Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет ННЦЗФН
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів редукції даних в системах інтелектуальної діагностики
(тема)

Виконав:
студент 2 курсу, групи СШІмз-20-1
Півнев Д.С.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)
Керівник доц. Дейнеко А.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет ННЦЗФН
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

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

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Півневу Дмитру Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів редукції даних в системах інтелектуальної діагностики

затверджена наказом університету від 25 жовтня 2021 р. № 168Стз

2. Термін подання студентом роботи до екзаменаційної комісії 08 грудня 2021 р.

3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел та відомих наукових проектів щодо розробки та дослідження методів редукції даних, специфікація мови C#, набір даних «kaggle», набір даних «diahnostuvannja-ta-prohnozuvannja»

4. Перелік питань, що потрібно опрацювати в роботі

1. Аналіз предметної галузі
2. Модифікація методу стиснення інформації з групуванням ознак
3. Модуль редукції даних для діагностичної системи та результати моделювання
4. Експериментальне моделювання методів редукції даних

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)_____

_____
_____
_____
_____
_____
_____
_____
_____
_____

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

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	доц. Дейнеко А.О.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на дипломну роботу	01.09.2021	
2	Аналіз предметної області та постановка задачі	02.09.2021 – 07.09.2021	
3	Теоретичні дослідження	08.09.2021 – 21.09.2021	
4	Розробка практичної реалізації	22.09.2021 – 12.10.2021	
5	Імітаційне моделювання	13.10.2021 – 31.10.2021	
6	Оформлення пояснювальної записки	1.11.2021 – 20.11.2021	
7	Захист перед ЕК	09.12.2021	

Дата видачі завдання _____ _____ 20 ___ р.

Студент _____
(підпис)

Керівник роботи _____ __доц. Дейнеко А.О._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 88 с., 11 рис., 11 табл., 2 дод., 23 джерела.

ВИБІРКА ДАНИХ, ДІАГНОСТИЧНА СИСТЕМА, ІНФОРМАТИВНІСТЬ ОЗНАКИ, НЕЙРОМЕРЕЖЕВЕ МОДЕЛЮВАННЯ, ПРОГРАМНИЙ МОДУЛЬ, РЕДУКЦІЯ ДАНИХ.

Об'єкт дослідження – процеси інтелектуального аналізу інформативності ознак в системах діагностики.

Предмет дослідження – методи редукції даних в системах інтелектуальної діагностики.

Мета роботи – розробка та дослідження алгоритмічних та програмних засобів, які реалізують методи редукції даних для інтелектуальних систем діагностики.

Методи дослідження – методи аналізу інформативності даних, методи редукції діагностичних моделей, методи нейромережевого моделювання.

Досліджено методи скорочення обсягу діагностичної інформації (скорочення кількості ознак, кількості екземплярів у вибірці) для редукції вибірок даних, що характеризують стан об’єктів діагностики. Запропоновано програмний модуль редукції даних для системи інтелектуальної діагностики.

РЕФЕРАТ

Пояснительная записка: 88 с., 11 рис., 11 табл., 2 прил., 23 источника.

ВЫБОРКА ДАННЫХ, ДИАГНОСТИЧЕСКАЯ СИСТЕМА, ИНФОРМАТИВНОСТЬ ПРИЗНАКИ, НЕЙРОСЕТЕВОЕ МОДЕЛИРОВАНИЕ, ПРОГРАММНАЯ МОДУЛЬ, РЕДУКЦИЯ ДАННЫХ.

Объект исследования – процессы интеллектуального анализа информативности признаков в системах диагностики.

Предмет исследования – методы редукции данных в системах интеллектуальной диагностики.

Цель работы – разработка и исследование алгоритмических и программных средств, реализующих методы редукции данных для интеллектуальных систем диагностики.

Методы исследования – методы анализа информативности данных, методы редукции диагностических моделей, методы нейросетевого моделирования.

Исследованы методы сокращения объема диагностической информации (сокращение количества признаков, количества экземпляров в выборке) для редукции выборок данных, характеризующих состояние объектов диагностики. Предложен программный модуль редукции данных для системы интеллектуальной диагностики.

ABSTRACT

Explanatory note: 88 p., 11 fig., 11 tabl., 2 ann., 23 sources.

DATA SAMPLING, DIAGNOSTIC SYSTEM, INFORMATIVITY OF THE SIGN, NON-NETWORK MODELING, SOFTWARE MODULE, DATA REDUCTION

The object of research – the computer analysis processes of the signs informativeness in diagnostic systems.

The subject of research – methods of data reduction in intelligent diagnostic systems.

The purpose of the work is to develop and research algorithmic and software tools that implement data reduction methods for intelligent diagnostic systems.

Research methods – methods of data informative analysis, methods of reduction of diagnostic models, methods of neural network modeling.

Methods of reducing the amount of diagnostic information (reducing the number of features, the number of copies in the sample) for the reduction of data samples that characterize the state of diagnostic objects have been studied. The software module of data reduction for the system of intelligent diagnostics is offered.

ЗМІСТ

Вступ 9
1 Аналіз предметної галузі 11
1.1 Загальна характеристика завдань інтелектуальної діагностики 11
1.2 Особливості застосування методів редукції даних в задачах інтелектуальної діагностики 13
1.3 Методи редукції даних та діагностичних ознак 17
1.3.1 Евристичні методи 17
1.3.2 Інформаційні методи 19
1.3.3 Статистичні методи 24
1.3.4 Імовірнісні методи 28
1.3.5 Нейромережеві методи 28
1.4 Постановка задач дослідження 31
2 Модифікація методу стиснення інформації з групуванням ознак 32
2.1 Опис базового методу стиснення інформації з групуванням ознак 32
2.2 Групування ознак 34
2.3 Критерії оцінки ознак в групах 36
2.4 Модифікований метод стиснення інформації з групуванням ознак 39
3 Модуль редукції даних для діагностичної системи та результати моделювання 45
3.1 Структура та функції пропонованого модуля редукції даних 45
3.2 Програмна реалізація та інтерфейс користувача модуля редукції даних 46
3.3 Результати моделювання 57
Висновки 64
Перелік джерел посилання 65
Додаток А Фрагмент програми редукції даних 68
Додаток Б Відомість кваліфікаційної роботи 88

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ СКОРОЧЕНЬ І ТЕРМІНІВ

БШП – багатошаровий перцептрон;

ДО – діагностична ознака;

ДС – діагностична система;

ОС – операційна система;

ЦП – цільовий параметр;

ШНМ – штучна нейронна мережа;

DLL – dynamic link library – бібліотека динамічної компоновки;

PCA – principal component analysis – аналіз головних компонентів.

ВСТУП

Під час розробки та експлуатації технологічних процесів, виконанні біологічних та медичних досліджень широко застосовують методи неруйнівної комп’ютерної діагностики. Особливістю програмних комплексів і систем, що реалізують ці методи, є робота з великими обсягами даних, за якими виконується діагностика. Оскільки задачі, для яких виконується діагностичне моделювання, можуть суттєво відрізнятися (задачі технічної та біомедичної діагностики, дослідження впливу зовнішніх факторів на якість виробів тощо), то в кожному з випадків необхідно для побудови діагностичної моделі виконувати накопичення, зберігання і обробку інформації про об’єкт. Для складних систем і об’єктів з багатьма зовнішніми факторами, а також великою кількістю екземплярів даних, процес побудови моделі є доволі складною і трудомісткою задачею.

Одним з можливих варіантів розв’язання цієї задачі є виконання редукції даних, тобто скорочення обсягу даних. Загальні положення теорії редукції наголошують, що для збереження загальних рис і властивостей об’єкта дослідження (моделювання) не потрібно накопичувати усю інформацію про нього, а навпаки визначати і відокремлювати виключно ті ознаки, що сприяють найбільш повному опису об’єкта, без додавання зайвої інформації [1], [2].

Для моделювання в системах технічної діагностики використовуються вибірки значень ознак, що характеризують стан складного об’єкта (процеса) та цільового параметра (головної характеристики об’єкта) [3], [4].

Для практичного вирішення задач неруйнівної діагностики і прогнозування стану складних об’єктів і процесів за певними ознаками доцільно дослідити методи редукції діагностичних моделей та їх реалізації з використанням засобів обчислювального інтелекту.

Метою кваліфікаційної роботи є розробка та дослідження алгоритмічних та програмних засобів, які реалізують методи редукції даних

для комп’ютерних систем діагностики, що дозволяє зменшувати кількість діагностичних ознак та розмір вибірок.

Згідно з метою роботи поставлено наступні задачі:

– аналіз існуючих джерел за темою дослідження;

– дослідження методу стиснення інформації з групуванням ознак;

– розробка модуля редукції даних для діагностичної системи;

– моделювання досліджуваних методів.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Загальна характеристика завдань інтелектуальної діагностики

Термін «діагностика» походить від грецького слова «діагностікос», що означає розпізнавання, визначення. У процесі діагностики встановлюється діагноз, тобто визначається стан та працездатність досліджуваної системи.

В залежності від галузі застосування розрізняють чимало різновидів діагностики [5], [6]. Це, наприклад:

– технічна діагностика – наука про розпізнавання стану технічної системи, що включає широке коло проблем, пов'язаних з отриманням та оцінкою діагностичної інформації;

– медична діагностика – галузь науково-медичних знань, сутність якої складають теорія, методи і засоби постановки діагнозу про стан пацієнта;

– педагогічна діагностика – підрозділ педагогіки, що вивчає принципи і методи розпізнавання та встановлення ознак, які характеризують нормальний або з відхилом від норм перебіг педагогічного процесу;

– економічна діагностика –визначення характеру порушень і області господарської діяльності, у якій відбуваються відхилення від нормальної відносної динаміки різних процесів.

Згідно з завданням в кваліфікаційній роботі досліджуються питання, що пов’язані з різними напрямами діагностики.

Інтелектуальна діагностика вивчає методи отримання та оцінки діагностичної інформації, діагностичні моделі та алгоритми прийняття рішень.

Загальна структура діагностики характеризується двома основними напрямками: теорією контрольоспроможності та теорією розпізнавання.

Теорія контрольоспроможності передбачає розробку засобів та методів отримання діагностичної інформації, автоматизований контроль та пошук порушень нормального стану об'єкта діагностики.

Використання засобів неруйнівного контролю та методів первинної статистичної обробки даних виявляється недостатнім при побудові ефективних систем прийняття рішень у задачах діагностики, оскільки ці засоби не дозволяють будувати високоточні багатовимірні діагностичні моделі нелінійних процесів, а також отримувати знання з експериментальних даних та адаптуватися до змін у зовнішньому та внутрішньому середовищі об'єкта, що діагностується.

Перспективним напрямом розвитку всіх різновидів діагностики слід вважати інтелектуальну діагностику. Під інтелектуальною діагностикою розуміють сукупність методів та засобів, які дозволяють отримувати надійні та адекватні діагностичні моделі складних об'єктів і процесів за експериментальними даними, що мають високу ефективність і здатність адаптуватися до змін у зовнішній і внутрішній середовищах об'єкта діагностики.

Теоретичним базисом для здійснення інтелектуальної діагностики є теорія розпізнавання образів та методи обчислювального інтелекту.

Теорія розпізнавання містить розділи, пов'язані з побудовою алгоритмів розпізнавання, вирішальних правил та діагностичних моделей.

Інтелектуальна діагностика вивчає алгоритми розпізнавання стосовно завдань діагностики, які можуть розглядатися як завдання класифікації.

Алгоритми розпізнавання в інтелектуальній діагностиці частково ґрунтуються на діагностичних моделях, що встановлюють зв'язок між станами об'єкта діагностики та їх відображенням у просторі діагностичних сигналів. Важливою частиною проблеми розпізнавання є правила прийняття рішень (вирішальні правила). Вирішення діагностичної задачі (наприклад, віднесення станів об'єкта до нормальних, проблемних або незадовільних) завжди пов'язане з ризиком помилкових рішень. Зменшення відсотку таких помилок може бути досягнуто шляхом використання сучасних інтелектуальних методів побудови систем класифікації та розпізнавання.

Інтелектуальна діагностика стану складних об'єктів є системою, яка

повинна мати інформаційне, технічне та математичне забезпечення.

Інформаційне забезпечення включає способи отримання діагностичної інформації, її зберігання та систематизацію. Інформаційне забезпечення містить необхідний масив технічних відомостей, що обновлюються з часом.

Технічне забезпечення є сукупністю пристроїв отримання та обробки інформації (діагностичні прилади, датчики, сигналізатори тощо). Важливу частину технічного забезпечення сучасних систем діагностики становлять комп'ютерні засоби (мікропроцесори, мікроконтролери, перетворювачі тощо).

Математичне забезпечення включає математичні методи, алгоритми та програми розпізнавання та чисельного моделювання.

До основних етапів побудови систем інтелектуальної діагностики стану технічних об'єктів відносяться:

– проведення експерименту;

– обробка експериментальних даних;

– оцінка інформативності та відбір діагностичних ознак;

– побудова діагностичної моделі з використанням засобів обчислювального інтелекту.

Згідно з завданням на атестаційну роботу далі розглядатимемо питання побудови та дослідження методів оцінки інформативності і відбору діагностичних ознак, що дозволяють здійснювати редукцію (скорочення розмірності) розроблюваних діагностичних моделей.

1.2 Особливості застосування методів редукції даних в задачах інтелектуальної діагностики

Інформацію про стан об’єкту (системи, виробу тощо) несуть діагностичні ознаки (параметри). Чим складніше об’єкт діагностики, тим більше ознак необхідно вжити для опису його стану. Але надмірне збільшення кількості ознак значно утскладює процес побудови та використання діпгностичних моделей цих об’єктів. Такі моделі об’єктів буде

складно створити (або взагалі неможливо), взаємодія з ними навіть при застосуванні потужних комп'ютерних засобів буде виконуватися дуже повільно. У такому випадку дуже актуальною стає задача скорочення кількості ознак до мінімально необхідної для створення адекватної діагностичної моделі об'єкту. Спрощення (редукцію) наборів ознак об'єкту у складі діагностичної системи може бути реалізовано за допомогою модулю редукції даних. З діагностичної системи до неї будуть надходити набори діагностичних даних, що складаються з різноманітних ознак; модуль редукції має їх обробити і здійснити автоматичне вилучення надлишкових даних, але таким чином, щоб не зашкодити адекватності математичних моделей, які будуть створюватися за цими даними [7].

Теоретично, розв'язання задачі редукції даних можливо розподілити на декілька послідовних етапів [8], [9], [10]:

– етап оцінки інформативності та критеріїв залежності ознак;

– етап групування ознак;

– етап обробки груп ознак за оцінкою інформативності;

– етап формування нової, скороченої кількості ознак.

На першому етапі виконується оцінка інформативності ознак – величин, що кількісно характеризують придатність наборів ознак до розпізнавання класів об'єктів в задачах діагностики (придатні та дефектні тощо). Окрім інформативності, також можуть обчислюватися критерії оцінки за якими встановлюють межі зв'язків та взаємозалежність між ознаками [11].

На другому етапі, в залежності від особливостей застосованих методів редукції, здійснюється групування ознак за деяким обраним алгоритмом.

На третьому етапі після завершення групування з кожної групи взаємопов'язаних ознак за визначеними на першому етапі критеріями здійснюється відбір ознак за рівнем інформативності або критерієм оцінки для певної ознаки.

Завершальним етапом практичного розв'язання задачі редукції є вилучення ознак з набору даних, що на третьому етапі були оцінені, як

надлишкові чи навіть зайві [12].

Отже, головним принципом роботи модулю редукції є обробка даних що надходять до неї, таким чином, щоб мати можливість скоротити обсяг цих даних. Утворення оптимальної кількості ознак після обробки, дозволить діагностичній системі створювати швидкі для розрахунків та адекватні (точні) математичні моделі об’єктів, що діагностуються; надасть можливість виконувати обчислення в цих моделях з мінімально необхідними витратами машинного часу і інших обчислювальних ресурсів, нарешті дозволить створювати оптимальні за розміром архіви даних в системах керування якістю, оскільки не буде здійснюватися накопичення і зберігання надлишкових (зайвих) даних.

Вибір діагностичних параметрів (ознак), що є інформативними для побудови відповідних прогнозуючих (діагностичних) моделей оцінки стану технічних об’єктів, може здійснюватися за допомогою різних підходів та методів.

Наприклад, якщо розглядати типову для діагностичних систем прогнозовану оцінку якості виробів, то кількість показників, що описують усі сторони такої оцінки, може досягати кількох десятків і навіть кількох сотень. Очевидно, не всі ці показники рівнозначні, тобто не всі містять суттєву інформацію про якість виробу. Звичайно, виникає завдання: скільки ознак і які ознаки потрібно контролювати, щоб прийняти рішення про якість виробу, що задовольняє заданому критерію, тобто необхідно здійснювати відбір найбільш важливих для оцінки якості показників або ознак, які називаються інформативними. При цьому значущість або інформативність ознак потрібно оцінювати кількісно, тому що тільки в цьому випадку можна віддати перевагу тій чи іншій ознаці, що важко зробити при якісному підході до відбору ознак.

Слід зауважити, що фізичний підхід до виявлення інформативних ознак, заснований на аналізі фізичної моделі виробу, не протиставляється математичному. На попередньому етапі може використовуватися фізичний аналіз, на основі якого відбирається вся сукупність вимірюваних

ознак (загалом надмірна за заданим критерієм), в тій чи іншій мірі характеризують якість виробів. Далі застосовується математичний підхід, коли з усієї надлишкової сукупності відбираються найбільш інформативні (значущі) ознаки. Завдання, яке при цьому вирішується, можна сформулювати наступним чином. Нехай є намірна за кількістю сукупність ознак $x = \{x_1, x_2, ..., x_n\}$. Необхідно цих *n* ознак відібрати *v* найбільш інформативних за заданим критерієм, причому $v \le n$. Не уточнюючи критерію інформативності, можна висунути низку загальних вимог до виділеної сукупності інформативних ознак:

– множина відібраних ознак повинна мати достатню інформативність, тобто. додавання до неї ще однієї будь-якої ознаки не має призводити до значної зміни значення критерію інформативності;

– множина відібраних ознак повинна містити якомога меншу кількість ознак з метою мінімізації обчислюваних витат на подальше проведення діагностики;

– множина відібраних ознак може складатися з кількісних та якісних ознак, але має допускати кількісну оцінку (ранжування).

Сукупність ознак, що задовольняє вищезазначеним вимогам, будемо називати оптимальною сукупністю ознак.

Завдання вибору інформативних ознак може виникнути у процесі розробки, виготовлення та експлуатації виробів. При цьому у зв'язку зі специфічністю завдань у кожному випадку можуть застосовуватись різні методи. Необхідність визначення інформативних ознак з наявної сукупності обумовлена також тим, що якість прогнозу не інваріантна до системи ознак, що використовуються. У практичних випадках характеристики класів та прогнозуючі правила визначаються за експериментальними даними обмеженого обсягу, тому додавання неінформативних ознак призводить до більш сильного перетину представників класів у просторі ознак, що може погіршити якість прогнозу.

Під інформативністю ознак в системах техічної діагностики зазвичай

розуміють величину, що кількісно характеризує придатність ознак (або їх набору) для розпізнавання класів виробів, що задовальняють або не задовальняють необхідним вимогам. З фізичної точки зору, інформативність відображає ступінь взаємозв'язку обраних ознак з процесами, що призводять до суттєвих порушень якості виробів. Інформативність набору ознак дорівнює сумі інформативності окремих ознак лише за їх незалежності. Якщо ж ознаки залежні друг від друга, то інформативність набору виражається через інформативність окремих ознак за більш складними підходами. Дослідження показали, що у завданнях діагностичного прогнозування переважно доводиться мати справу із системою статистично залежних ознак.

Сукупність інформативних ознак в системах інтелектуальної діагностики визначається з використанням методів розпізнавання образів. Таке завдання інтерпретується як завдання визначення системи ознак, які при використанні в алгоритмі навчання алгоритмів діагностики, що реалізуються зазвичай за допомогою нейромережевих або нейро-нечітких моделей, забезпечують найкращий поділ класів навчальної вибірки. Бажання оптимізувати набір ознак так, щоб забезпечити найкращий поділ класів, при вибраному алгоритмі навчання та заданій навчальній вибірці призводить до необхідності повного перебору всіх підмножин вихідної системи ознак. У більшості випадків для цієї мети потрібно надмірно великий обсяг машинного часу через те, що для кожного підмножини ознак потрібно провести процес навчання.

Розглянемо докладніше різні підходи до побудови методів інформативної системи діагностичних ознак, що існують на сьогодні.

1.3 Методи редукції даних та діагностичних ознак

1.3.1 Евристичні методи

Для отримання найкращого (за вибраним критерієм) набору ознак

можна використовувати таке евристичне правило. З усіх ознак вибирається одна найбільш інформативна; далі, до першої ознаки додається така ознака з n-1, що залишилися, щоб інформативність пари ознак для прогнозування була найбільшою; потім до отриманої пари ознак додається найкращим чином нова ознака і так далі. Процес закінчується тоді, коли інформативність деякої сукупності ознак не перевищує суттєво інформативність сукупності, отриманої на попередньому кроці, або коли досягнуто необхідного рівня інформативності. Процедура такого відбору ознак називається методом скороченого перебору із додаванням ознак.

Процес відбору ознак можна вести і у зворотному порядку: спочатку з вихідної сукупності n ознак вибирається найбільш інформативне підмножина (n-1) ознак, потім з цієї підмножини відбирається найбільш інформативна підмножина (n-2) ознак із усіх можливих підмножин тощо. Процедура такого відбору ознак називається методом скороченого перебору з виключенням ознак.

Поєднуючи ідеї алгоритмів перебору і виключенням та додаванням ознак, можна отримати метод комбінованого перебору.

На початку роботи алгоритму комбінованого перебору по черзі оцінюється ефективність класифікації кожної ознаки окремо. Після перебору всіх ознак у проміжний набір додається та ознака, для якої ефективність класифікації виявиться найбільшою. Потім до проміжного набору додають по черзі ті ознаки, які ще не увійшли в набір та оцінюють ефективність класифікації. Остаточно набір заносять ту ознаку, при додаванні якої ефективність класифікації виявиться найбільшою в порівнянні з ефективністю за його відсутності. Цей процес триває до того часу, доки виявиться, що серед ознак немає жодного, включення якого поліпшило б результат класифікації. Після закінчення процесу додавання ознак починає працювати блок виключення ознак набору. З проміжного набору виключається ознака і з'ясовується, чи це призводить до підвищення ефективності класифікації. Остаточно з набору викидається та ознака,

виняток якої максимально покращує або, у крайньому випадку, не погіршує класифікацію. Процес виключення повторюється до того часу, поки з проміжного набору ознак виявиться неможливим виключити жодного, не погіршуючи ефективності класифікації. Після закінчення процесу виключення ознак знову робиться спроба додати до набору ознак такі ознаки, які б поліпшили ефективність класифікації. Процедура відбору ознак закінчується, коли до системи відібраних ознак виявляється неможливим додати жодної, що поліпшує класифікацію, ані виключити жодної ознаки, не погіршуючи класифікації об'єктів навчальної вибірки.

Перевагою евристичного підходу до відбору ознак є відносна простота реалізації евристичних процедур. Недоліками цього підходу є відсутність математичного обґрунтування евристичних процедур та тривалість процесу ітеративного перебору ознак.

1.3.2 Інформаційні методи

Порівняльна оцінка інформативності ознак x_k та x_s, k, s = 1, ..., n, може бути здійснена також на основі визначення кількості інформації, яку отримує система в процесі розпізнавання об'єктів у результаті визначення кожної з цих ознак.

Нехай об'єкт, що розпізнається, може належати лише одному з m класів, апріорні ймовірності віднесення цього об'єкта до певного класу позначимо $P(K_j)$, j=1, 2, ..., m,, а умовні щільності розподілу значень ознак $f_j(x_k)$, $f_j(x_s)$.

Якщо ознака x_k набуває дискретних значень з ймовірностями $P\left(x_{k_q}\right)=\sum_{j=1}^{m} P\left(K_j\right)P\left(x_{k_q} \mid K_j\right)$, q = 1, ..., n_k, то інформативність ознаки x_k дорівнює:

$$J_{x_k} = -\sum_{i=1}^{m} P(K_i)\log P(K_i) + \sum_{j=1}^{m} P(K_j)\sum_{i=1}^{m}\sum_{q=1}^{n_k} P(x_{k_q} \mid K_j) P(K_i \mid x_{k_q}) \log P(K_i \mid x_{k_q}). \qquad (1.1)$$

.

Аналогічні вирази можуть бути отримані для ознаки x_s,. Будемо вважати, що інформативність ознаки x_k вище, ніж інформативність ознаки x_s, якщо $J_{xk} > J_{xs}$. Інформативність ознак не є постійною величиною і не є безумовною величиною, а навпаки, у загальному випадку кількість інформації, яку отримує система розпізнавання в результаті вимірювання кожної даної ознаки, залежить від того, які ознаки були визначені раніше і які значення вони набули. Це однаково стосується як статистично залежних, так і статистично незалежних ознак. Наведений підхід складає основу методу відбору ознак за кількістю привнесеної інформації.

Контроль з інформаційного погляду дозволяє зняти невизначеність стану об'єкта, яка кількісно характеризується ентропією його стану. Нехай координати x_s, s = l, ..., k вектору стану X – незалежні величини. Тоді, використовуючи відому властивість, що полягає в тому, що ентропія сукупності незалежних величин дорівнює сумі ентропії цих величин, можна записати:

$$H_\Sigma = \sum_{s=1}^{k} H_s(x), \qquad (1.2)$$

де H_Σ – ентропія стану об'єкта, що діагностується;

$H_s(x)$ – безумовна ентропія s-ї ознаки.

Кількість інформації, яку несе ознака стану, можна оцінити виразом

$$J_s(x_s, \Sigma) = H_\Sigma - H_s(\Sigma / x_s), \qquad (1.3)$$

де $H_s(\Sigma / x_s)$ – умовна ентропія стану об'єкта після контролю ознаки xs.

Вибір ознак слід починати з ознаки x_s, яка несе максимальну кількість інформації.

Наведений підхід складає основу методу відбору ознак за мінімумом ентропії.

Упорядкування ознак за ступенем інформативності можна здійснювати також і за величиною дисперсії розподілу ознаки, адже чим менша дисперсія ознаки, тим щільніший розподіл і тим більша ймовірність того, що стани порівнюваних об'єктів належать до одного класу, який характеризується певним ступенем працездатності чи терміном служби. Цей метод можна рекомендувати при ймовірнісному прогнозуванні, коли обчислюються та аналізуються величини дисперсії компонентів прогнозованого процесу

Переважна більшість діагностичних ознак пов’язані за якимось правилом чи функцією, і цей зв’язок може бути більш або менш тісним. В якості міри зв’язку у такому разі можливо використовувати співвідношення: відстані та ступеня зв’язку.

Цілком логічним є висновок, що існують деякі сукупності (групи) взаємопов’язаних між собою параметрів (ознак). Звідси можливо вивести припущення, що параметри найбільш сильно пов’язані між собою в тому випадку, якщо вони підпорядковані однаковим впливам та зовнішнім факторам. Прийнявши гіпотезу, за якою параметри, що належать до однієї групи, можуть бути сильно пов’язані між собою, а параметри які належать до різних груп пов’язані між собою відносно слабкими зв’язками, можливо побудувати розбиття усієї сукупності параметрів на певні групи.

Такий підхід складає основу методу стиснення інформації з групуванням ознак. Заключним етапом в цьому методі є стиснення інформації (скорочення її обсягу). Ефект стиснення створюється за рахунок виключення з груп визначеної кількості параметрів згідно методики обробки групи або в залежності від зв’язків між параметрами в групі.

Інформаційні методи діагностики передбачають можливість проведення досліджень з оцінювання так званої діагностичної значимості обстеження

об’єктів діагностики.

Згідно з діагностичною значимістю обстеження за ознакою x_j для класу (діагнозу) K_i можна вважати величину інформації, що вноситься всіма реалізаціями ознаки x_j для віднесення до класу K_i.

Для *m*-розрядної ознаки

$$Z_{K_i}(x_j) = \sum_{s=1}^{m} P(x_{js} / K_i) Z_{K_i}(x_{js}), \tag{1.4}$$

де x_{js} – s-й розряд *j*-ї ознаки.

Діагностична значимість обстеження враховує всі можливі реалізації ознаки і є математичним очікуванням величини інформації, що вноситься окремими реалізаціями. Якщо величина відноситься тільки до одного класу K_i., то її називають частковою діагностичною значимістю обстеження за ознакою x_j.

Слід зазначити, що $Z_{K_i}(x_j)$ визначає незалежну діагностичну значимість обстеження. Вона відповідає випадку, коли обстеження проводиться першим чи коли результати інших обстежень невідомі. Величина $Z_{K_i}(x_j)$ може бути записана у трьох еквівалентних формах:

$$Z_{K_i}(x_j) = \sum_{s=1}^{m} P(x_{js} / K_i) \log_2 \frac{P(x_{js} / K_i)}{P(x_{js})}; \tag{1.5}$$

$$Z_{K_i}(x_j) = \sum_{s=1}^{m} P(x_{js} / K_i) \log_2 \frac{P(K_i / x_{js})}{P(K_i)}; \tag{1.6}$$

$$Z_{K_i}(x_j) = \sum_{s=1}^{m} P(x_{js} / K_i) \log_2 \frac{P(K_i x_{js})}{P(K_i) P(x_{js})}. \tag{1.7}$$

Відзначимо, що обстеження, що має невелику діагностичну значимість для одного класу, може мати суттєву значимість для іншого. Введемо поняття загальної діагностичної значимості обстеження за ознакою x_j для всієї системи класів K, визначивши її як кількість інформації, що вноситься обстеженням у систему класів:

$$Z_K(x_j) = \sum_{i=1}^{q} P(K_i) Z_{K_i}(x_j) = \sum_{i=1}^{q}\sum_{s=1}^{m} P(K_i) P(x_{js}/K_i) \log_2 \frac{P(x_{js}/K_i)}{P(x_{js})}, \qquad (1.8)$$

де q – кількість класів (діагнозів).

Величина $Z_{K_i}(x_j)$ є очікуване (середнє) значення інформації, що може бути внесено обстеженням можливості віднесення до невідомого заздалегідь класу, що належить аналізованої системі (сукупності) класів. В іншій формі

$$Z_K(x_j) = \sum_{i=1}^{q}\sum_{s=1}^{m} P(K_i x_{is}) \log_2 \frac{P(K_i x_{js})}{P(K_i) P(x_{js})}. \qquad (1.9)$$

Загальна діагностична цінність одночасного обстеження комплексу v ознак загальна діагностична цінність буде дорівнювати:

$$Z_K(X) = Z_K(x_1 x_2 .. x_v) = \sum_{i=1}^{q}\sum_{s=1}^{m_1}\sum_{p=1}^{m_2} .. \sum_{\sigma=1}^{m_v} P(K_i x_{1s} .. x_{v\sigma}) \log_2 \frac{P(K_i x_{1s} x_{2p} .. x_{v\sigma})}{P(K_i) P(x_{1s} x_{2p} .. x_{v\sigma})}, \qquad (1.10)$$

де v – кількість ознак;

m_v – кількість розрядів v-ї ознаки.

Аналогічні співвідношення справедливі й для часткової діагностичної значимості комплексу ознак:

$$Z_{K_i}\left(x_1 x_2 .. x_v\right) = \sum_{s=1}^{m_1}\sum_{p=1}^{m_2}...\sum_{\sigma=1}^{m_v} P\left(x_{1s} x_{2p} .. x_{v\sigma} / K_i\right) \log_2 \frac{P\left(x_{1s} x_{2p} .. x_{v\sigma} / K_i\right)}{P\left(x_{1s} x_{2p} .. x_{v\sigma}\right)}. \quad (1.11)$$

Загальна діагностична значимість комплексу ознак при проведенні послідовного обстеження. Якщо проведено обстеження за комплексом ознак $X^{(n)}$ і потрібно вибрати новий комплекс ознак для одночасного обстеження $X^{(m)}$ з найбільшою діагностичною значимостю, слід виходити з величини:

$$Z_K(X^{(m)}/X^{(n)}) = Z_K(x_1 / X^{(n)}) + Z_K(x_2 / X^{(n)} x_1) + ... + Z_K(x_m / X^{(n)} x_1 x_2 .. x_{m-1}). \quad (1.12)$$

Перевагою інформаційного підходу є суворого математичне обгрунтування оцінки інформативності ознак. Недоліком є залежність багатьох методів цього підходу від розрядності ознак (необхідність квантувати значення ознак на фіксовану кількість дискретних розрядів).

1.3.3 Статистичні методи

За наявності достатнього обсягу статистичних даних може бути проведена додаткова обробка з метою кількісної оцінки інформативності сукупності наявних початкових ознак стану об'єкта.

Оцінка інформативності початкових ознак про потенційну стабільність параметрів-критеріїв придатності (параметри-критерії не виходять за встановлені межі) проводиться за допомогою апарату лінійного регресійно-кореляційного аналізу шляхом підрахунку коефіцієнтів парної або множинної кореляції між екстремальними значеннями параметра-критерію в процесі випробування (або значеннями параметра-критерію в кінці випробувань) та початковими ознаками стану об'єктів.

Інформативність ознак оцінюється шляхом аналізу одержаних коефіцієнтів парної кореляції. З початкових ознак вибираються ті, які мають

значні коефіцієнти парної кореляції з параметром-критерієм. Потім із цих ознак відбираються ті, які мають найменші коефіцієнти парної кореляції між собою. За сукупністю відібраних ознак підраховуються коефіцієнти лінійної регресії та коефіцієнт множинної кореляції цієї сукупності початкових ознак з параметром-критерієм.

Такий підхід складає основу методу відбору ознак за статистичною оцінкою інформативності.

Ще одним методом відбору інформативних ознак є метод покрокової регресії, коли незалежні змінні одна за одною включаються до підмножини згідно з попередньо заданим критерієм. У той же час деяка змінна може бути замінена іншою змінною, яка не входить до набору або видалена з нього. Сукупність критеріїв, що визначають, які змінні включати, замінювати та видаляти, називається покроковою процедурою. За стандартною покроковою процедурою (F-метод) включення та видалення змінних здійснюється за допомогою статистики критерію, а саме *t*-критерію для перевірки рівності нулю часткового коефіцієнта кореляції. У більшості пакетів статистичних програм обчислюється квадрат цієї статистики, який має *F*-розподіл і називається або *F*-включенням, або *F*-видаленням.

Змінна x_{i_1}, якій відповідає найбільше значення F-включення (або, що еквівалентно, найбільша величина квадрата коефіцієнта кореляції з *Y*), вибирається як найкращий предиктор для *Y*. Потім обчислюються відповідне рівняння найменших квадратів, таблиця дисперсійного аналізу (таблиця 1.1). та множинний коефіцієнт кореляції.

Метод головних компонент (МГК, англ.: principal component analysis, PCA) – метод факторного аналізу в статистиці, який використовує ортогональне перетворення множини спостережень з можливо пов'язаними змінними (сутностями, кожна з яких набуває різних числових значень) у множину змінних без лінійно кореляції, які називаються головними компонентами.

Таблиця 1.1 – Таблиця дисперсійного аналізу для метода покрокової регресії

Джерело дисперсії	Сума квадратів	Ступені свободи	Середній квадрат	F-відношення
Регресія	$SS_D = \sum_{i=1}^{p} b_i \sum (x_{ij} - \bar{x}_i) y_i$	$v_D = p$	$MS_D = \frac{SS_D}{v_D}$	$F = \frac{MS_D}{MS_R}$
Відхилення від регресії	$SS_R = SS_T - SS_D$	$v_R = N - p - 1$	$MS_R = \frac{SS_R}{v_R}$	
Повна	$SS_T = \sum_{i=1}^{N} (y_i - \bar{y})^2$	$v_T = N - 1$		

Задача аналізу головних компонент має чотири базових версії: апроксимувати дані лінійними многовидами меншої розмірності; знайти підпростори меншої розмірності, в ортогональній проекції на які розкид даних (тобто середньоквадратичне відхилення від середнього значення) максимальний; знайти підпростори меншої розмірності, в ортогональній проекції на які середньоквадратична відстань між точками максимальна; для даної багатовимірної випадкової величини побудувати таке ортогональне перетворення координат, внаслідок якого кореляції між окремими координатами перетворяться в нуль.

Перші три версії оперують скінченними множинами даних. Вони еквівалентні і не використовують жодної гіпотези про статистичне породження даних. Четверта версія оперує випадковими величинами. Скінченні множини з'являються тут як вибірки з даного розподілу, а розв'язання трьох перших завдань – як наближення до фактичного розкладу. При цьому виникає додаткове і не цілком тривіальне питання про точність цього наближення.

Розглянемо застосування методу головних компонент для створення спрощеної моделі, яка повинна максимально точно описувати вихідні данні. Досить часто чинники мають значну залежність один від одного та їх одночасна наявність є надлишковою. Деякі ознаки/фактори мають неявну залежність. Знаючи залежності та їх силу за допомогою методу головних компонент можна виразити декілька чинників через один, тобто об’єднати, і

працювати з більш простою моделлю. Уникнути втрати інформації при цьому неможливо, але метод головних компонент надає можливість їх мінімізувати. Даний метод апроксимує *n*-мірний простір («хмару») спостережень до еліпсоїда, піввісі якого будуть майбутніми головними компонентами і при проекції на такі осі (зменшення розмірності) зберігається найбільша кількість інформації [12, 13]. Метод головних компонент заснований на конструюванні чинників (як метод відбору інформативних ознак). Головна ідея методу полягає в об’єднанні декількох корельованих змінних в одну, яка представлятиме собою лінійну комбінацію вихідних змінних. Метою методу є дослідження внутрішньої структури досліджуваної системи величин, стиснення цієї системи без істотної втрати інформації шляхом виявлення невеликого числа чинників, які пояснюють мінливість і взаємозв'язок величин [14].

Метод Random forest (англ. випадковий ліс) – ансамблевий метод машинного навчання для класифікації, регресії та інших завдань, який працює за допомогою побудови численних дерев прийняття рішень під час тренування моделі й продукує моду для класів (класифікацій) або усереднений прогноз (регресія) побудованих дерев [15]. Недоліком є схильність до перенавчання. Випадкові ліси, отримані в результаті застосування методу RF, можуть бути природним чином використані для оцінки важливості змінних в задачах регресії, класифікації та інтелектуальної діагностики.

Перевагою статистичного підходу є урахування статистичного зв'язку між ознаками та вихідним параметром, а також урахування характеристик розподілу даних величин.

Недоліком статистичного підходу є відносна складність реалізації та тривалість роботи окремих методів, пов'язані з необхідністю обчислення великої кількості різних статистичних характеристик.

1.3.4 Імовірнісні методи

Порівняльна оцінка інформативності ознак може бути зроблена навіть у разі, коли умовні щільності їх розподілів $f_i(x_j)$ невідомі, проте відомі перші та другі моменти цих розподілів. Оцінка, заснована на використанні цих даних, можлива у зв'язку з тим, що ознаки x_j можуть бути умовно поділені на дві групи.

До першої групи належать ознаки, значення яких змінюються при переході від одного об'єкта даного класу до іншого об'єкту і дуже помітно змінюються при переході від об'єкта одного класу до об'єктів інших класів.

До другої групи належать ознаки, значення яких чутливі до переходів від одного об'єкта даного класу до іншого об'єкта і лише трохи змінюються при переходах від об'єктів одного класу до об'єктів інших класів.

Ознаки, що належать до першої групи, корисніші за ознаки, що належать до другої групи. Кількісна оцінка інформативності ознак x_j, $j = 1, 2, ..., n$ може бути проведена за порівнянням компактності розташування об'єктів відносно осі досліджуваної ознаки.

Такий підхід складає основу методу відбору ознак за порівнянням їх імовірнісних характеристик. До класу імовірнісних методів відбору ознак належить також метод порівняння апостеріорних імовірностей, згідно з яким розраховуються апостеріорні імовірності віднесення ознак до різних класів інформативності.

Перевагою імовірнісного підходу є урахування імовірнісного зв'язку між ознаками та вихідним параметром. Недоліком розглянутого підходу є необхідність знання апріорних ймовірностей класів та попадання значень ознак у певні інтервали.

1.3.5 Нейромережеві методи

Перспективними для вирішення задачі оцінки інформативності та

відбору ознак є штучні нейронні мережі (ШНМ), які здатні отримувати знання з даних [3].

Зокрема моделями ШНМ, що часто використовуються, для вирішення завдань розпізнавання образів є багатошарові перцептрони (БШП).

Ступінь впливу *i*-ї ознаки на вихідний сигнал *y* для одношарового перцептрону визначається, в основному, значеннями ваг і ознак. Для БШП, що має один вихід, тобто містить на останньому шарі тільки один нейрон, ступінь впливу *i*-ї ознаки x_i, значення якої надходять на входи нейронів першого шару БШП, вихідний сигнал у також, в основному, визначатися значеннями ваг і ознак БШП.

Кількісна оцінка інформативності (значущості) *i*-ї ознаки буде визначатися як сума інформативностей *i*-х входів нейронів першого шару БШП. Для оцінки інформативності входів БШП необхідно оцінити не лише їх значимість щодо нейрона, а й значимість виходу нейрона як входу нейронів наступного шару і далі до нейронів останнього шару. Очевидно, що реалізувати подібну процедуру простіше у зворотному порядку, рухаючись від останнього шару до першого, оскільки обчислення значимості входів нейрона останнього шару буде здійснюватися аналогічно одношаровому перцептрону, а значимості входів нейронів попереднього шару можна буде визначити як добуток часткових значимостей входів відносно виходів та значимості виходів, як входів нейронів наступного шару.

Для реалізації такого підходу можна використовувати процедуру, подібну до відомого алгоритму зворотного поширення помилки (Error Backpropagation Method), що використовується при навчанні БШП.

Метод зворотної оцінки значимості ознак на основі БШП наведено в роботі [11].

Згідно з результатами цієї роботи можна припустити, що цей метод дозволяє достатньо ефективно визначати якісну оцінку значущості діагностичних ознак.

Підходи до відбору інформативних діагностичних ознак, розглянуті у

цьому підрозділі, були проаналізовані з урахуванням низки критеріїв. Порівняльна характеристика деяких з розглянутих методів відбору інформативних ознак наведена у таблиці 1.2.

Таблиця 1.2 – Порівняльна характеристика методів відбору інформативних ознак

Критерії порівняння	Методи відбору ознак				
	M1	M2	M3	M4	M5
математичне обгрунтування	низьке	високе	середнє	середнє	середнє
відносна складність реалізації	низька	середня	висока	низька	висока
відносна швидкість роботи	низька	висока	низька	висока	середня
вимоги до даних	низькі	високі	низькі	низькі	середні

В таблиці 1.2 використано такі позначення: M1 – метод комбінованого перебору (евристичний підхід); M2 – метод стиснення інформації з групуванням ознак (інформаційний підхід); M3 – метод головних компонентів (статистичний підхід); M4 – методу відбору ознак за порівнянням їх імовірнісних характеристик (імовірнісний підхід); M5 – метод зворотної оцінки значимості ознак на основі БШП (нейромережевий підхід). Як очевидно з табл. 1.2, жоден з розглянутих підходів не має абсолютної переваги перед усіма іншими. Тому на практиці слід використовувати той підхід, який краще відповідає найважливішій вимозі (критерію), що пред'являється до розв'язуваного завдання.

В кваліфікаційній роботі, за результатами порівняння можливостей методів редукції і вимог, що вони висувають, для побудови та досліджень були обрані редукційні методи стиснення інформації з групуванням ознак, метод головних компонентів (PCA) та нейромережевий метод зворотної

оцінки значимості ознак на основі БШП. Слід відзначити, що ці методи є зручні для розробки алгоритмів редукції та програмної реалізації, а при їх реалізації комп'ютерній реалізації не потрібне використання додаткових ресурсів для зберігання та обробки даних.

1.4 Постановка задач дослідження

Метою кваліфікаційної роботи є розробка та дослідження алгоритмічних та програмних засобів, які реалізують методи редукції даних для інтелектуальних систем діагностики, що дозволяє зменшувати кількість діагностичних ознак та розмір вибірок.

Згідно з метою роботи були поставлені наступні задачі:

– аналіз існуючих джерел за темою дослідження;

– дослідження методу стиснення інформації з групуванням ознак;

– розробка модуля редукції даних для діагностичної системи;

– моделювання досліджуваних методів.

2 МОДИФІКАЦІЯ МЕТОДУ СТИСНЕННЯ ІНФОРМАЦІЇ З ГРУПУВАННЯМ ОЗНАК

2.1 Опис базового методу стиснення інформації з групуванням ознак

Нехай є n параметрів (ознак об’єкту), що контролюються і залежать від часу $x_i=x_i(t)$, де i=1,2, … , n; $0 \le t \le T$, які описують зміни параметрів в часі в кожній точці i. Час t приймає дискретні значення $t_p=t_1, t_2, \dots, t_k$; $t_1=0$; $t_k=T$, p=1, 2, … , k.

Отже, параметр x_i можливо представити у вигляді вектору $x_i=\{x_i(t_1), x_i(t_2), \dots, x_i(t_k)\}$, де кожен компонент вектору описує значення i-го параметра у фіксовані моменти часу, тобто утворює екземпляр вибірки.

Переважна більшість параметрів контролю пов’язані за якимось правилом чи функцією, і цей зв’язок може бути більш або менш тісним. В якості міри зв’язку у такому разі можливо використовувати співвідношення: відстані та ступеня зв’язку.

Поняття відстані $R(x_i, x_j)$ між двома параметрами x_i і x_j для простору X можливо представити у вигляді формули [1]:

$$R(x_i, x_j) = \sqrt{\sum_{p=1}^{k} (x_i^p - x_j^p)^2}, i \neq j, \tag{2.1}$$

де $R(x_i, x_j)$ – матриця показників відстані;

x_i, x_j – параметри між якими обчислюється відстань;

p – кількість моментів часу.

Таким чином, виконується побудова матриці відстаней, що виключає головну діагональ ($i \neq j$). Для визначення ступеня зв’язку між параметрами x_i і x_j можливо використати потенціальну функцію, що буде залежати від відстані та спадатиме з її зменшенням [11]:

$$K(x_i, x_j) = \frac{1}{1 + \alpha\, R^2(x_i, x_j)}\ , \qquad (2.2)$$

де $K(x_i, x_j)$ – матриця показників ступеня зв'язку між параметрами;

$R(x_i, x_j)$ – матриця показників відстані;

α – вільний коефіцієнт, від якого залежить швидкість зменшення потенціальної функції (задається в межах $0 \le \alpha \le 1$).

Отже, якщо із розгляду фізичної природи об'єкту випливає існування зв'язків між параметрами, то цілком логічним буде висновок, що існують і деякі сукупності (групи) взаємопов'язаних між собою параметрів. Звідси можливо вивести припущення, що параметри найбільш сильно пов'язані між собою в тому випадку, якщо вони підпорядковані однаковим впливам та зовнішнім факторам.

Прийнявши гіпотезу, за якою параметри, що належать до однієї групи, можуть бути сильно пов'язані між собою, а параметри які належать до різних груп пов'язані між собою відносно слабкими зв'язками, можливо побудувати розбиття усієї сукупності параметрів на певні групи.

Таким чином, загальна послідовність дій в методі з виконання процесу групування параметрів є наступною:

– розрахунок матриці зв'язків між параметрами;
– побудова за нею сукупності зв'язаних параметрів;
– виконання розбиття цієї сукупності на окремі групи параметрів.

Після розрахунку матриці зв'язків необхідно за нею сформувати сукупність параметрів, на якій надалі буде виконуватися розбиття на групи. Для створення такої послідовності параметрів можливо використати одну з багатьох можливих варіацій алгоритмів групування, а саме – метод групування ознак за допомогою графу груп. Сутність методу групування ознак полягає в тому, що за матрицею зв'язків між параметрами будується граф, який формальними прийомами за значенням межі зв'язку роздроблюють на окремі підграфи. Ці підграфи і є групами взаємопов'язаних

параметрів.

Заключним етапом в наведеному методі стиснення інформації є, власне, саме стиснення інформації (скорочення її обсягу). Ефект стиснення створюється за рахунок виключення з груп визначеної кількості параметрів згідно методики обробки групи або в залежності від зв'язків між параметрами в групі. Послідовність дій виконання є наступною:

– обробка груп параметрів з виключенням тих, що є не досить цінними за обраним критерієм оцінки, або оцінка їхньої інформативності нижче встановленого значення межі;

– визначення за сукупністю параметрів, що залишилися, обсягу параметрів, який не підлягає виключенню;

– вилучення з вхідних даних зайвої інформації, формування скороченої за обсягом вибірки.

2.2 Групування ознак

Як було зазначено вище, метод групування параметрів (ознак) дозволяє виконувати групування параметрів за ступенем зв'язку між ними. Побудова сукупності параметрів відбувається у вигляді графа, в вузлах якого розміщують значення параметрів, а ребрам графа відповідають значення з матриці зв'язків між параметрами. Граф будується за наступною методикою:

– існує матриця зв'язків, що містить *n* параметрів (ознак). В матриці виконується пошук максимального недіагонального члену – $\rho(x_i, x_j)=\rho_1$, ($i \neq j$). Номера i, j, утворюють початок графа. На ребрі, що поєднує параметри x_i та x_j, ставиться значення їхнього зв'язку ρ_1;

– в матриці виконується пошук найбільшого після ρ_1 члену в *j*-му стовпці і найбільшого після ρ_1 члену в *i*-му рядку матриці зв'язків. Найбільший з них поєднується з x_i, а над ребром ставиться зв'язок $\rho(x_i, x_j)=\rho_2$;

– знаходяться параметри, що найбільш тісно пов’язані з двома останніми параметрами, що розглядалися (пов’язані з x_i зв’язками ρ_1, ρ_2). З їхніх коефіцієнтів зв’язку обирається найбільший і утворює ребро зв’язку ρ_3;

– побудова графа продовжується спочатку, за матрицею зв’язків знаходиться параметр, що найбільш тісно пов’язаний з одним з двох параметрів, що розглядалися на попередньому кроці. Побудова завершується після того, коли число елементів графа буде дорівнювати n (тобто оброблені усі параметри).

Тобто після утворення, граф буде відображувати зв’язок між усіма значеннями параметрів вибірки (рисунок 2.1).

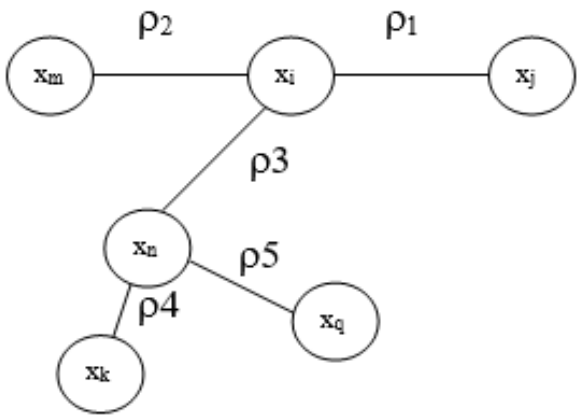

Рисунок 2.1 – Приклад графа, що будується за методом групування ознак

При комп’ютерній реалізації цього методу його можливо суттєво спростити, виконавши групування параметрів до одного вектору (масиву записів із збереженням інформації пор їхні зв’язки). Зробити таке групування параметрів можливо наступним чином:

– на кожному кроці знаходяться два найбільших коефіцієнта, що пов’язують граничні (крайні) параметри графу, які були отримані на попередньому кроці;

– з цих двох коефіцієнтів обирається найбільший, і граф доповнюється параметром, який відповідає найбільшому коефіцієнту зв’язку;

– дії повторюються доти, доки не вичерпається кількість параметрів.

Приклад графу, що будується в такий спосіб, наведено на рисунку 2.2.

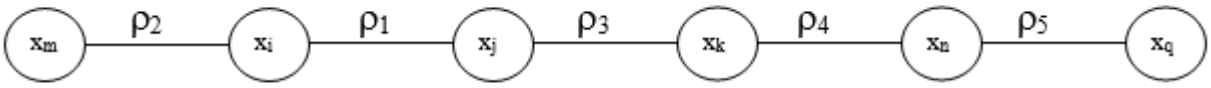

Рисунок 2.2 – Приклад графа, що будується за спрощеним методом

Отже, виконавши групування параметрів за ступенем їхнього зв’язку, можливо виконати їхнє розподілення. Для цього в методі застосовують граничне значення коефіцієнта зв’язку – межі. Змінюючи значення коефіцієнта зв’язку можливо корегувати кількість елементів у групах або кількість самих груп.

Розбиття графа, відбувається на тих ділянках де зв’язок між параметрами менше ніж граничний коефіцієнт зв’язку.

Таким чином, за методом виконується групування параметрів до певної структури (графу) за ступенем зв’язку між ними, після чого виконується розбиття графа на підграфи (групи) в місцях де зв’язки є меншими ніж припустимо за граничним коефіцієнтом зв’язку.

2.3 Критерії оцінки ознак в групах

Під критеріями оцінки параметрів (ознак) слід розуміти показники, що надають інформацію про ступені зв’язків між параметрами, про значення інформативності параметру. Параметри, для яких за розрахунками інформативність або критерій оцінки для групи малий, з неї виключаються.

Оцінка модуля коефіцієнту парної кореляції є найбільш поширеним способом оцінки ступеня зв’язку параметрів. Інформативність параметрів за критерієм модуля кореляції можливо оцінити наступним чином. Позначимо:

S – кількість екземплярів у виборці, s – номер поточного екземпляра вибірки, N – кількість параметрів (ознак), y^s – цільове значення, що зіставляється s-му екземпляру вибірки (номер класу або числова характеристика). Модуль коефіцієнта парної кореляції (I^A_i) розраховується за формулою [16]:

$$I_i^A = \left| \frac{\sum_{s=1}^{S}(x_i^s - \bar{x}_i)(y^s - \bar{y})}{\sqrt{\sum_{s=1}^{S}(x_i^s - x_i)^2 \sum_{s=1}^{S}(y^s - \bar{y})^2}} \right|, \qquad (2.3)$$

де x^s_i – значення i-ї ознаки s-го екземпляра вибірки;

$\bar{x}_i$, $\bar{y}$ – середнє арифметичне значень i-ї ознаки і цільового параметра відповідно; i=1…N;

N – кількість ознак;

S – кількість екземплярів.

Критерії оцінки параметрів для групи можливо представити за допомогою непараметричних показників зв’язку між параметрами: коефіцієнта кореляції Фехнера, коефіцієнта кореляції знаків, коефіцієнта асоціації та коефіцієнта Юла.

Коефіцієнт кореляції Фехнера заснований на принципі обліку знаків відхилень варіант від їхніх середніх арифметичних, і розраховується за формулою [17]:

$$r_{\Phi} = \frac{C - H}{C + H}, \qquad (2.4)$$

де C – число збігів однакових, як позитивних, так і негативних знаків різниць $(x_i - \bar{x})$ та $(y_i - \bar{y})$;

H – число незбіжних знаків, відповідно.

При використання коефіцієнту Фехнера у якості критерія оцінки для параметрів, можливо користувати лише значенням модуля цього коефіцієнту.

Коефіцієнт кореляції знаків можливо розрахувати наступним чином [5]:

$$R_{XY} = \frac{P(XY) - P(X)P(Y)}{\sqrt{P(X)P(Y)(1 - P(X)(1 - P(Y))}} , \qquad (2.5)$$

де P(XY) – число збігів позитивних знаків в загальній серії випробувань, віднесене до їх числа n;

P(X), P(Y) – частковість позитивних знаків для кожної ознаки .

Ступень тісноти зв’язку між якісними параметрами Y та X, що групуються до кореляційної таблиці Пірсона, можливо визначити за формулою коефіцієнта асоціації [5, 8]:

$$r_A = \frac{a\,d - b\,c}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d))}} , \qquad (2.6)$$

де a, b, c, d – чисельність груп (варіант) за таблицею.

Коефіцієнт асоціації Юла також розраховують за значеннями таблиці Пірсона [18]. Розрахунок коефіцієнта виконується за формулою:

$$r_Q = \frac{a\,d - b\,c}{a\,d + b\,c} , \qquad (2.7)$$

де a, b, c, d – чисельність груп (варіант) за таблицею.

Значення коефіцієнтів асоціації, як і коефіцієнт Фехнера, можливо для зручності при практичному користуванні обирати за модулем. Отже, після виконання розрахунків за вищенаведеними коефіцієнтами, можливо визначити ті ознаки в групах, що не є достатньо пов’язані з іншими членами групи, цільовим параметром або є малоінформативними. Такі ознаки цілком

можливо виключити з кожної групи, а потім вилучити з вхідних даних їхню інформацію, скорочуючи таким чином обсяг діагностичних даних.

2.4 Модифікований метод стиснення інформації з групуванням ознак

Схему алгоритму редукції даних, що реалізує модифікований метод стиснення інформації згрупуванням ознак, наведено на рисунку 2.3.

Сутність модифікації процедури редукції полягає в застосуванні наведеного вище спрощеного способу побудови графу звязків між ознаками та використанні багатошарового персептрону (БШП) для оцінювання інформативності ознак, сформованих за модифікованим методом.

Мережі типу БШП найбільш вдало реалізують ідею керованого навчання. Для управління навчанням необхідно мати дані, які мають відому класифікацію і обсяг, достатній для проведення навчання та тестування БШП. Крім того, перед подачею даних в БШП бажано провести попередню обробку.

Середньоквадратична похибка для багатошарових мереж набагато складніша, ніж для одношарових, і має багато локальних мінімумів. Якщо алгоритм зворотного поширення збігається, то це ще не означає, що отриманий розв’язок є оптимальним. Щоб гарантувати оптимальність отриманого розв’язку, необхідно виконати декілька спроб з різними початковими умовами. У більшості випадків багатошарова мережа пов’язується із множиною прикладів її навчання, яка, зазвичай, подана великою кількістю можливих пар «вхід-ціль». Важливо, щоб мережа вміла узагальнювати одержані дані.

Фрагмент програмної реалізації модифікованого методу наведено в додатку А. Загальну структуру БШП наведено на рисунку 2.4.

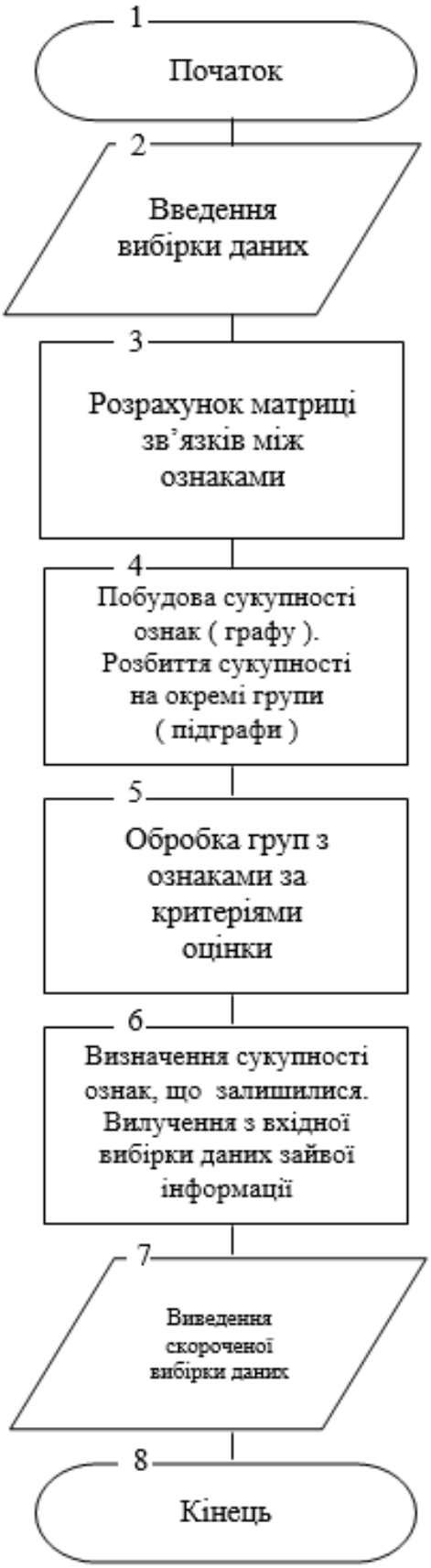

Рисунок 2.3 – Схема алгоритму редукції даних за модифікованим методом стиснення інформації з групуванням ознак

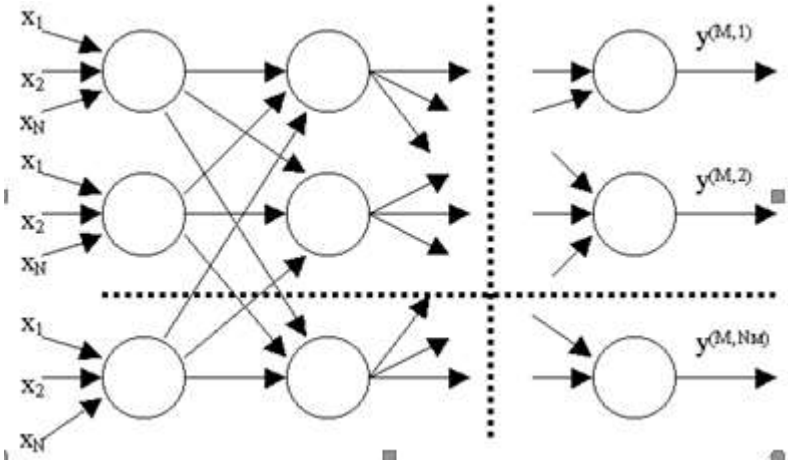

Рисунок 2.4 – Загальна структура БШП

Кожен і-й нейрон μ-го шару (μі-й нейрон) перетворює вхідний вектор x(μ,i) у вихідну скалярну величину y(μ,i). Це перетворення складається з двох етапів: спочатку обчислюється дискримінантна функція net(μ,i), яка далі перетворюється на вихідну величину y(μ,i).

Реалізацію методу рекомендується починати за допомогою двошарового персептрону, перший шар якого містить 4 нейрони, а другий шар 1 нейрон. При цьому усі нейрони мають сигмоїдну функцію активації ψ(x)=1/(1+e-x).

На етапі навчання на входи БШП подаються значення факторів з навчальної виборки для конкретних об’єктів діагностики.

Ступінь впливу і ознаки на вихідний сигнал для перцептронів визначається, в основному, значеннями ваг і ознак. Для БШП, що має один вихід, тобто містить на останньому шарі тільки один нейрон, ступінь впливу і ознаки, значення якої надходять на входи нейронів першого шару БШП, на вихідний сигнал також буде визначатися значеннями ваг та ознак БШП.

Кількісна оцінка інформативності (значущості) *i*-ї ознаки буде визначатися як сума інформативностей *i*-х входів нейронів першого шару БШП. Для оцінки інформативності входів БШП необхідно оцінити як їх значимість щодо нейрона, так і значимість виходу нейрона як входу нейронів

наступного шару до нейронів останнього шару. Очевидно, що реалізувати подібну процедуру простіше у зворотному порядку, рухаючись від останнього шару до першого, оскільки обчислення значимості входів нейрона останнього шару буде здійснюватися аналогічно одношаровому перцептрону, а значимості входів нейронів попереднього шару можна буде визначити як добуток приватних значимостей , як входи нейронів наступного шару.

Для реалізації такого підходу будемо використовувати процедуру, подібну до широко відомого алгоритму зворотного поширення помилки (Error Backpropagation Method), що використовується при навчанні БШП.

При побудові моделей складних об'єктів та процесів за точковими даними на основі БШП виникає завдання розбиття вихідної вибірки даних на навчальну та контрольну.

Традиційно навчальну та контрольну вибірки виділяють за допомогою випадкових чисел або лінійним розбиттям вихідної вибірки. Однак при цьому в контрольну вибірку можуть потрапити такі екземпляри, які сильно віддалені в метричному просторі ознак екземплярів, що потрапили в навчальну вибірку. У цьому випадку модель, побудована за навчальною вибіркою, погано працюватиме навіть для контрольної вибірки, не кажучи вже про адекватність цієї моделі об'єкту, що досліджується.

Тому виникає завдання використання алгоритму, що дозволяє розбивати вихідну вибірку таким чином, щоб навчальна вибірка містила всі екземпляри, що знаходяться в вузлових точках, представлені у вихідній вибірці, а контрольна вибірка містила екземпляри близькі (в сенсі відстані) до відповідних екземплярів навчальної вибірки.

З іншого боку, під час навчання нейронних мереж однією з найважливіших завдань є скорочення часу навчання мережі при забезпеченні заданого рівня точності прогнозування, що може бути досягнуто за рахунок скорочення навчальної вибірки шляхом видалення з неї екземплярів, що не перебувають у вузлових точках (тобто надлишкових) прикладів).

Для вирішення обох вищеописаних завдань пропонується

використовувати наступний алгоритм розбиття вихідної вибірки на навчальну та тестову.

Крок 1. Ініціалізація параметрів алгоритму розбиття вихідної вибірки. Задати вихідну вибірку екземплярів ххх. та зіставлені ним номери класів або значення прогнозованого параметра, а також L – кількість розбиття вихідної вибірки. Занести змінну N – кількість ознак, що характеризують екземпляри, а змінну M – кількість екземплярів вихідної вибірки. Для завдань класифікації прийняти ширину допустимого інтервалу варіації прогнозованого параметра dt = 0, для задач чисельної оцінки прогнозованого параметра прийняти dt = | max(твих.) – min(твих.) | / L де min(a) і max(a) – мінімальне і максимальне значення вектора a, відповідно. Встановити лічильник newind=1.

Крок 2. Обчислити відстані між екземплярами вихідної вибірки:

$$R(p,q)=\begin{cases}\sqrt{\sum_{i=1}^{N}\left(x_i^p-x_i^q\right)^2}, p\neq q;\\ RealMax, p=q,\end{cases} \qquad p = 1,..., M,\ q = 1,.., M \qquad (2.8)$$

де RealMax – максимальне число, що оброблюється в процесорі;

x^p_i –значення i-ї ознаки p-го екземпляру.

Крок 3. Знайти в матриці відстаней R мінімальний елемент minx та його індекси q і p, а також максимальний елемент maxx, за умови, що при знаходженні мінімуму та максимуму тут і далі ігноруються елементи, що дорівнюють RealMax.

Крок 4. Прийняти a = | maxx – minx | /(2L).

Крок 5. Якщо minx<Realmax, то перейти на крок 6, інакше – перейти на крок 13.

Крок 6 Прийняти: $x_{нав}$ (newind) = $x_{вих}$ (q), $t_{нав}$ (newind) = $t_{вих}$ (q), де $x_{нав}$ і $t_{нав}$. – масиви екземплярів навчальної вибірки та зіставлених їм прогнозованих значень, відповідно. Встановити: newind = newind + 1, значення мінімального поточного елемента в рядку teck=R(q,p). Знайти мінімальний елемент mminx

та його індекси mqmin та mpmin серед елементів *q*-го рядка матриці *R*.

Крок 7. Якщо mminx<Realmax, то перейти на крок 8, інакше – перейти на крок 11.

Крок 8. Встановити значення покажчика віддаленого екземпляра зі стовпця deleted = 0 (у матриці *R* нумерація рядків та стовпців має починатися з 1).

Крок 9. Якщо |teck – *R*(q,mpmin)| ≤ a, то перейти крок 10, інакше прийняти: deleted = mpmin, *R*(q,mpmin) = Realmax, R(mpmin,q) = Realmax і перейти крок 11.

Крок 10. Якщо $|t_{вих}$ (mpmin) – $t_{вих}$ (*q*)| ≤ a, то прийняти: *R*(v, mpmin) = realmax, R(mpmin, v) = realmax, *v*=1,.., *M*, інакше – прийняти : $x_{нав}$ (newind) = $x_{вих}$ (mpmin), $t_{нав}$ (newind) = $t_{вих}$ mpmin), newind = newind+1, R(v, mpmin) = Realmax, R(mpmin, v) = Realmax, *v*=1,.., *M*.

Крок 11. Знайти мінімальний елемент mminx та його індекси mqmin та mpmin серед елементів *q*-го рядка матриці *R*.

Крок 12. Перейти до кроку 7.

Крок 13. Прийняти: *R* (*v*, *q*) = Realmax, *R* (*q*, *v*) = Realmax, *v*=1,.., *M*, покажчик віддаленого екземпляра з рядка dstr = *q*. Знайти у матриці відстаней *R* мінімальний елемент minx та його індекси *q* та *p*.

Крок 14. Перейти до кроку 5.

Крок 15. Якщо (deleted dstr) і (deleted > 0), тоді прийняти: $x_{нав}$ (newind) = $x_{вих}$ (deleted); $t_{нав}$ newind) = $t_{вих}$ (deleted).

Крок 16. Зупинка.

В результаті виконання даного алгоритму для вихідної вибірки $x_{вих}$ і зіставленого їй набору значень $t_{вих}$ ми отримаємо навчальну вибірку $x_{нав}$ і зіставлений їй набір значень $t_{нав}$. Залишок екземплярів з $x_{вих}$ і $t_{вих}$, що не ввійшли в $x_{нав}$ і $t_{нав}$, складе контрольну вибірку.

3 МОДУЛЬ РЕДУКЦІЇ ДАНИХ ДЛЯ ДІАГНОСТИЧНОЇ СИСТЕМИ ТА РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

3.1 Структура та функції пропонованого модуля редукції даних

Пропонована структура модуля редукції даних має не лише бути складовою частино діагностичної системи, але також використовуватися автономно від системи, як самостійний засіб. Згідно з результатами попередніх розділів, модуль редукції даних має виконувати наступні функції:

– виконувати введення зовнішніх даних (вибірок), що надходять до неї від інших модулів діагностичної системи (або з інших джерел при автономному використанні підсистеми);

– здійснювати обробку введених даних за математичними методами редукції (методом стиснення інформації з групуванням ознак) з використанням БШП;

– виконати обробку результатів роботи математичних методів;

– формувати і виводити оброблені вибірки даних.

Модуль також має доповнюватися інтерфейсом користувача. Призначення модуля інтерфейсу користувача – надавати користувачу можливість керувати процесом редукції, здійснювати вибір вхідних та збереження вихідних даних, візуально представляти роботу підсистеми, надавати допомогу користувачеві тощо.

Таким чином, структурна схема модуля має містити: блок введення–виведення даних, блок математичної обробки даних, блок інтерфейсу користувача (рисунок 3.1).

Блок введення–виведення даних призначений для виконання дій з даними, що вводяться до підсистеми або виводяться з неї у певному форматі. Модуль виконує введення даних до підсистеми і виведення даних, що надходять від блоку математичної обробки після здійснення редукції.

Блок математичної обробки даних – це основний блок модуля редукції. Його призначення виконувати редукцію даних і робити обробку результатів редукції. Блок поєднує у собі математичну і практичну частини; математична обчислює яку саме вхідну інформацію і як можливо спростити, а практична вилучає зайву (згідно розрахунків) інформацію з вхідних даних.

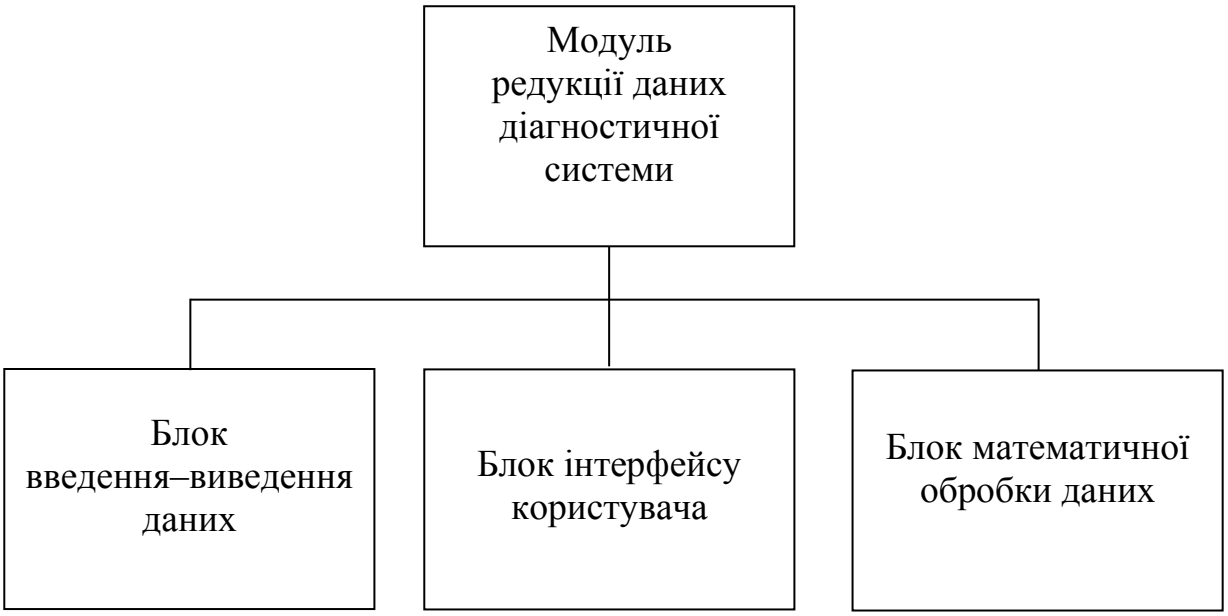

Рисунок 3.1 – Структурна схема модуля редукції діагностичної системи

3.2 Програмна реалізація та інтерфейс користувача модуля редукції даних

Згідно з структурною схемою (рисунок 3.1) програму, що реалізує функції модуля редукції даних, доцільно поділити на 3 програмних модуля, що надають відповідно графічний інтерфейс користувача, виконують введення-виведення та математичну обробку даних.

Модуль інтерфейсу користувача розроблюється як модуль-оболонка, до якого приєднуються після автоматичного завантаження засобами ОС модулі введення-виведення та математичної обробки даних. Виконання програми в такому вигляді надасть усі переваги модульної розробки, а саме:

чітке розподілення функціональних частин, можливість вдосконалення модулів введення-виведення та математичної обробки даних без зміни модуля інтерфейсу користувача, можливість використання декількох варіантів модулів (наприклад, різних за версіями реалізації і можливостями) під керівництвом одного модуля інтерфейсу користувача.

Зміст модуля інтерфейсу користувача: керівні процедури виклику інтерфейсів модулів введення-виведення та математичної обробки даних, процедури взаємодії з користувачем, внутрішні процедури і дані.

Зміст модуля введення-виведення даних: процедури інтерфейсу модуля, внутрішні процедури виконання введення та виведення даних підсистеми.

Зміст модуля математичної обробки даних: процедури інтерфейсу модуля, процедури редукції даних (реалізація методу стиснення інформації з групуванням ознак), внутрішні процедури і дані.

Отже, програмна реалізація модуля редукції даних діагностичної системи буде складатися з однієї програми-додатку для ОС Windows (модуля інтерфейсу користувача) та 2-х модулів (відповідно, введення-виведення та математичної обробки даних), що виконуються у вигляді модулів- бібліотек динамічної компоновки DLL.

Схема взаємодії програмних модулів при роботі модуля редукції даних наведена на рисунок 3.2.

Для виконання розробки програмної реалізації модуля редукції даних діагностичної було обрано інтегроване програмне середовище розробки (ІСР) програм – додатків для ОС Windows – Inprise (Borland) Delphi 5.0 Enterprise Edition (мова Object Pascal). Як альтернативні засоби для розробки було розглянуто наступні ІСР та мови програмування: Microsoft Visual Basic 6.3, Microsoft Visual C++ 6.0, пакет MATLAB 5.0.

Вибір на користь ІСР Delphi зроблено, тому що в Delphi є достатні можливості щодо розробки, відлагодження і тестування програм; засоби мови Object Pascal дозволяють ефективно використовувати динамічний розподіл

оперативної пам’яті, керувати виділенням пам’яті при використанні масивів даних, змінних за обсягом (динамічних масивів); швидкість виконання обчислень в програмах створених в Delphi є достатньою для обробки великих обсягів даних за обраним методом редукції [6].

Вбудовані в ІСР Delphi засоби розробки програмних модулів (DLL модулів) значно спрощують і прискорюють процес розробки і відлагодження модульних програм, що в даному випадку при виконанні модульної розробки діагностичної системи є дуже важливим.

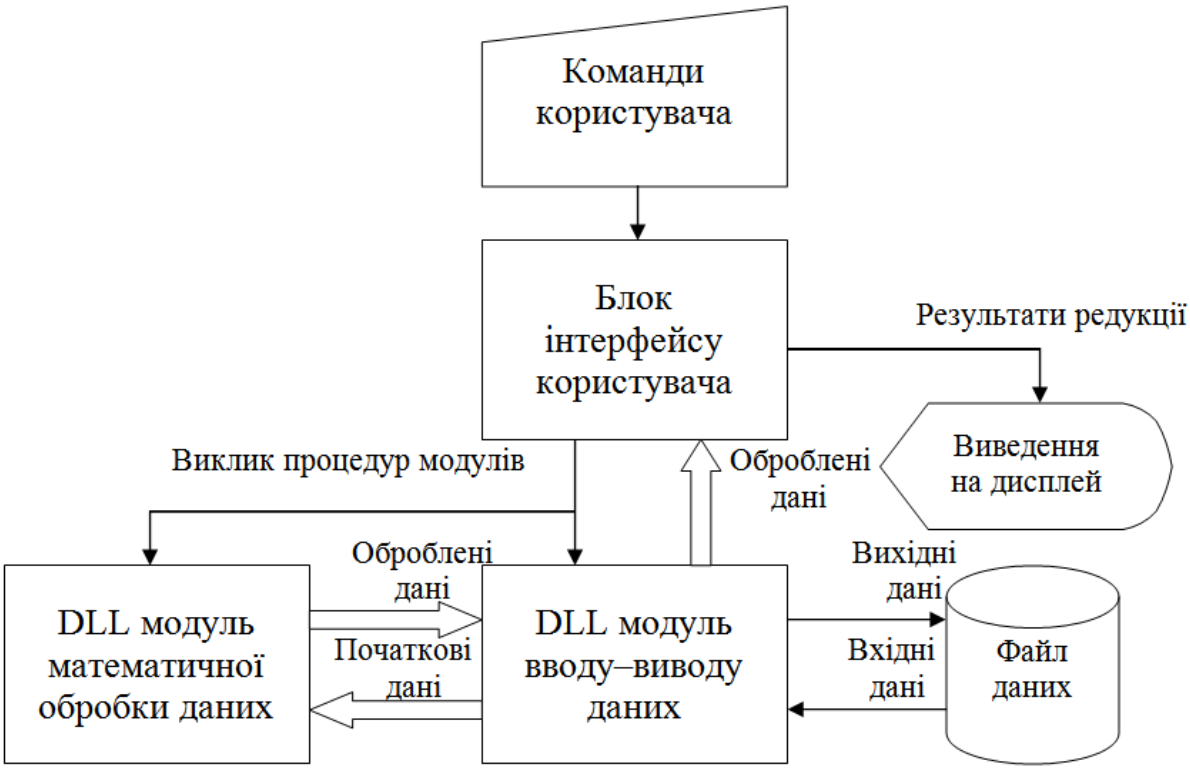

Рисунок 3.2 – Схема взаємодії програмних модулів

Вхідні та вихідні дані передаються до модуля та виводяться з нього у вигляді текстових файлів спеціального формату в кодуванні ASCII. Формат файлів даних та інтерфейсів модулів DLL для модуля редукції наведено в табл. 3.1, де прийнято такі позначення:

– R_ММО – модуль математичної обробки даних;

– R_IMO – модуль вводу-виводу даних;

– Z – масив керуючих редукцією коефіцієнтів (α та інші);

– Mth – масив номерів методів відбору з групи;

– M1 – вибірка ознак (масив);

– M2 – вектор цільових параметрів.

Таблиця 3.1 – Формат файлів даних та інтерфейсів модулів DLL для модуля редукції

Вхідний-вихідний файл з вибіркою			
	Екземпляр 1	...	Екз.-р S
Ознака 1	Значення	x	x
...	x	x	x
Ознака N	x	x	Значення
Цільовий параметр	Значення	...	Значення

Модуль DLL	Інтерфейс модуля
R_MMO	MMO->Init_Set(Z,Mth)
	MMO->StatusInfo() -> info
	MMO->Reduction(M1,M2)-повертає M1, M2
	MMO->AllFreeEnd()
R_IOM	IOM-> StatusInfo() -> info
	IOM-> GetIn(M1,M2, filename)
	IOM-> PushOut(M1,M2, filename)

Розробка інтерфейсу користувача – це дуже важлива стадія процесу розробки підсистеми, від того як буде взаємодіяти з підсистемою користувач, залежить ефективність роботи модуля редукції даних в цілому. Блок інтерфейсу користувача має надавати наступні можливості,щодо керування програмою і відображення інформації:

– мати сторінку з інформацією про модулі підсистеми;

– мати сторінку з засобами введення, зберігання і відображення даних з файлів вибірок;

– мати засоби керування процесом редукції за модифікованим методом стиснення інформації з групуванням ознак;

– мати інформаційну сторінку з допомогою для користувача;

– оскільки передбачається можливість застосування модуля редукції даних автономно від діагностичної системи у вигляді самостійної програми,

інтерфейс програми має бути розроблений з можливістю зміни людської мови спілкування з користувачем.

На сторінках мають бути розміщені інформаційні поля і таблиці для відображення текстової інформації і даних, а також стандартні для ОС Windows засоби керування програмою. Приклади сторінок форм інтерфейсу наведено на рисунках 3.3–3.5.

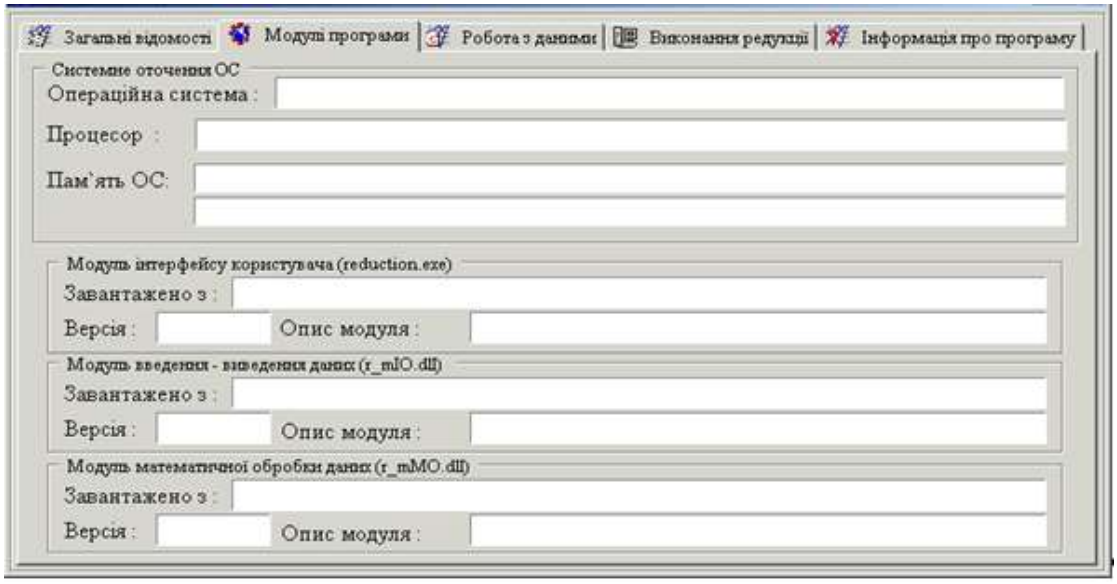

Рисунок 3.3 – Сторінка форми інтерфейсу «Модулі програми»

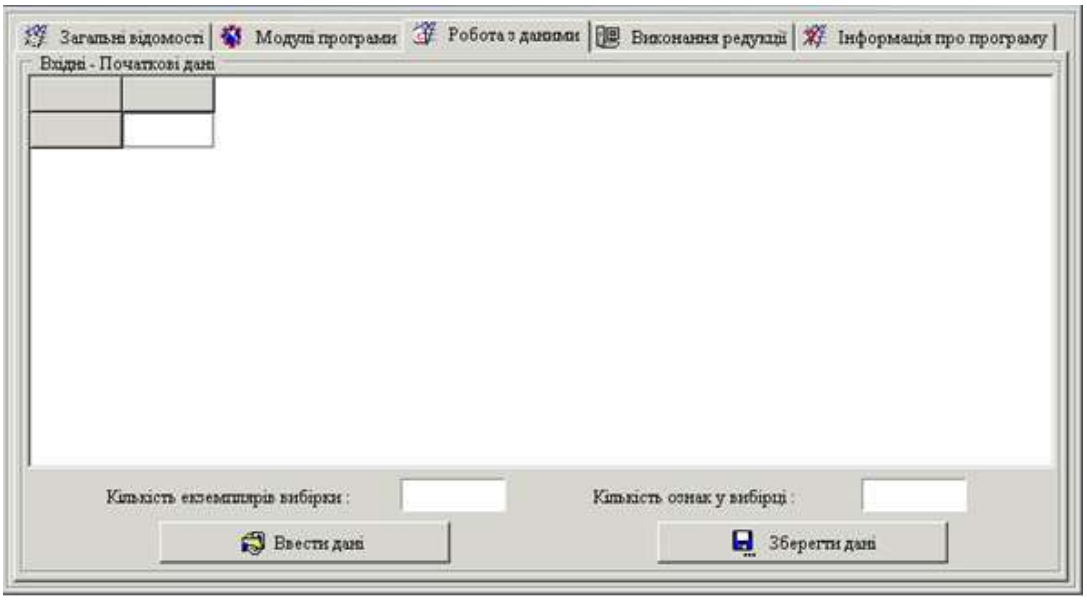

Рисунок 3.4 – Сторінка форми інтерфейсу «Робота з даними»

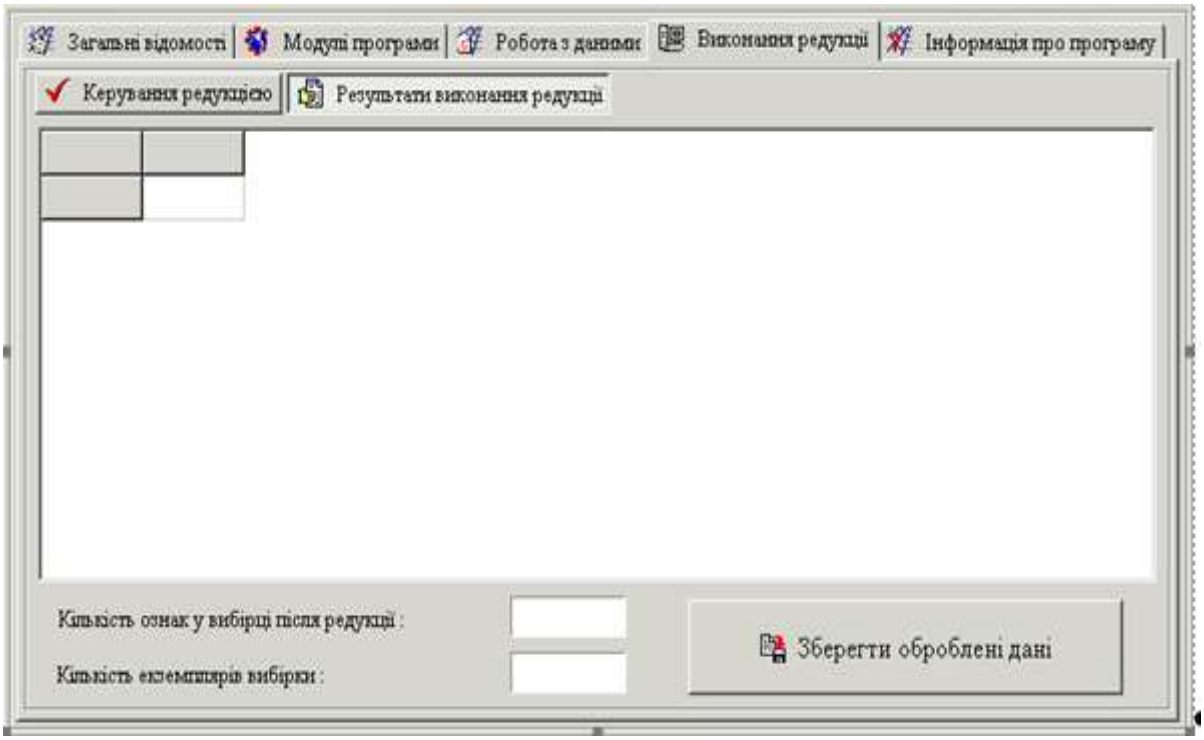

Рисунок 3.5 – Сторінка форми інтерфейсу «Результати виконання редукції»

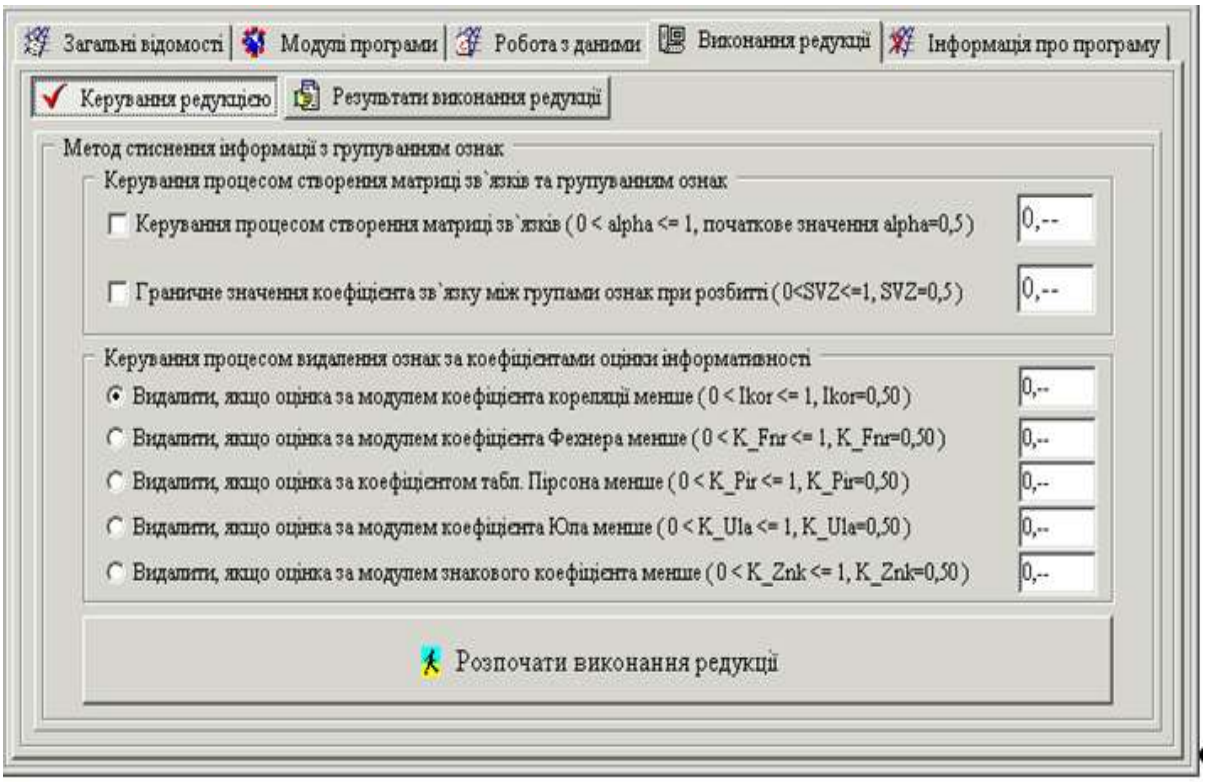

Рисунок 3.6 – Приклад виконання операції «Керування редукцією»

Особливі вимоги програма висуває до обсягу оперативної пам’яті комп’ютера, а в разі її нестачі до можливості застосування віртуальної пам’яті ОС Windows. Під час роботи програми усі дані завантажуються у вигляді динамічних масивів до оперативної пам’яті. Розміщення даних в пам’яті у такому вигляді дозволяє суттєво прискорити розрахунки при математичній обробці даних і ефективно використовувати пам’ять комп’ютера, якщо розмір файлів даних змінюється. Взаємодія модулів програми між собою виконується за допомогою передачі покажчиків на динамічні масиви даних. Для розподілу ресурсів оперативної пам’яті використовується менеджер пам’яті фірми Inprise (Borland) borlndmm.dll, а для віртуальної пам’яті – внутрішній менеджер пам’яті ОС Windows.

Для початку сеансу роботи з програмою необхідно стандартними засобами ОС Windows запустити графічну оболонку підсистеми – файл Reduction.exe. Після завантаження модулів програми до пам’яті комп’ютера на екран буде здійснено виведення графічного інтерфейсу і розпочато цикл очікування команд керування від користувача. Керування програмою виконується за допомогою стандартних елементів програм-додатків ОС Windows.

Розглянемо склад та опис файлів програмного пакету програми. Основними модулями програми є модулі r_mUI.pas (блок інтерфейсу користувача), r_mIO.pas (DLL – модуль введення-виведення даних) та r_mMO.pas (DLL – модуль математичної обробки даних). Для виконання взаємодії з DLL – модулями програми для кожного з них розроблено інтерфейс, що складається з процедур, за допомогою яких виконується виклик модуля та передача даних [19].

Опис інтерфейсів DLL – модулів, а також вибірковий опис головних процедур та функції модуля інтерфейсу користувача наведено в таблиці 3.2.

Вибірковий опис типів даних та і динамічних масивів програми наведено в таблиці 3.3.

Таблиця 3.2 – Вибірковий опис процедур і функцій модулів програми

Формат виклику	Призначення
Модуль інтерфейсу користувача r_mUI (Reduction.exe)	
MUI_StatusInfo(StatCode)	Функція отримання інформації про модуль Вхідні дані: StatCode:byte; Вихідні дані: рядок з інформацією про модуль
TotalSystemInfo	Процедура отримання характеристик ПЕОМ Вихідні дані: інформація про засоби ПЕОМ
P3_LoadIOClick(Sender), P3_SaveIOClick(Sender)	Процедури введення та виведення даних з програми
P41_ReductionStartButtonClick (Sender)	Процедура виконання процесу редукції
P42_SaveResultsClick(Sender)	Процедура збереження результатів процесу редукції
Модуль математичної обробки даних r_mMO.dll	
MMO_Initialization(Math)	Процедура ініціалізації модуля
MMO_StatusInfo(StatCode)	Функція отримання інформації про модуль Вхідні дані: StatCode:byte; Вихідні дані: рядок з інформацією про модуль
MMO_Reduction(In_MMO)	Функція виконання редукції за даними вибірки Вхідні дані:In_MMO:TViborka – початкові дані; Вихідні дані: дані типу TViborka;
MMO_AllFreeMMO	Процедура звільнення динамічних масивів модуля
Модуль введення – виведення даних r_mIO.dll	
MIO_StatusInfo(StatCode)	Функція отримання інформації про модуль Вхідні дані: StatCode:byte; Вихідні дані: рядок з інформацією про модуль
MIO_GetIN(In_MMO, filename)	Функція введення даних модуля Вхідні дані: In_MMO:TViborka – дані, filename – шлях та ім’я файла даних; Вихідні дані: завантажена інформація з файлу даних (файли з розширенням .dat), дані типу TViborka

Таблиця 3.3 – Вибірковий опис даних і динамічних масивів програми

Найменування	Призначення
TViborka	Запис, що описує вибірку
TViborka.PRZ	Динамічний масив значень ознак
TViborka.ZPa	Динамічний масив значень цільових параметрів
TViborka.RedPRZ	Динамічний масив номерів ознак
TViborka. K_pr	Кількість ознак у вибірці
TViborka. K_ex	Кількість екземплярів вибірки
TGraffity	Запис, що описує елементи графу груп ознак
TGraffity.n	Номер ознаки у вибірці (стовпець)
TGraffity.s	Номер ознаки у вибірці (рядок)
TGraffity.Prizn	Значення ознаки x(n,s)
TGraffity.Con1	Зв’язок з попереднім елементом графу
TGraffity.Con2	Зв’язок з наступним елементом графу
In_MMO, Out_MMO	Змінні типу TViborka для введення, виведення даних з/до модулів введення-виведення та математичної обробки
Prizn	Загальний масив значень ознак, до якого вводяться значення ознак з файлу вибірки
ZAPa	Загальний масив значень цільового параметру, до якого вводяться значення параметрів з файлу вибірки
RedPrizn	Загальний масив номерів ознак, що були виключені за методом редукції
R_mas	Матриця відстаней
K_mas	Матриця зв’язків
Graffity	Масив значень типу TGraffity (масив значень графу груп)
Math	Масив керуючих значень за методом редукції
StatCode	Змінна коду статуса модуля (інформація, версія, тощо)
filename	Змінна, що містить шлях та ім’я файлу вибірки
f	Файлова змінна для роботи з текстовими файлами

В програмі існує 2 типа змінних: глобальні та локальні. Глобальні – це масиви та матриці, покажчики на які передаються при виконанні виклику процедур та функцій модулів. Локальні – це переважною більшістю лічильники для циклів математичних розрахунків та введення–виведення даних (для модуля інтерфейсу – змінні формування таблиць даних, для модуля введення-виведення – змінні обліку обсягу введеної або виведеної інформації).

Програма розроблена з урахуванням можливості виникнення ситуацій критичного стану. При появі таких ситуації програма виконує виведення на дисплей (монітор) повідомлень згідно становища, що склалося на момент появи ситуації. Розглянемо вхідні та вихідні дані програми.

Вхідні дані – текстові файли вибірок (.dat), команди користувача. Вихідні дані – текстові файли вибірок (.dat), інформація, що виводяться програмою на дисплей (таблиці, текстові повідомлення, тощо).

Структурну будову файлів вхідних і вихідних даних програми наведено в таблиці 3.4. Дані у файлах (значення ознак, цільового параметру, символи сигнатури, перелік номерів ознак) описуються в ASCII кодуванні. Значення ознак у файлі повинні відокремлюватися один від одного знаком пробілу – ‘ ’, якщо значення ознаки не ціле число, то його запис повинен виконуватися за допомогою знаку коми – ‘,’.

Таблиця 3.4 – Структура файлів даних програми

Сигнатура, номер	Цільовий параметр	Дані	Ознака 1	…	Ознака N
Data file (Coded by WISE`Soft`2000) – сигнатура (позначення) файлу даних					
Число стовпців – число стовпців ознак вибірки (N)					
Число рядків – число рядків екземплярів вибірки (S)					
#1	Значення ЦП 1	Екземпляр 1	значення	...	значення
#...	...	...	...	...	...
#S	Значення ЦП S	Екземпляр S	значення	...	значення
@ – сигнатура кінця даних					
! – у вихідному файлі вибірки за сигнатурою наводиться перелік номерів ознак, які було вилучено з вхідної вибірки					

У разі виникнення критичних ситуацій, а також для повідомлення користувача про стан виконання редукції, програма здійснює виведення на екран повідомлень.

Повідомлення, що виводяться програмою поділяються на 3 типи: повідомлення про похибки, попереджуючі повідомлення, інформаційні

повідомлення. Повідомлення про похибки з'являються у випадку обробки критичної ситуації або стану програми, вони сповіщають користувача про причини відмови або збою виконання програми. Попереджуючі повідомлення з'являються при відстеженні дій користувача програми, у тому випадку коли визначено ситуацію, що може призвести до збою виконання програми. Інформаційні повідомлення з'являються при потребі надати користувачу інформацію про стан або роботу програми. Перелік повідомлень, що з'являються при роботі програми наведено в таблиці 3.5.

Таблиця 3.5 – Перелік повідомлень програми

Текст повідомлення	Дії користувача
Повідомлення про похибки	
1 Похибка завантаження модулів програми – Підсистема редукції даних	
1.1 Не знайдено файл r_mIO.dll, робота програми не може бути продовжена !	Перевірити присутність файлів r_mIO.dll, r_mMO.dll, borlndmm.dll в каталозі, де здійснено запуск програми редукції
1.2 Не знайдено файл r_mMO.dll, робота програми не може бути продовжена !	
1.2 Не знайдено файл borlndmm.dll, робота програми не може бути продовжена !	
2 Похибка завантаження даних програми – Підсистема редукції даних	
2.1 При здійсненні завантаження виявлено похибку в структурі файла даних !	Перевірити структуру файлу даних
Попереджуючі повідомлення	
1 Попередження програми – Підсистема редукції даних	
1.1 Обсяг вибірки дуже великий (більше 12000 ознак на 1000 екземплярів), можлива нестача ресурсів ПЕОМ при виконанні обчислень !	Виконати розбиття файлу даних на декілька частин. Виконати редукцію з кожною частиною окремо
Інформаційні повідомлення	
1. Інформація програми – Підсистема редукції даних	
1.1 Розпочато математичні розрахунки. Натисніть "OK" і очікуйте появи повідомлення про кінець процесу обчислень !	Виконати інструкцію програми
1.2 Розрахунки завершено ! Натисніть "OK" !	
1.3 Файл ReadMe.txt не знайдено !	Перевірити присутність файлу ReadMe.txt в каталозі

3.3 Результати моделювання

Моделювання методів редукції даних здійснювалося з використанням експериментальних даних щодо ризику виникнення у серцево-судинних хворих поєднаної патології (АГ та ІХС – артеріальної гіпертензії та ішемічної хвороби серця), наведених в датасеті [18].

Одним з важливих завдань діагностики захворювання цього типу є попередня оцінка коефіцієнта ризику β^{y}, що відображує ймовірність наявності у серцево-судинних хворих поєднаної патології зазначеного типу.

Ефективність своєчасного виникнення подібної патології, значною мірою залежить від наявності відповідної моделі медичної діагностики, що дозволяла б оперативно враховувати значення найбільш інформативних діагностичних ознак (факторів).

Як фактори, що найбільш повно відображають процес виникнення хворих поєднаної патології АГ та ІХС, згідно з [18] використовуються такі ознаки: 1 – середньодобовий артеріальний тиск; 2 – діастолічний артеріальний тиск; 3 – пульсовий артеріальний тиск; 4 – систолічний артеріальний тиск; 5 – частота сердцевих скорочень; 6 – індекс площі гіпертензії; 7 – індекс рігідності артерій; 8 – внутріміокардіальна напруга; 9 – маса міокарда лівого шлуночка; 10 – обсяг порожнини лівого шлуночка; 11 – ступень нічного зниження середньодобового артеріального тиску; 12 – ступень денного зниження середньодобового артеріального тиску; 13 – периферічний опір судин; 14 – індекс маси тіла; 15 – середньодобовий пульсовий артеріальний тиск; 16 – швидкість циркулярного скорочення волокон міокарду.

Завданням цього дослідження була оцінка значущості (інформативності) факторів, що використовуються при побудові моделі діагностики поєднаної патології АГ та ІХС.

Для побудови такої моделі доцільним є використання БШП та модифікованого методу стиснення даних, розглянутих в розділі 2. Після

побудови нейромережевої моделі необхідно здійснити аналіз інформативності факторів, на основі яких здійснюється оцінка значення коефіцієнта ризику β^y з метою спрощення моделі та підвищення її достовірності. Для цього пропонується використовувати комбінацію наведеного вище модифікованого методу стиснення даних з групуванням ознак та нейромережевої оцінки інформативності ознак.

Навчання нейронної мережі здійснювалося для 16 розглянутих вище ознак за навчальними даними, наведеними в [18].

Моделювання здійснювалося за допомогою двошарового персептрону, перший шар якого містив 4 нейрони, а другий шар – 1 нейрон. Усі нейрони мали сигмоїдну функцію активації ψ(x)=1/(1+e-x).

Матриця вагових коефіцієнтів, отримана в результаті навчання БШП, представлена в таблиці 3.6.

Таблиця 3.6 – Матриця вагових коефіцієнтів

j – номер входу нейрону	i – номер нейрона в шарі				μ – номер шару
	1	2	3	4	
0	14,1894	-5,1831	-10,8277	-5,3176	1
1	-21,1447	4,6313	-0,7001	0,9456	
2	-7,8715	-3,2767	0,4041	3,4354	
3	-11,5879	35,883	6,7379	5,2636	
4	-2,8764	15,19	-1,1862	2,0939	
5	34,8108	-4,2685	0,3106	-2,1849	
6	-14,0461	11,9728	-0,4762	2,2898	
7	-1,3948	28,2757	-36,5509	19,7894	
8	-4,4415	-0,8422	-1,7184	3,4117	
9	-5,6667	-8,1141	13,0482	1,4346	
10	6,1819	-7,5646	-2,5819	-1,7134	
11	-1,5746	-10,854	-3,4903	2,2829	
12	10,664	4,4055	5,3613	-8,4166	
13	-0,7028	0,5237	-1,0671	1,2218	
14	-21,3477	15,2453	-8,8919	5,6718	
15	-0,9858	7,8709	15,4614	-2,6364	
16	0,4963	-7,2036	30,9704	0,4219	
0	-34,3184				2
1	37,3173				
2	-32,6482				
3	29,7206				
4	63,3830				

На входи БШП подавалися значення факторів. На вихід мережі подавалося значення коефіцієнта ризику для відповідного зразка.

Як цільову функцію при навчанні використовувався мінімум середньоквадратичної помилки мережі для всієї вибірки goal=10-6.

Навчання БШП проводилося з використанням алгоритму Левенберга-Марквардта. При навчанні БШП крок навчання дорівнював 0.00001, а максимальна кількість циклів навчання epochs = 500.

Кількість витрачених циклів навчання БШП склала 339, середньоквадратична помилка – 9.94262.

Результати нейромережевого моделювання коефіцієнта ризику наведено у таблиці 3.7. Тут $\beta^{у}_{експ}$ – значення коефіцієнта ризику, отримане експериментально, $\beta^{у}_{розр}$ – розрахункове значення коефіцієнта ризику, отримане за допомогою БШП.

Таблиця 3.7 – Результати нейромережевого моделювання

номер екземп-ляру	$\beta^{у}_{експ.}$	$\beta^{у}_{розр.}$	номер екземп-ляру	$\beta^{у}_{експ.}$	$\beta^{у}_{розр.}$	номер екземп-ляру	$\beta^{у}_{експ}$	$\beta^{у}_{розр}$
1	1,16	1,16	21	1,45	1,45	41	1,58	1,58
2	1,27	1,27	22	1,5	1,5	42	1,6	1,6
3	1,38	1,38	23	1,61	1,61	43	1,56	1,56
4	1,54	1,54	24	1,64	1,64	44	1,66	1,66
5	1,46	1,46	25	1,44	1,44	45	1,6	1,6
6	1,35	1,35	26	1,13	1,1304	46	1,59	1,59
7	1,39	1,39	27	1,29	1,29	47	1,55	1,55
8	1,35	1,35	28	1,29	1,29	48	1,56	1,56
9	1,21	1,21	29	1,32	1,32	49	1,53	1,53
10	1,51	1,51	30	1,42	1,42	50	1,2	1,2
11	1,51	1,51	31	1,57	1,57	51	1,19	1,19
12	1,38	1,38	32	1,47	1,47	52	1,15	1,1499
13	1,37	1,37	33	1,43	1,43	53	1,13	1,1306
14	1,21	1,21	34	1,4	1,4	54	1,83	1,83
15	1,38	1,38	35	1,48	1,48	55	1,75	1,75
16	1,38	1,38	36	1,52	1,52	56	1,23	1,23
17	1,38	1,38	37	1,48	1,48	57	1,22	1,22
18	1,37	1,37	38	1,49	1,49	58	1,39	1,39
19	1,33	1,33	39	1,45	1,45	59	1,35	1,35
20	1,27	1,27	40	1,6	1,6			

Після отримання нейромережевої моделі здійснювалася нейромережева оцінка інформативності факторів, а також було знайдено коефіцієнти кореляції факторів та коефіцієнта ризику. Результати оцінки інформативності факторів наведено у таблиці 3.8.

Таблиця 3.8 – Результати оцінки інформативності ознак

Номер ознаки	Коефіцієнти кореляції ознак та прогнозованого параметра	Нейромережева оцінка інформативності ознак	Рівень інформативности ознаки
1	-0.088	0.0456	малоінформативна
2	0.288	0.0380	інформативна
3	0.3076	0.1033	інформативна
4	-0.0769	0.0374	малоінформативна
5	-0.199	0.0736	інформативна
6	-0.2947	0.0512	інформативна
7	-0.042	0.2096	інформативна
8	0.1979	0.0314	інформативна
9	0.1665	0.0459	малоінформативна
10	0.2139	0.0330	інформативна
11	-0.4946	0.0345	інформативна
12	0.2442	0.0814	інформативна
13	0.0428	0.0108	малоінформативна
14	0.1308	0.0993	інформативна
15	-0.0919	0.0491	малоінформативна
16	0.2875	0.0558	інформативна

На основі отриманих оцінок інформативностей факторів приймалися рішення про поділ факторів на дві групи: інформативні та малоінформативні. До інформативних належали ті фактори, інформативності яких перевищували певні порогові значення, а до малоінформативних належали ті фактори, інформативності яких були меншими від порогових значень. При прийнятті рішення про інформативність ознак приймався такі поріг значущості для нейромережевих оцінок інформативностей ознак, що дорівнював 0.0625.

Після прийняття рішень про інформативність ознак з навчальної

множини були виключені малоінформативні ознаки (1, 4, 9, 13, 15). Потім здійснювалося повторне моделювання коефіцієнта зміцнення на основі БШП. При цьому всі параметри БШП та процесу навчання були такими ж, як і в попередньому випадку.

Результати нейромережевого моделювання після виключення малоінформативних ознак свідчать про те, що похибка розрахунку коефіцієнта зміцнення порівняно з попереднім випадком майже не змінилася, проте швидкість навчання при фіксованій кількості циклів навчання та швидкість роботи БШП підвищилися порівняно з попереднім випадком, де враховувалися всі ознаки.

Також в роботі було здійснено моделювання роботи методу головних компонентів (PCA) та методу RF для визначення інформативності окремих ознак, що мають вплив на загальний стан серцево-судинної системи пацієнта [20], [21], [22].

Вихідні дані для моделювання були взяті з датасету Heart Failure Prediction https://www.kaggle.com/fedesoriano/heart-failure-prediction [23].

Як фактори, що відображають загальний стан серцево-судинної системи пацієнта (HeartDisease) використовувались такі ознаки:

– вік пацієнта (age);

– стать пацієнта (sex);

– тип болю в грудях (ChestPainType): типова стенокардія (TA), атипова стенокардія (ATA), неангінальний біль (NAP), безсимптомний (ASY);

– артеріальний тиск у стані спокою (RestingBP);

– холестерин: сироватковий холестерин (Cholesterol: serum cholesterol);

– рівень цукру в крові натще (FastingBS);

– ЕКГ у стані спокою: результати електрокардіограми в спокої (RestingECG: resting electrocardiogram results): нормальний, наявність аномалій зубця ST-T (інверсії зубця T та/або елевація або депресія ST > 0,05 мВ), вірогідна або визначена гіпертрофія лівого шлуночка;

– максимальна частота серцевих скорочень (MaxHR);

– стенокардія, спричинена фізичним навантаженням (ExerciseAngina);

– числова оцінка рівню депресії (Oldpeak: oldpeak = ST);

– нахил сегмента ST піку (ST_Slope): вгору, рівно, вниз.

Використання методу головних компонентів (PCA) дозволило отримати результати ранжирування інформативності окремих ознак, що мають вплив на загальний стан серцево-судинної системи пацієнта, наведені в таблиця 3.9.

Таблиця 3.9 – Результати ранжирування інформативності ознак за методом PCA

	0	1
0	PC0	Cholesterol
1	PC1	MaxHR
2	PC2	RestingBP
3	PC3	Age
4	PC4	Oldpeak

Згідно з отриманими за методом PCA результатами найважливішими ознаками для оцінювання загального стану серцево-судинної системи пацієнта є: Cholesterol (рівень холестерину), MaxHR (максимальна частота серцевих скорочень), RestingBP (результати електрокардіограми в спокої), Age (рік), Oldpeak (числова оцінка рівню депресії).

Досліджувані ознаки по завершенню роботи методу RF («випадковий ліс») отримали нормовані значення параметру значущості атрибутів (від 0 до 1), що наведені на рисунку 3.7

Згідно з отриманими за методом RF результатами найважливішими ознаками для оцінювання загального стану серцево-судинної системи пацієнта є: Cholesterol (рівень холестерину), MaxHR (максимальна частота серцевих скорочень), ChestPainType (тип болю в грудях): RestingBP (результати електрокардіограми в спокої), Oldpeak (числова оцінка рівню депресії), Age (рік).

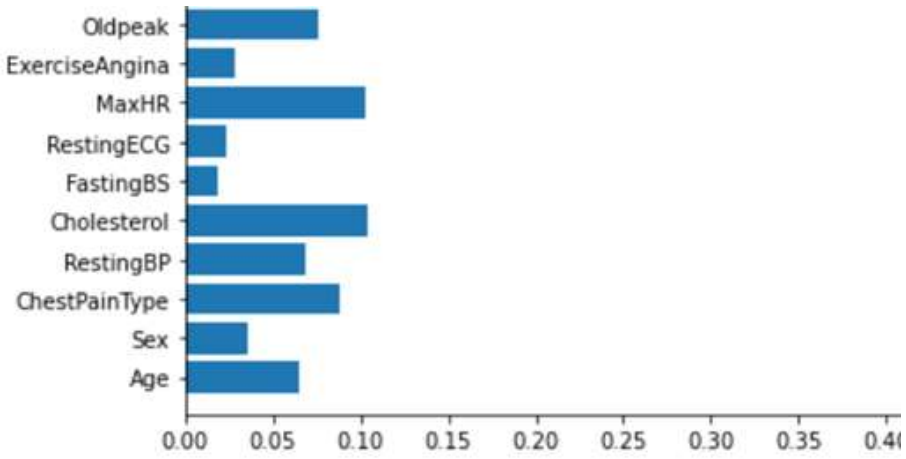

Рисунок 3.7 – Результати ранжирування інформативності ознак за методом Random Forest

Перспективами розвитку проведених досліджень (щодо діагностики стану серцево-судинної системи пацієнта) може бути нейромережеве моделювання впливу розглянутих факторів на стан серцево-судинної системи пацієнта.

Таким чином, результати моделювання процедур визначення інформативних параметрів та редукції даних для діагностики стану серцево-судинної системи пацієнта за методом стиснення даних з групуванням ознак та методом PCA підтверджують працездатність запропонованих в роботі діагностичних засобів.

ВИСНОВКИ

Результатом виконання кваліфікаційної роботи магістра є розробка та дослідження алгоритмічних та програмних засобів, які реалізують методи редукції даних для інтелектуальних систем діагностики, що дозволяє зменшувати кількість діагностичних ознак та розмір вибірок.

Аналіз існуючих джерел за темою дослідження дозволив виявити недоліки та переваги існуючих методів редукції даних в системах інтелектуальної діагностики.

Відзначено перспективність використання в цих системах методу стиснення інформації з групуванням ознак в поєднанні з нейромережевим підходом.

Запропоновано модифікацію цього методу, сутність якої полягає в застосуванні спрощеного способу побудови графу звязків між ознаками та використанні багатошарового персептрону для оцінювання інформативності ознак, сформованих за модифікованим методом.

Здійснено програмну реалізацію модифікованого методу редукції та запропоновану структуру модулю редукції даних, що може бути використаний як в складі діагностичної системи, так і автономно.

Наведено результати нейромережевого моделювання процедур визначення інформативних параметрів та редукції даних для діагностики стану серцево-судинної системи пацієнта, що підтверджують працездатність запропонованих в роботі діагностичних засобів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Олійник А.О., Субботін С.О., Олійник О.О. Інтелектуальний аналіз даних : навчальний посібник. Запоріжжя: ЗНТУ, 2012. 278 с.

2. Троцько В.В. Методи штучного інтелекту: навчально-методичний і практичний посібник. Київ: Університет економіки та права «КРОК», 2020. 86 с.

3. Бодянський Є.В., Пелешко Д.Д., Винокурова О.А., Машталір С.В., Іванов Ю.С. Аналіз та обробка потоків даних засобами обчислюваного інтелекту: монографія. Львів: Видавництво Львівської політехніки, 2016. 236 с.

4. Роїк М.В., Присяжнюк О.І., Денисюк В.О. Огляд програмних засобів статистичного аналізу даних. *Ефективна економіка*. 2017. 7. URL: http://www. nayka.com.ua/op=1&z=56 (дата звернення 10.10.2021)

5. Черняк О.І., Захарченко П.В. Інтелектуальний аналіз даних. К.: Знання, 2014. 599 с.

6. Дубровин В.И., Субботин С.А., Богуслаев А.В., Яценко В.К. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: Монография. Запорожье: ОАО "Мотор-Сич", 2003. 279 с.

7. Ганчук А.А., Соловйов В.М., Чабаненко Д.М. Методи прогнозування. Черкаси: Брама-Україна, 2012. 140 с.

8. Бахрушин В.Є. Методи аналізу даних: навч. посібник для студентів. Запоріжжя: КПУ, 2011. 268 с.

9. Троцько В.В. Методи штучного інтелекту: навчально-методичний і практичний посібник. Київ: Університет економіки та права «КРОК», 2020. 86 с.

10. Ian H. Witten, Frank Eibe, Mark A. Hall. Data mining: practical machine learning tools and techniques. 3rd ed. Morgan Kaufmann Publishers, 2011. 630 p.

11. Субботин С.А. Формирование и редукция выборок для интеллектуального анализа данных. *Радіоелектроніка, інформатика, управління*. 2013. № 1. С. 113-118.

12. Компонентний аналіз URL: https://stud.com.ua/93349/statistika/komponentniy_analiz (дата звернення: 17.10.2021).

13. Blazquez D., Domenech J. Big Data sources and methods for social and economic analyses / Technological Forecasting and Social Change, 2018. Vol. 130. P. 99-113.

14. Технический справочник по алгоритму временных рядов (Майкрософт) URL: http://msdn.microsoft.com/ru-ru/library/bb677216.aspx (дата звернення: 10.10.2021).

15. Polamuri S. How the Random Forest algorithm works in machine learning. URL: https://data/aspirant.com/2017/05/22/random-forest-algorithm-machine-learing (application date: 13.10.2021).

16. Карякина А.А., Мельников А.В. Сравнение моделей прогнозирования оттока клиентов интернет-провайдеров. *Машинное обучение и анализ данных*, 2017. Том 3, № 4. С. 250-256.

17. Поморова О.В., Гнатчук Є.Г., Севастьянов М.Г. Дослідження ефективності методів опрацювання нечіткої діагностичної інформації в інтелектуальних системах технічного діагностування / Вісник Хмельницького національного університету. №3. 2010. С. 181-185.

18. Діагностування та прогнозування. URL: http://www.tsatu.edu.ua/tstt/wp-content/uploads/sites/6/-diahnostuvannja-ta-prohnozuvannja-.pdf. (дата звернення: 15.10.2021).

19. Шумейко А. А., Сотник С. Л. Интеллектуальный анализ данных (Введение в Data Mining.) Днепропетровск: Белая Е.А., 2012. 212 с.

20. Ronaghan S. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. URL: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-

andfeature-importance-in-scikit-learn-and-spark-f2861df67e3 (application date: 13.10.2021).

21. Trost J. Getting Started with DGA Domain Detection Research. URL: http://www.covert.io/gettingstarted-with-dga-research (application date: 13.10.2021).

22. Gorban A.N, Kegl B., Wunsch D., Zinovyev A.Y. Principal Manifolds for Data Visualisation and Dimension Reduction, Series: Lecture Notes in Computational Science and Engineering, 58, Springer, Berlin. Heidelberg. New York, 2007, XXIV , 340 p.

23. Dataset: Heart Failure Prediction URL: https://www.kaggle.com/fedesoriano/heart-failure-prediction (application date: 16.10.2021).