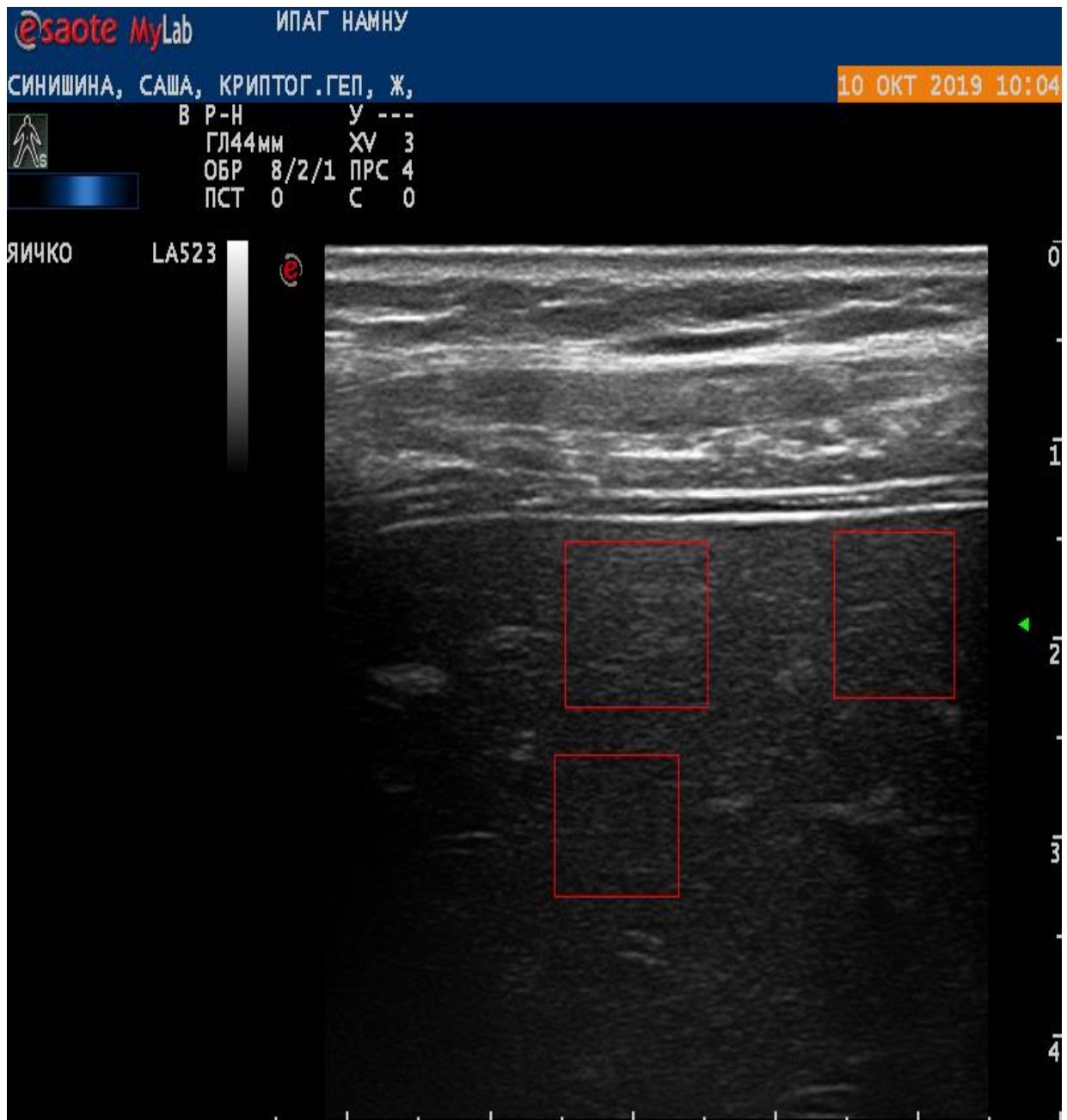


Рисунок 3.6 Зображення УЗД печінки в нормі(лінійним у посиленому режимі)

«На них чітко видно червоний прямокутник, який є областю інтересу (ОІ). Її виділяють лікарі (на їх думку дана область є найбільш інформативною в визначенні діагнозу). Таких ОІ на зображенні може бути нескінченна кількість. »[25].

Відповідно до цього були сформовані 3 різні підвибірки даних зображень, кількість областей інтересу по кожній із яких вказана в таблиці 3.1.

Таблиця 3.1

Опис загальної вибірки даних зображень

Датчик	Назва підвибірки	Кількість областей інтересу (Норма:Патологія)
Конвексний	КД	304 (197:107)
Лінійний у стандартному режимі	ЛСД	154 (80:74)
Лінійний у посиленому режимі	ЛПД	124 (35:89)

Порахувавши для кожної із підвбірок текстурні ознаки авторів кафедри БМК [1-10], отримано наступне: 352 ознаки для класифікації зображень конвексного датчику; 354 ознаки для класифікації зображень лінійного датчику (стандартний режим); 342 ознаки для класифікації зображень лінійного датчику (посилений режим). Із цього випливає тема дослідження, а власне знаходження оптимального варіанту селекції ознак, щоб можна було ефективно виконати задачу бінарної класифікації «Норма:Патологія».

Як видно із таблиці 3.1, у вибірках конвексного і посиленого лінійного датчиків присутня незбалансованість класів. Через це побудований класифікатор може виявитись абсолютно неефективним. Слід відмітити, що можуть відрізнятись і приклади помилкової класифікації. Як правило, невірна класифікація міноритарного класу (зазвичай патології) обходиться дорожче, ніж помилкова класифікація мажоритарного класу. Одним із підходів для вирішення даної проблеми є використання різних стратегій семплінгу.

3.3. Метод семплінгу

Найпопулярнішим методом семплінгу є алгоритм SMOTE [20] (Synthetic Minority Over-sampling Technique – техніка синтетичного оверсемплінгу меншості). Дана стратегія заснована на ідеї генерації певної кількості штучних прикладів, які були б «схожі» на вже існуючі в міноритарному класі, але при цьому не дублювали б їх. Для створення нового запису знаходять різницю $d = X_b - X_a$, де X_a, X_b – вектори ознак сусідніх прикладів a і b із міноритарного класу. Їх знаходять за допомогою алгоритму найближчого сусіду (KNN). У даному випадку необхідно і достатньо для прикладу b отримати набір із k сусідів, із якого у подальшому буде обрана запис b . Інші кроки алгоритму KNN не потрібні.

Після цього із d шляхом множення кожного його елемента на випадкове число в інтервалі $(0, 1)$ отримують d' . Вектор ознак нового прикладу обчислюється шляхом сумування X_a і d' . Алгоритм SMOTE дозволяє задавати кількість записів, яке необхідно штучно згенерувати. Ступінь подібності прикладів a і b можна регулювати шляхом зміни значення k (числа найближчих сусідів). На рис. 3.8 схематично зображено те, як у двовимірному просторі ознак можуть розташовуватись штучно згенеровані приклади.

Даний підхід має недолік у тому, що «всліпу» збільшує щільність прикладами в області слабо представленого класу (рис. 3.7). У випадку, коли міноритарні приклади рівномірно розподілені серед мажоритарних, і мають низьку щільність, алгоритм SMOTE тільки сильніше перемішує класи.

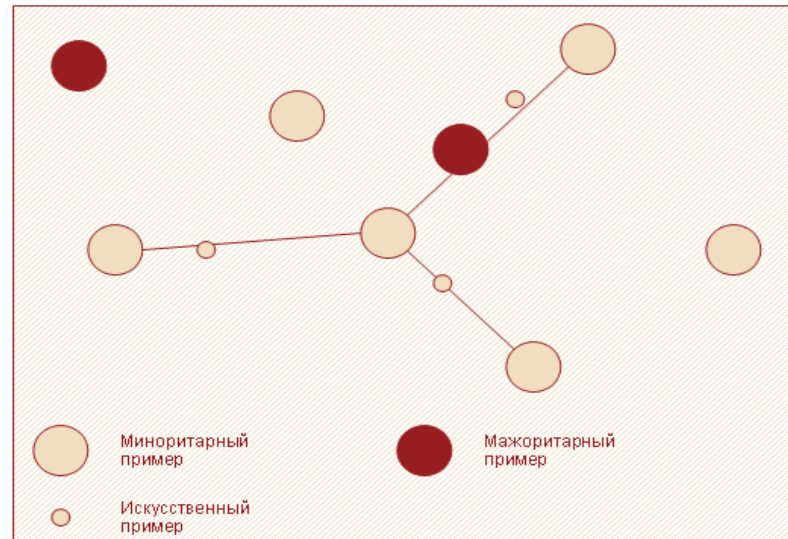


Рисунок 3.7. Штучно створені нові приклади міноритарного класу

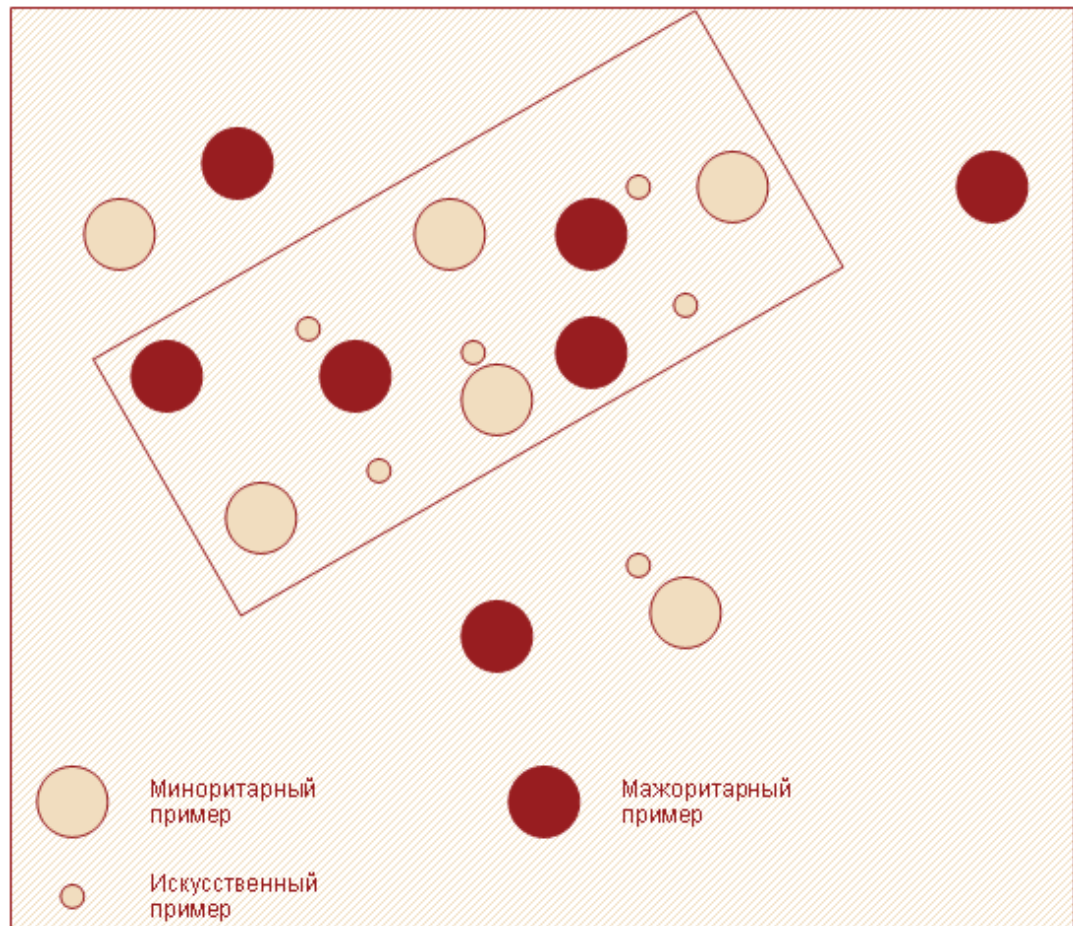


Рисунок 3.8 Негативний вплив алгоритму SMOTE

Як рішення даної проблеми був запропонований алгоритм адаптивного штучного збільшення числа прикладів міноритарного класу ASMO (Adaptive

Synthetic Minority Oversampling).

Алгоритм ASMO наступний:

1. Якщо для кожного i -го прикладу міноритарного класу із k найближчих сусідів g ($g \leq k$) належить до мажоритарного, то набір даних вважається «розсіяним». У цьому випадку використовують ASMO, інакше – SMOTE (як правило, g задають рівним 20).

2. Використовуючи тільки приклади міноритарного класу, виділити декілька кластерів (наприклад, алгоритмом k -means).

3. Згенерувати штучні записи в межах окремих кластерів на основі усіх класів. Для кожного прикладу міноритарного класу знаходять m найближчих сусідів, і на основі них (також як і в SMOTE) створюються нові записи.

Така модифікація алгоритму SMOTE робить його більш адаптивним до різних наборів даних із незбалансованими класами. Загальне представлення ідеї даної стратегії показане на рис. 3.9.

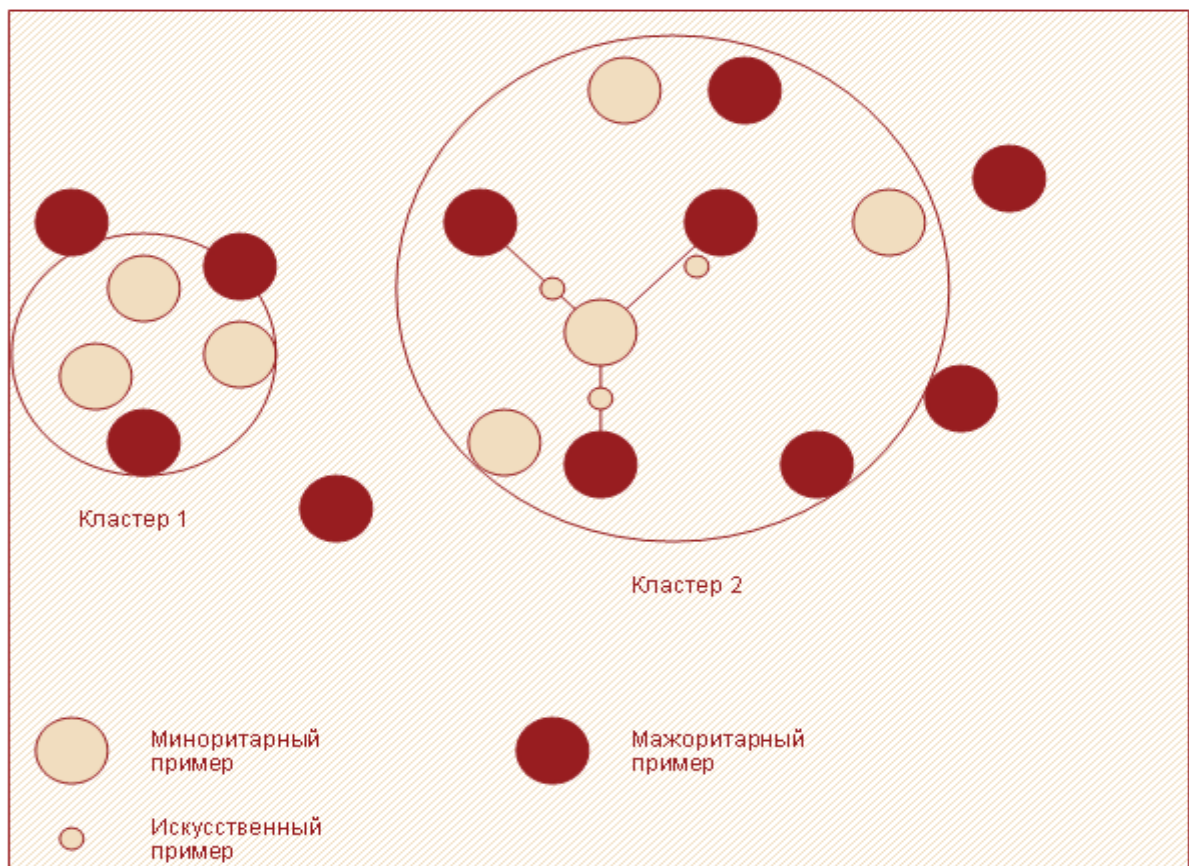


Рисунок 3.9. Основна ідея алгоритму ASMO

Висновки до розділу 3

Відбір (обирання) ознак (англ. Feature Selection) – це процес відбору підмножини вагомих незалежних (вхідних) ознак для їх використання в моделюванні. Слід відрізняти це від «виділення ознак», яке створює нові ознаки, як функції від оригінальних.

РОЗДІЛ 4

АНАЛІЗ ЕФЕКТИВНОСТІ КРИТЕРІЇВ ВІДБОРУ

Для знаходження оптимального варіанту селекції ознак були запропоновано наступні 5 критеріїв:

1. Внутрішньокласова дисперсія (треба знайти мінімум).

Розраховується за формулою (рис. 4.1):

$$\frac{1}{K} \sum_{q=1}^K \frac{1}{n_q} \sum_{i=1}^{n_q} \sum_{p=1}^m \left(x_{ip} - \bar{x}_p^q \right)^2$$

Рисунок 4.1 Внутрішньокласова дисперсія

де: x_{ip} – це p -та координата i -ой точки (їх усього n_q штук) в q -ом кластері (їх усього K штук), $i = 1, \dots, n_q$, $p = 1, \dots, m$.

2. Міжкласова дисперсія (треба знайти максимум). Розраховується за формулою (рис. 4.2):

$$\frac{1}{C_K^2} \sum_{i=1}^K \sum_{j=i+1}^K \sum_{p=1}^m \left(\bar{x}_p^i - \bar{x}_p^j \right)^2$$

Рисунок 4.2 Міжкласова дисперсія

де: x_{ip} – це p -та координата i -ой точки (їх усього n_q штук) в q -ом кластері (їх усього K штук), $i = 1, \dots, n_q$, $p = 1, \dots, m$.

3. Відношення внутрішньокласової і міжкласової дисперсій (треба знайти мінімум). Розраховується за формулою (рис. 4.3):

$$\min_{K, x_i \in X_j} \left(\frac{\frac{1}{K} \sum_{q=1}^K \frac{1}{n_q} \sum_{i=1}^{n_q} \sum_{p=1}^m (x_{ip} - \bar{x}_p^q)^2}{\frac{1}{C_K^2} \sum_{i=1}^K \sum_{j=i+1}^K \sum_{p=1}^m (\bar{x}_p^i - \bar{x}_p^j)^2} \right)$$

Рисунок 4.3 Відношення внутрішньокласової і міжкласової дисперсій

4. Кореляційний відбір ознак (CFS – Correlation Feature Selection). Треба знайти максимум.

Кореляційний відбір ознак [18] оцінює підмножину незалежних ознак на базі наступної гіпотези: «Хороші підмножини ознак мають ознаки, які сильно корелюють з залежною змінною, але не корелюють один із одним». Дана гіпотеза дозволяє вирішити проблему мультиколінеарності. Ідея полягає в знаходженні такої підмножини із k незалежних ознак, яка дає максимальне значення критерію оцінки підмножини ознак S . Критерій визначається наступною формулою (рис. 4.4):

$$S_k = \frac{\overline{kr_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Рисунок 4.4 Кореляційний відбір ознак

де: $\overline{r_{cf}}$ – середнє значення модулів кореляцій усіх незалежних ознак з залежною змінною, $\overline{r_{ff}}$ – середнє значення модулів кореляцій усіх ознак між собою, k – кількість ознак в підмножині.

Для обрахування даного критерію буде використовуватись кореляція Спірмана, оскільки розподіл незалежних ознак не є нормальним (ознаки не розподілені за функцією Гауса), а залежна змінна є бінарною.

5. Критерій якості ансамблю ознак (QCFE – Quality Criterion of Features' Ensemble).

Пропонований авторами кафедри БМК комбінований критерій, об'єднує у собі 2 показники – зв'язок ансамблю ознак з класом, та незалежність ознак в ансамблі між собою (рис. 4.5).

$$QCFE = \frac{\alpha}{2k} \sum_{p=1}^k |K_{y, x_{ip}}| + \frac{(1-\alpha)}{2 + \frac{1}{C_k^2} \sum_{p=1}^k \sum_{q=1}^p |K_{x_{ip}, x_{iq}}|}$$

Рисунок 4.5 Критерій якості ансамблю ознак

де: K – коефіцієнт кореляції, y – клас об'єкту класифікації, x_i – ознака об'єкту, k – розмір ансамблю ознак, α – вага залежності ознак, C – кількість комбінацій (із комбінаторики).

Слід зазначити, що перед тим як знаходити оптимум даного критерію, кожна ознака зводиться до бінарних чисел (0, 1) згідно їхніх оптимальних порогових функцій. Саме тому для обрахування даного критерію буде використовуватись кореляція Кендала.

Щоб вияснити, який із перелічених критеріїв оцінки ансамблю ознак є оптимальний для отримання найкращих моделей моделювання класифікації патології печінки за знімками УЗД, був запропонований наступний алгоритм дій:

а) Задається бажана кількість ознак k в ансамблі. Мінімальною кількістю буде 5, а максимальною – 20. Дані числа були обрані з тою логікою, щоб моделі вийшли з оптимальною складністю, та не були перенавчені.

б) Для кожного k від 5 до 20 за допомогою генетичного алгоритму знаходиться ансамбль, який дає екстремум критерію із пп. 1-5 (для кожного критерію окремо).

с) Після знаходження оптимальних ансамблів для кожного k , за ними розраховуються моделі класифікації патології печінки. Щоб отримати об'єктивні результати, будуть використовуватись 10 різних алгоритмів класифікації на навчальних (60%), валідаційних (20%) і тестових (20%) вибірках. На тестових вибірках розраховуються F1-score кожної моделі класифікації, після чого береться середнє арифметичне для кожного k . Таким чином для кожного критерію обирається той k , де середнє F1-score усіх моделей є найбільшим.

Обраними алгоритмами класифікації стали:

1. k -найближчих сусідів (KNN).
2. Метод опорних векторів (Support Vector Machine – SVM) з лінійною функцією ядра.
3. SVM з функцією радіального базису (Radial Basis Function – RBF).
4. Гаусівський процес.
5. Дерево прийняття рішень (ДПР).
6. Випадковий ліс.
7. Багатошаровий перцептрон (БШП), тобто нейронна мережа.
8. Адаптивний бустинг (AdaBoost).
9. Наївний Байєсівський класифікатор.
10. Квадратичний дискримінантний аналіз (КДА).

Результати наведені у наступних таблицях:

1. Внутрішньокласова дисперсія:

КД зображена на рис. 4.6

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.98	0.98	0.98	0.94	0.95	0.93	0.84	0.9	0.78
Лінійний SVM	0.81	0.87	0.76	0.84	0.86	0.85	0.74	0.81	0.69
RBF SVM	0.99	1	0.98	0.96	0.99	0.91	0.92	0.99	0.81
Гаусівський процес	1	1	1	0.99	0.99	1	0.93	0.96	0.91
ДПР	0.89	0.93	0.84	0.87	0.93	0.8	0.81	0.89	0.72
Випадковий ліс	0.9	0.95	0.83	0.89	0.95	0.83	0.84	0.92	0.73
БШП	0.83	0.88	0.79	0.85	0.88	0.85	0.77	0.82	0.72
AdaBoost	0.94	0.96	0.92	0.92	0.95	0.89	0.87	0.92	0.81
Наївний Байес	0.48	0.99	0.09	0.51	0.99	0.11	0.51	0.97	0.12
КДА	0.54	0.81	0.28	0.51	0.81	0.22	0.52	0.79	0.27

Рисунок 4.6 Відношення дисперсій(КД)

ЛСД зображена на рис. 4.7

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.98	0.99	0.97	0.98	1	0.98	0.9	0.96	0.84
Лінійний SVM	0.93	0.95	0.92	0.92	0.95	0.92	0.9	0.96	0.84
RBF SVM	1	1	1	0.97	1	0.96	0.96	1	0.91
Гаусівський процес	1	1	1	1	1	1	1	1	1
ДПР	0.99	1	0.97	0.97	1	0.96	0.97	1	0.93
Випадковий ліс	0.98	1	0.95	0.95	1	0.94	0.97	1	0.93
БШП	0.95	0.97	0.93	0.92	0.95	0.92	0.95	1	0.89
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.9	0.89	0.9	0.86	0.86	0.89	0.87	0.94	0.8
КДА	0.9	0.88	0.94	0.88	0.81	0.94	0.87	0.9	0.84

Рисунок 4.7 Відношення дисперсій(КД)

ЛПД зображена на рис. 4.8

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.99	0.99	0.99	0.97	1	0.98	0.92	0.9	0.94
Лінійний SVM	0.88	0.74	0.98	0.78	0.5	0.98	0.86	0.71	0.96
RBF SVM	1	1	1	1	1	1	0.95	0.86	1
Гаусівський процес	0.39	1	0.15	0.25	1	0.13	0.37	1	0.15
ДПР	0.99	0.97	1	1	1	1	0.98	0.95	1
Випадковий ліс	0.97	0.93	0.99	0.93	0.88	0.98	0.95	0.86	1
БШП	0.94	0.92	0.96	0.87	0.88	0.94	0.97	0.9	1
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.56	0.96	0.37	0.43	1	0.37	0.51	1	0.31
КДА	0.5	1	0.28	0.39	1	0.31	0.46	1	0.26

Рисунок 4.8 Відношення дисперсій(КД)

2. Міжкласова дисперсія

КД зображена на рис. 4.9

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.97	0.99	0.95	0.92	0.95	0.88	0.83	0.87	0.78
Лінійний SVM	0.83	0.92	0.72	0.8	0.95	0.63	0.76	0.88	0.62
RBF SVM	1	1	1	0.99	1	0.97	0.94	1	0.84
Гаусівський процес	1	1	1	0.99	0.98	1	0.92	0.94	0.91
ДПР	0.93	0.95	0.92	0.9	0.92	0.88	0.86	0.92	0.8
Випадковий ліс	0.89	0.99	0.76	0.89	1	0.75	0.83	0.96	0.67
БШП	0.95	0.98	0.9	0.92	0.98	0.85	0.86	0.89	0.83
AdaBoost	0.98	0.99	0.95	0.95	0.99	0.9	0.9	0.96	0.81
Наївний Байєс	0.69	0.91	0.46	0.73	0.94	0.51	0.66	0.89	0.41
КДА	0.73	0.92	0.51	0.75	0.93	0.56	0.67	0.89	0.44

Рисунок 4.9 Відношення дисперсій(КД)

ЛСД зображена на рис. 4.10

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.99	1	0.97	0.97	1	0.96	0.93	1	0.87
Лінійний SVM	0.9	0.95	0.83	0.88	0.95	0.87	0.91	1	0.82
RBF SVM	0.99	1	0.98	0.98	1	0.98	0.98	1	0.96
Гаусівський процес	1	1	1	0.98	1	0.98	0.98	1	0.96
ДПР	0.96	1	0.9	0.92	1	0.91	0.96	1	0.91
Випадковий ліс	0.98	0.98	0.98	0.97	0.9	1	0.96	1	0.91
БШП	0.95	0.97	0.94	0.95	0.95	0.96	0.98	1	0.96
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.73	0.94	0.51	0.52	0.81	0.42	0.62	1	0.31
КДА	0.79	1	0.56	0.63	1	0.49	0.69	1	0.42

Рисунок 4.10 Відношення дисперсій(КД)

ЛПД зображена на рис. 4.11

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.99	0.97	0.99	0.93	0.78	1	0.88	0.76	0.96
Лінійний SVM	0.92	0.85	0.96	0.8	0.67	0.94	0.93	0.81	1
RBF SVM	1	1	1	1	1	1	0.95	0.86	1
Гаусівський процес	1	1	1	1	1	1	1	1	1
ДПР	0.99	0.97	1	1	1	1	0.98	0.95	1
Випадковий ліс	0.91	0.77	1	0.89	0.67	1	0.93	0.81	1
БШП	0.97	0.93	0.99	0.97	0.89	1	0.93	0.86	0.98
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.81	0.57	0.98	0.7	0.33	0.98	0.77	0.52	0.96
КДА	0.86	0.77	0.93	0.76	0.56	0.94	0.84	0.71	0.94

Рисунок 4.11 Відношення дисперсій(КД)

3.Відношення дисперсій

КД зображена на рис. 4.12

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.95	0.97	0.93	0.86	0.92	0.79	0.77	0.87	0.66
Лінійний SVM	0.82	0.87	0.78	0.82	0.89	0.74	0.75	0.82	0.67
RBF SVM	1	1	1	1	1	1	0.96	1	0.91
Гаусівський процес	1	1	1	0.99	0.99	1	0.97	0.98	0.95
ДПР	0.92	0.98	0.82	0.89	0.94	0.82	0.85	0.95	0.72
Випадковий ліс	0.86	0.95	0.73	0.81	0.94	0.65	0.75	0.92	0.56
БШП	0.88	0.93	0.83	0.83	0.92	0.72	0.82	0.88	0.75
AdaBoost	0.94	0.97	0.9	0.9	0.96	0.84	0.86	0.92	0.8
Наївний Байєс	0.71	0.68	0.79	0.78	0.79	0.79	0.67	0.61	0.81
КДА	0.76	0.69	0.9	0.78	0.76	0.82	0.71	0.69	0.77

Рисунок 4.12 Відношення дисперсій(КД)

ЛСД зображена на рис. 4.13

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	0.98	0.99	0.96	0.95	1	0.94	0.9	0.96	0.84
Лінійний SVM	0.93	0.95	0.9	0.86	0.95	0.85	0.92	1	0.84
RBF SVM	0.98	1	0.96	0.97	1	0.96	0.98	1	0.96
Гаусівський процес	0.96	0.98	0.94	0.94	1	0.92	0.97	1	0.93
ДПР	0.97	0.99	0.94	0.94	1	0.92	0.95	0.96	0.93
Випадковий ліс	0.95	0.98	0.93	0.89	0.9	0.91	0.89	0.94	0.84
БШП	0.94	0.95	0.93	0.91	0.95	0.91	0.96	1	0.91
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.72	0.95	0.48	0.5	0.81	0.38	0.6	1	0.29
КДА	0.78	1	0.54	0.63	1	0.49	0.68	1	0.4

Рисунок 4.13 Відношення дисперсій(ЛСД)

ЛПД зображена на рис. 4.14

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	1	0.99	1	0.97	1	0.98	0.93	0.9	0.96
Лінійний SVM	0.89	0.78	0.98	0.86	0.75	0.96	0.91	0.76	1
RBF SVM	1	1	1	1	1	1	0.95	0.86	1
Гаусівський процес	1	1	1	0.97	1	0.98	0.97	1	0.96
ДПР	1	1	1	1	1	1	1	1	1
Випадковий ліс	0.97	0.92	1	0.96	0.88	1	0.85	0.62	1
БШП	0.99	0.97	1	1	1	1	0.98	0.95	1
AdaBoost	1	1	1	0.97	1	0.98	0.97	1	0.96
Наївний Байєс	0.63	0.97	0.47	0.46	1	0.4	0.57	0.95	0.43
КДА	0.91	1	0.88	0.79	1	0.85	0.92	1	0.91

Рисунок 4.14 Відношення дисперсій(ЛПД)

4. CFS

КД зображена на рис. 4.15

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.98	0.99	0.98	0.92	0.97	0.85	0.83	0.89	0.77
Лінійний SVM	0.83	0.89	0.76	0.87	0.89	0.87	0.77	0.79	0.77
RBF SVM	0.95	0.97	0.92	0.92	0.97	0.85	0.87	0.92	0.8
Гаусівський процес	0.85	0.91	0.79	0.86	0.9	0.83	0.8	0.82	0.8
ДПР	0.9	0.92	0.89	0.88	0.9	0.89	0.84	0.86	0.83
Випадковий ліс	0.88	0.95	0.81	0.87	0.94	0.78	0.81	0.87	0.77
БШП	0.85	0.9	0.8	0.85	0.9	0.8	0.79	0.82	0.78
AdaBoost	0.94	0.95	0.92	0.91	0.95	0.87	0.84	0.89	0.8
Наївний Байєс	0.61	0.99	0.25	0.66	0.99	0.3	0.57	0.96	0.22
КДА	0.29	0.02	1	0.27	0.03	1	0.27	0.01	1

Рисунок 4.15 CFS(КД)

ЛСД зображена на рис. 4.16

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	0.98	0.99	0.97	0.95	1	0.94	0.89	0.96	0.82
Лінійний SVM	0.93	0.95	0.9	0.88	0.95	0.87	0.95	1	0.89
RBF SVM	1	1	1	1	1	1	1	1	1
Гаусівський процес	1	1	0.99	1	1	1	0.98	1	0.96
ДПР	0.98	1	0.96	0.97	1	0.96	0.98	1	0.96
Випадковий ліс	0.99	1	0.97	0.98	1	0.98	0.96	1	0.91
БШП	0.97	0.97	0.98	0.97	0.95	0.98	0.98	1	0.96
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.89	0.88	0.9	0.83	0.86	0.85	0.89	0.92	0.87
КДА	0.9	0.89	0.93	0.87	0.9	0.89	0.9	0.94	0.87

Рисунок 4.16 CFS(ЛСД)

ЛПД зображена на рис. 4.17

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	1	0.99	1	0.93	0.88	0.98	0.95	0.95	0.96
Лінійний SVM	0.81	0.59	0.97	0.73	0.5	0.94	0.85	0.67	0.98
RBF SVM	0.95	0.93	0.96	0.87	0.88	0.94	0.95	0.86	1
Гаусівський процес	0.94	0.92	0.96	0.87	0.88	0.94	0.97	0.9	1
ДІР	1	1	1	1	1	1	1	1	1
Випадковий ліс	0.97	0.95	0.99	0.97	1	0.98	0.97	0.9	1
БШП	0.89	0.8	0.95	0.76	0.62	0.92	0.95	0.86	1
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.51	0.12	0.99	0.56	0.12	0.98	0.51	0.1	1
КДА	0.45	0.04	1	0.58	0.12	1	0.51	0.1	1

Рисунок 4.17 CFS(ЛПД)

5. QCFE

КД зображена на рис. 4.18

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чутл.	Спец.	F1	Чутл.	Спец.	F1	Чутл.	Спец.
KNN	0.97	0.99	0.94	0.92	0.98	0.85	0.81	0.92	0.67
Лінійний SVM	0.8	0.89	0.7	0.8	0.93	0.64	0.75	0.86	0.62
RBF SVM	0.99	1	0.99	0.98	0.99	0.97	0.94	0.98	0.88
Гаусівський процес	1	1	1	0.99	0.98	1	0.94	0.97	0.91
ДПР	0.9	0.98	0.78	0.87	0.98	0.75	0.86	0.98	0.69
Випадковий ліс	0.87	0.98	0.7	0.81	0.97	0.63	0.84	0.97	0.67
БШП	0.87	0.94	0.79	0.86	0.97	0.73	0.82	0.89	0.73
AdaBoost	0.95	0.98	0.91	0.94	0.98	0.88	0.86	0.91	0.81
Наївний Байєс	0.67	0.59	0.85	0.7	0.61	0.83	0.63	0.53	0.81
КДА	0.74	0.66	0.92	0.77	0.69	0.88	0.71	0.61	0.89

Рисунок 4.18 QCFE (КД)

ЛСД зображена на рис. 4.19

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	0.96	0.99	0.93	0.86	0.95	0.85	0.85	0.98	0.71
Лінійний SVM	0.77	0.83	0.71	0.64	0.62	0.7	0.79	0.9	0.69
RBF SVM	0.99	0.99	0.98	0.97	0.95	0.98	0.97	0.98	0.96
Гаусівський процес	0.99	1	0.98	0.98	1	0.98	0.98	1	0.96
ДПР	0.95	0.96	0.94	0.87	0.86	0.91	0.88	0.9	0.87
Випадковий ліс	0.94	0.94	0.96	0.91	0.81	0.98	0.9	0.94	0.87
БШП	0.89	0.92	0.85	0.82	0.86	0.83	0.85	0.9	0.8
AdaBoost	1	1	1	0.98	1	0.98	0.95	1	0.89
Наївний Байєс	0.59	0.91	0.3	0.51	0.81	0.4	0.64	0.96	0.38
КДА	0.87	0.98	0.73	0.81	1	0.75	0.82	0.96	0.69

Рисунок 4.19 QCFE (ЛСД)

ЛПД зображена на рис. 4.20

Класифікатор	Навчальна вибірка (60%)			Валідаційна вибірка (20%)			Тестова вибірка (20%)		
	F1	Чут.	Спец.	F1	Чут.	Спец.	F1	Чут.	Спец.
KNN	0.97	0.97	0.98	0.89	0.89	0.9	0.86	0.9	0.87
Лінійний SVM	0.58	0.17	1	0.51	0.11	1	0.51	0.1	1
RBF SVM	1	1	1	1	1	1	0.95	0.86	1
Гаусівський процес	1	1	1	0.98	1	0.98	0.92	1	0.91
ДПР	0.87	0.68	1	0.71	0.37	1	0.83	0.57	1
Випадковий ліс	0.91	0.75	1	0.77	0.47	1	0.8	0.52	1
БШП	0.87	0.69	0.99	0.69	0.37	0.98	0.87	0.67	1
AdaBoost	1	1	1	1	1	1	1	1	1
Наївний Байєс	0.57	0.8	0.49	0.53	0.79	0.41	0.67	0.95	0.57
КДА	0.88	0.97	0.87	0.82	0.89	0.8	0.87	0.9	0.89

Рисунок 4.20 QCFE (ЛПД)

Порівняння критеріїв для відбору ознак наведене у таблиці 4.1 :

Таблиця 4.1

Порівняння критеріїв для відбору ознак

Критерій	Максимальне F1			Середнє арифметичне F1			Середнє зважене F1
	КД	ЛСД	ЛПД	КД	ЛСД	ЛПД	
Внутрішньокласова дисперсія	0.93	1	1	0.78	0.94	0.8	0.82
Міжкласова дисперсія	0.94	1	1	0.82	0.9	0.92	0.86
Відношення дисперсій	0.97	1	1	0.81	0.89	0.91	0.85
CFS	0.87	1	1	0.74	0.95	0.87	0.82
QCFE	0.94	0.98	1	0.82	0.86	0.83	0.83

Під середньозваженим F1 розуміється сума середніх арифметичних F1 помножених на вагові коефіцієнти (які залежать від розмірності вибірок). Як видно за таблицею вище, по метриці найкращим критерієм виявилась міжкласова дисперсія. Хоча доволі близький результат показав критерій відношення між внутрішньокласовою і міжкласовою дисперсіями.

Висновки до розділу 4

В даному розділі були описані критерії для знаходження оптимального варіанту селекції ознак, а саме:

1. Внутрішньокласову дисперсію
2. Міжкласову дисперсію
3. Відношення внутрішньокласової і міжкласової дисперсій
4. Послідовний вибір ознак

5. Рекурсивне виключення ознак з крос-валідацією

6. Комбінований критерій()

Дані критерії було використано на 1 підвибірці(конвексного датчику), після чого результати були проаналізовані і зроблені висновки(результат показав, що найбільше виділяються рекурсивне виключення ознак з крос-валідацією та комбінований критерій)

РОЗДІЛ 5

СТАРТАП АНАЛІЗ ПРОЕКТУ

Стартап-проект за темою магістерської дисертації є продовженням бакалаврського диплому [5], виконано під наглядом доцента кафедри біомедичної кібернетики Павловим Володимиром Анатолійовичем.

Метою стартапу було знаходження оптимального варіанту селекції ознак, щоб можна було ефективно виконати задачу бінарної класифікації «Норма:Патологія».

Виконавці стартап-проекту:

1. Кожара Катерина Миколаївна (рис. 5.1) – виконавець, магістр кафедри біомедичної кібернетики.



Рисунок 5.1. Виконавець стартап-проекту

2. Павлов Володимир Анатолійович (рис. 5.2) – науковий консультант, доцент кафедри біомедичної кібернетики, кандидат технічних наук, доцент.



Рисунок 5.2. Науковий керівник стартап-проекту

Відбір (обирання) ознак (англ. Feature Selection) – це процес відбору підмножини вагомих незалежних (вхідних) ознак для їх використання в моделюванні. Слід відрізнити це від «виділення ознак», яке створює нові ознаки, як функції від оригінальних. Дана процедура необхідна для того, щоб не виникало проблем прокляття розмірності [11] та мультиколінеарності [12].

Розв’язання подібних проблем дозволить отримати більш якісні та адекватні моделі прогнозування стану пацієнта за ультразвуковими знімками, що у свою чергу забезпечить підтримку лікарям під час прийняття рішень. Це визначає перспективу створення програмного застосунку наприклад такого, що представлений на рис. 5.3.

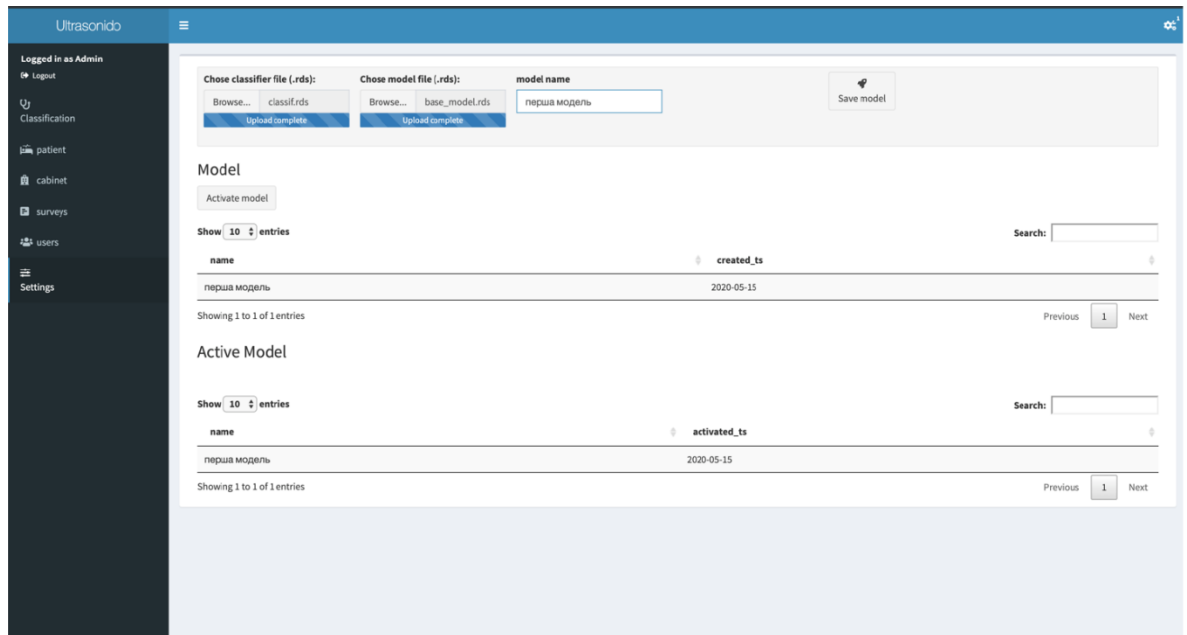


Рисунок 5.3. Приклад програмного застосунку

Висновки до розділу 5

Було проведено стартап-аналіз проекту магістерської дисертації. Також були представлені виконавці та приклад програмного застосунку.

ЗАГАЛЬНІ ВИСНОВКИ

В результаті виконання роботи магістерської дисертації були виконані поставлені задачі для досягнення головної мети:

1. Були проаналізовані різні джерела інформації.
2. Проаналізовано алгоритми класифікації, також основні методи класифікації та описано задачу наукового дослідження.
3. Була описана постановка задачі та вхідні дані для роботи.
4. Проведено дослідження, та відтворенні результати.

Критерієм оцінки моделі була точність на тестовій вибірці, і як показав результат, RFECV і QCFE найбільше виділяються серед усіх інших. Проте це поки що проміжковий результат, який буде покращено у майбутньому¹.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бабенко В. Класифікація ультразвукових зображень методом генетичного лісу дерев оптимальної складності (дипломна робота магістра). 2021. С. 1–125.
2. Гончарук М. Оптимізація параметрів перетворення УЗ зображень в задачі класифікації патологія-норма при дифузних захворюваннях печінки. 2020. С. 1–69.
3. 3.Грішко Д. Рішення задачі реконструкції та класифікації УЗ зображень печінки у просторі параметрів моделі ковзного вікна. 2021. С. 1–114.
4. 4.Іванченко А. Система диференціальної діагностики захворювань – аутоімунного гепатиту та хвороби Вільсона. 2021. С. 1–63.
5. Круглий В. Система підтримки прийняття рішень при діагностуванні фібротичних змін печінки. 2021. С. 1–110.
6. 6.Настенко Є., Павлов В., Носовець О., Круглий В., Гончарук М., Карлюк А., Грішко Д., Трофименко О., Бабенко В. Застосування текстурного аналізу у вирішенні задачі класифікації медичних зображень. *Біомедична інженерія і технологія*. 2020. Вип. 0, № 4. С. 69–82.
7. Настенко Є., Дикан І., Тарасюк Б., Павлов В., Носовець О., Бабенко В., Круглий В., Дибя М., Солодущенко В. Класифікація станів печінки при дифузних захворюваннях на основі статистичних показників текстури ультразвукових зображень та МГУА. *Індуктивне моделювання складних систем*. 2019. Вип. 11. С. 54–66.
8. 8.Трофименко О. Рішення задачі реконструкції та класифікації УЗ зображень печінки у просторі параметрів моделі просторової розгортки. 2021. С. 1–107.
9. Nastenکو I., Maksymenko V., Dykan I., Nosovets O., Tarasiuk B., Pavlov V., Babenko V., Kruhlyi V., Soloduschenko V., Dyba M., Umanets V.

- Liver Pathological States Identification in Diffuse Diseases with Self-Organization Models Based on Ultrasound Images Texture Features. Zbarazh, Ukraine:2020.
10. Nastenکو I., Maksymenko V., Galkin A., Pavlov V., Nosovets O., Dykan I., Tarasiuk B., Babenko V., Umanets V., Petrunina O., Klymenko D. Liver Pathological States Identification with Self-organization Models Based on Ultrasound Images Texture Features. *Advances in Intelligent Systems and Computing V. Ред. Shakhovska N., Medykovskyy M. O. Cham:Springer International Publishing, 2021. pp. 401–418.*
 11. Babenko V., Nosovets O., Nastenکو I., Pavlov V., Iakymchuk V., Matviichuk O., Suvorov M. Forming the System with the Functionality of Clinical Pharmacist for Personalized Treatment Strategy Searching. 2022. DOI: 10.1007/978-981-16-2377-6_47
 12. Bellman, R. E. (2003). “Dynamic Programming”, Publ. Courier Corporation, 340p.
 13. Goldberger, A. S. (1991). “A Course in Econometrics”, Publ. Harvard University Press, 405p.
 14. Roffo, G., Melzi, S. & Cristani, M. (2015). “Infinite Feature Selection”, 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 4202-4210. DOI: 10.1109/ICCV.2015.478.
 15. Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J. & Yuan, T.-F. (2015). “Detection of Subjects and Brain Regions Related to Alzheimer’s Disease Using 3D MRI Scans Based on Eigenbrain and Machine Learning”, *Frontiers in Computational Neuroscience*, Vol. 9, pp. 1–15. DOI: 10.3389/fncom.2015.00066.
 16. Roffo, G. & Melzi, S. (2017). “Features Selection via Eigenvector Centrality”, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10312 LNCS, pp. 19–35. DOI: 10.1007/978-3-319-61461-8_2.

- 17.Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y. & Y. Liu (2011). “Genetic Algorithm-Based Efficient Feature Selection for Classification of Pre-MiRNAs”, *Genetics and Molecular Research*, Vol. 10, No. 2, pp. 588–603. DOI: 10.4238/vol10-2gmr969.
- 18.Zhang, Y., Wang, S., Phillips, P. & Ji, G. (2014). “Binary PSO with Mutation Operator for Feature Selection Using Decision Tree Applied to Spam Detection”, *Knowledge-Based Systems*, Vol. 64, pp. 22–31. DOI: 10.1016/J.KNOSYS.2014.03.015.
- 19.Hall, M. A. (1999). “Correlation-Based Feature Selection for Machine Learning”, 109 p.
- 20.Zare, H., Haffari, G., Gupta, A. & Brinkman, R. R. (2013). “Scoring Relevancy of Features Based on Combinatorial Analysis of Lasso with Application to Lymphoma Diagnosis”, *BMC Genomics*, Vol. 14, pp. 1–9. DOI: 10.1186/1471-2164-14-S1-S14.
- 21.N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, 321-357, 2002. DOI: 10.1613/jair.953
- 22.Non-invasive diagnosis of liver fibrosis and cirrhosis [Електронний ресурс]–Режим доступу до ресурсу: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631961/ttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631961/>.
- 23.Симптоми гепатиту В [Електронний ресурс] – Режим доступу до ресурсу: <http://orcssd.com.ua/ua/patients/hepatitis/diagnostics>.
- 24.Логістична регресія [Електронний ресурс] – Режим доступу до ресурсу: https://courses.prometheus.org.ua/assets/courseware/v1/cdf163c83c64f8357ddbcdac82f7d624/c4x/IRF/ML101/asset/%D0%A2%D0%B8%D0%B6%D0%B4%D0%B5%D0%BD%D1%8C_1_%D0%BA%D0%BE%D0%BD%D1%81%D0%BF%D0%B5%D0%BA%D1%82.pdf.

- 25.Метод Групового Урахування Аргументів [Електронний ресурс] –
Режим доступу до ресурсу: <http://www.gmdh.net/gmdhu.htm>.
26. Кожара К.М., Настенко Є.А., Павлов В.А., Бабенко В.О. АЛГОРИТМ
САМООРГАНІЗАЦІЇ АНСАМБЛЮ У ВИРІШЕННІ ЗАДАЧІ
КЛАСИФІКАЦІЇ «НОРМА-ПАТОЛОГІЯ» НА ЗОБРАЖЕННЯХ УЗД
ПЕЧІНКИ. IVміжнародна науково-практична інтернет-конференція
Сучасні виклики і актуальні проблеми науки, освіти та виробництва:
міжгалузеві диспути [Електронний ресурс] // 15.05.2020. – 2020.
– Режим доступу до ресурсу:
https://openscilab.org/wpcontent/uploads/2020/05/suchasni-vikliki-i-aktualni-problemi-nauki-osviti-ta-virobnictva_2020_05_15_tezy.pdf – 330
с.
- 27.Бабенко В. Класифікація ультразвукових зображень методом
генетичного лісу дерев оптимальної складності (дипломна робота
магістра). 2021. С. 1–125.
- 28.Orlacchio A, Bolacchi F, Petrella MC, Pastorelli D, Bazzocchi G, Angelico
M, et al. Liver contrast enhanced ultrasound perfusion imaging in the
evaluation of chronic hepatitis C fibrosis: preliminary results. Ultrasound
Med Biol. 2011;37:1–6. [PubMed] [Google Scholar].
- 29.Ishibashi H, Maruyama H, Takahashi M, Shimada T, Kamesaki H,
Fujiwara K, et al. Demonstration of intrahepatic accumulated microbubble
on ultrasound represents the grade of hepatic fibrosis. Eur
Radiol. 2012;22:1083–1090. [PMC free article] [PubMed] [Google
Scholar].
- 30.Textural Features for Image Classification [Електронний ресурс] –
Режим доступу до ресурсу:
<https://ieeexplore.ieee.org/document/4309314/references#references>.
31. Кожара К. Алгоритм самоорганізації ансамблю ознак у вирішенні
задачі класифікації «норма-патологія» на зображеннях УЗД
печінки.2020. С.1-78.