МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
імені В. Н. КАРАЗІНА

Ю. В. Холін
Я. М. Пушкарьова
А. В. Пантелеймонов
А. Н. Некос

ХЕМОМЕТРИЧНІ МЕТОДИ В РОЗВ'ЯЗАННІ ЗАДАЧ ЯКІСНОГО ХІМІЧНОГО АНАЛІЗУ ТА КЛАСИФІКАЦІЇ ФІЗИКО-ХІМІЧНИХ ДАНИХ

Монографія

Харків – 2016

УДК 543.061+543.08
ББК 24.4
Х 71

Рецензенти:

В. В. Іванов – доктор хімічних наук, професор, професор кафедри хімічного матеріалознавства, Харківський національний університет імені В. Н. Каразіна;

С. А. Неділько – доктор хімічних наук, професор, професор кафедри неорганічної хімії, Київський національний університет імені Тараса Шевченка.

Затверджено до друку рішенням Вченої ради
Харківського національного університету імені В. Н. Каразіна
(протокол № 10 від 30 серпня 2016 р.)

Холін Ю. В.

Х 71 Хемометричні методи в розв'язанні задач якісного хімічного аналізу та класифікації фізико-хімічних даних : монографія / Ю. В. Холін, Я. М. Пушкарьова, А. В. Пантелеймонов, А. Н. Некос. – Х. : ХНУ імені В. Н. Каразіна, 2016. – 184 с.

ISBN 978-966-285-411-4

В монографії обговорено зміст та актуальні завдання сучасного якісного хімічного аналізу та хемометричні методи, що найчастіше використовують для обробки хіміко-аналітичних і фізико-хімічних даних. Особливу увагу приділено засобам контролю автентичності продуктів харчування і напоїв, сільськогосподарської сировини, лікарських засобів, ідентифікації об'єктів довкілля. Розглянуто застосування апарату штучних нейронних мереж та нечітких множин для розв'язання задач якісного хімічного аналізу (ідентифікації аналітів та кластеризації багатопараметричних масивів даних).

Для фахівців у царинах хемометрії, якісного хімічного аналізу, фізичної хімії.

Іл. 36, табл. 68, бібліогр. 333 назв.

УДК 543.061+543.08
ББК 24.4

ISBN 978-966-285-411-4

Зміст

Вступ

Неперервний техногенний тиск на умови мешкання людини, зростаючий запит на контроль автентичності продуктів споживання надають особливої значущості якісному хімічному аналізу. Уявлення про зміст якісного хімічного аналізу протягом останніх десятиріч зазнало суттєвих змін. Сьогодні його трактують як процедуру *класифікації* об'єктів за їх ознаками. Якісний хімічний аналіз розв'язує задачі *виявлення* (встановлення присутності певного аналіту в пробі), *ідентифікації* (*дискримінації*) (ототожнення аналіту з відомою індивідуальною речовиною або групою речовин, віднесення зразка до одного із заздалегідь установлених класів) і *кластеризації* (визначення сукупностей зразків з близькими характеристиками при відсутності навчальних вибірок). Результати аналізу розглядають як рекомендації для прийняття управлінських рішень.

Ідентифікація аналітів та кластеризація об'єктів ґрунтуються на обробці багатовимірних масивів експериментальних даних, отриманих інструментальними методами (хроматографічними, спектроскопічними, системами «електронний язик», «електронний ніс» тощо). Для того щоб забезпечити високу надійність класифікації, згадані масиви даних слід обробляти з використанням ефективних хемометричних методів.

Протягом останніх років на хімічному та екологічному факультетах Харківського національного університету імені В. Н. Каразіна виконуються дослідження, спрямовані на розробку ефективних хемометричних засобів аналізу багатовимірних масивів хімічних даних, в першу чергу для розв'язання задач ідентифікації аналітів та кластеризації об'єктів в якісному хімічному аналізі. В ході робіт на задачі ідентифікації та кластеризації було поширено ключове метрологічне поняття «надійність результатів якісного аналізу», розроблено низку алгоритмів класифікації, що ґрунтуються на апаратах теорій штучних нейронних мереж та нечітких множин, запропоновано засоби обробки багатовимірних масивів експериментальних даних, стійкі до наявності в даних пропусків та викидів. Було показано, що створені алгоритми класифікації не лише ефективно розв'язують досить традиційні задачі якісного аналізу (перевірка автентичності

продуктів споживання, ідентифікація зразків природних вод за хімічним складом тощо), але й дозволяють із прийнятною надійністю ідентифікувати географічне походження рослинної сировини за відомостями про вміст у ній важких і перехідних металів.

У монографії викладено загальні підходи до розв'язання актуальних проблем якісного аналізу та конкретні алгоритми ідентифікації та кластеризації, а також наведено результати прикладання цих алгоритмів до конкретних хімічних задач. Особлива увага приділяється класифікації з використанням штучних нейронних мереж.

Виконувані ними функції можна розділити на декілька основних груп: апроксимації й інтерполяції; розпізнавання та класифікації образів; стиснення даних; прогнозування; ідентифікації; управління. Штучні нейронні мережі мають такі цінні властивості, як здатність до навчання (мережу можна навчити розв'язанню необхідної задачі, виконавши алгоритм навчання), здатність до узагальнення (після навчання мережа може працювати із зашумленими або спотвореними даними, даючи правильний результат на виході), свобода від апріорних припущень щодо статистичних характеристик вихідних даних.

Однією з найскладніших задач класифікації є кластеризація (розбиття на декілька однорідних груп) масиву багатопараметричних даних за відсутності еталонів (навчальних вибірок) та інформації про кількість однорідних класів. У монографії описано запропонований авторами алгоритм обробки таких даних на основі поєднання нейронної мережі Кохонена, ймовірнісної мережі та процедури перехресної оцінки достовірності (крос-валідації). Цю процедуру випробувано при класифікації органічних розчинників за набором їхніх фізико-хімічних параметрів та класифікації річкових і джерельних вод і рослинної сировини за вмістом у них важких і перехідних металів. Одержані класифікації виявилися змістовними, а розроблений алгоритм можна розглядати як ефективний засіб експлораторного аналізу фізико-хімічних та хіміко-аналітичних даних.

З великою приємністю висловлюємо вдячність проф. В. І. Вершиніну (Омський державний університет ім. Ф. М. Достоєвського), проф. В. В. Іванову (ХНУ імені В. Н. Каразіна), проф. М. О. Мчедлову-Петросяну (ХНУ імені В. Н. Каразіна) за цінні поради, зауваження та допомогу в роботі. Автори вдячні М. О. Оніжуку за допомогу в підготовці рукопису, проф. О. І. Юрченку, н.с. Н. П. Титовій та А. Г. Гарбуз (ХНУ імені В. Н. Каразіна) за надання експериментальних даних про вміст металів у зразках природних вод і рослинній сировині.

Глава 1

АКТУАЛЬНІ ЗАДАЧІ СУЧАСНОГО ЯКІСНОГО ХІМІЧНОГО АНАЛІЗУ ТА ПІДХОДИ ДО ЇХ РОЗВ'ЯЗАННЯ

1.1. Основні завдання та зміст сучасного якісного хімічного аналізу

Вважають, що уявлення про якісний хімічний аналіз виникло у XVIII столітті завдяки працям шведського хіміка Торнберна Улафа Бергмана (1735–1784), який відділив якісний аналіз від кількісного і розробив першу схему якісного аналізу [1].

У XX сторіччі усталеним став погляд на якісний аналіз як на розділ аналітичної хімії, присвячений встановленню складу речовин і матеріалів (вмісту атомів, іонів, молекул, ізотопів елементів, функціональних груп) за допомогою хімічних і фізичних методів. Із розвитком інструментальних методів, комп'ютеризації аналізу і створенням великих за розмірами баз даних із фізико-хімічних властивостей речовин, ростом практичних запитів на аналіз проб складного складу (побутової продукції, об'єктів навколишнього середовища, лікарських засобів, харчів) зміст якісного хімічного аналізу змінився, а потреба у ньому суттєво зросла.

Проблематиці сучасного якісного хімічного аналізу присвятив спеціальний номер авторитетний журнал «Trends in Analytical Chemistry» [2]. Було розглянуто питання термінології, контролю якості аналізу, підходів до оцінки невизначеності в якісному аналізі, проблеми

ідентифікації хімічних сполук. Редактори випуску М. Валкарсел та С. Карденас зазначали [3], що свого часу якісний аналіз був пов'язаний із виявленням катіонів та аніонів у розчинах (подібне твердження зустрічається і в монографії В. І. Вершиніна, Б. Г. Дерендяєва та К. С. Лебедєва [4]), але ситуація кардинально змінилася, коли стало зрозумілим, що важливі практичні результати можна отримати не лише на основі сенсорного сприйняття ознак хімічних реакцій, але й на основі обробки масивів експериментальних даних, одержаних інструментальними методами (хроматографія, низка спектроскопічних методів, сенсорні системи «електронний язик» та «електронний ніс») [3, 5].

Отже, протягом останніх двох десятиріч зміст якісного аналізу принципово змінився. Суттєво зросла і його роль. Це обумовлено зростаючою потребою в масовому аналізі складних сумішей у таких областях, як аналіз об'єктів довкілля, перевірка автентичності медико-біологічних препаратів, продуктів харчування, харчової сировини, виявлення токсикантів, наркотиків, вибухонебезпечних речовин.

Сучасний якісний аналіз розв'язує задачі *виявлення* аналітів; *ідентифікації* зразків за набором їх характеристик; *кластеризації* даних (визначення сукупностей об'єктів із близькими характеристиками за відсутності навчальних вибірок). Характерними є такі висловлювання щодо змісту якісного аналізу:

- «Качественный анализ – анализ, в котором вещества идентифицируют или классифицируют на основе их химических или физических свойств, таких как химическая реакционная способность, растворимость, молекулярный вес, температура плавления, испускание или поглощение излучения, масс-спектры, полупериоды полураспада ядер и т.д.» [6];
- «The objectives of qualitative analysis of multispecies solutions consist of discrimination, classification, or identification of different samples» [7];
- «A recent direction in computerized qualitative analysis is solving classification problems related to the assignment of a studied material as an integral object to one or another class» [8].

Результатом розв'язання всіх цих завдань є класифікація об'єктів аналізу: при виявленні − розподіл зразків на групи, що містять аналіт у концентрації, яка перевищує порогову, і не містять його; при ідентифікації − висновок про тотожність досліджуваного зразка й еталона або про приналежність зразка деякому класу об'єктів на

основі відповідності їхніх властивостей (класифікація з навчанням) [6]; при кластеризації − розподіл масиву аналізованих зразків на групи об’єктів із близькими характеристиками.

У зв’язку з цим все більш виразною стає тенденція розглядати якісний аналіз як процедуру класифікації об’єктів за їх ознаками («Qualitative analysis: the classification of objects against specified criteria to meet an agreed requirements» [9]). Результати аналізу розглядають як рекомендації для ухвалення управлінських рішень.

Раніше в якісному аналізі використовували методики з бінарним відгуком, що безпосередньо приводять до висновку про виявлення або ідентифікацію аналіту. В методиках з бінарним відгуком аналітичний сигнал найчастіше реєструють органолептично (в більшості випадків – візуально). Застосовуються і методики, що використовують для вимірювання сигналу інструментальні методи. Тоді висновок формують на основі порівняння значення аналітичного сигналу з деяким пороговим значенням. Як указував Б. Л. Мільман [10], «следует проводить различие между идентификацией аналита и его обнаружением, поскольку последнее представляет собой регистрацию аналитического сигнала без решающего заключения о его природе».

Сьогодні в якісному аналізі домінують підходи, в яких аналітичний висновок формується на основі обробки багатовимірних масивів первинних кількісних даних, отриманих інструментальними методами (хроматографічними, спектроскопічними, системами «електронний язик», «електронний ніс» тощо). В. І. Вершинін запропонував називати ці масиви даних «спектрами» [8].

Алгоритми класифікації діляться на дві групи: алгоритми «без навчання» і «з навчанням».

Алгоритми «без навчання» застосовують для знаходження однорідних груп об’єктів (кластеризації), алгоритми «з навчанням» – для визначення приналежності об’єктів до заздалегідь установлених класів (ідентифікації) [11].

Процедури класифікації повинні задовільно працювати при класифікації сполук, подібних за будовою і властивостями (при перекриванні класів), а також у випадку погано описаних властивостей, коли масиви даних містять пропуски, порушено гіпотезу про нормальний розподіл експериментальних похибок або інформація про розподіл похибок взагалі відсутня.

Для забезпечення високої надійності класифікації аналітів обробляти такі масиви слід із використанням ефективних методів аналізу даних, зокрема хемометричних.

1.2. Метрологічні проблеми якісного аналізу

Принципи метрології якісного хімічного аналізу істотно відрізняються від метрологічних засад кількісного аналізу [12]. Дійсно, для ключового поняття метрології кількісного аналізу – невизначеності (uncertainty) – в якісному аналізі аналогів запропонувати неможливо.

Метрологічні проблеми якісного аналізу розглядали, головним чином, для задач із бінарним відгуком (наприклад, для відповіді на питання «аналіт виявлений / аналіт невиявлений», «ТАК / НІ» [13]). Деякі метрологічні характеристики процедур, що приводять до бінарних відгуків, наведено в табл. 1.1 [5, 6, 14, 15].

Таблиця 1.1

Метрологічні характеристики процедур із бінарним відгуком

Характеристика	Англійський термін, позначення	Визначення
Позитивний результат	Positive, P	Наявність аналітичного сигналу
Негативний результат	Negative, N	Відсутність аналітичного сигналу
Правильний позитивний результат	True positive, TP	Аналіт, присутній у пробі, виявлено; аналіт ідентифіковано правильно
Хибний позитивний результат	False positive, FP	Відсутній аналіт виявлено; аналіт ідентифіковано неправильно
Правильний негативний результат	True negative, TN	Відсутній аналіт не виявлено (не ідентифіковано)
Хибний негативний результат	False negative, FN	Присутній аналіт не виявлено (не ідентифіковано)
Кількість правильних позитивних результатів	n_{TP}	

Продовження табл. 1.1

Кількість хибних позитивних результатів	n_{FP}	
Кількість правильних негативних результатів	n_{TN}	
Кількість хибних негативних результатів	n_{FN}	
Частота хибних позитивних результатів	False positive rate, FPR	$\mathrm{FPR} = \frac{100 \cdot n_{\mathrm{FP}}}{n_{\mathrm{FP}} + n_{\mathrm{TN}}}, \%$
Специфічність (показник правильних негативних результатів)	Specificity (true negative rate), Sp	$\mathrm{Sp} = \frac{100 \cdot n_{\mathrm{TN}}}{n_{\mathrm{TN}} + n_{\mathrm{FP}}}, \%$
Частота хибних негативних результатів	False negative rate, FNR	$\mathrm{FNR} = \frac{100 \cdot n_{\mathrm{FN}}}{n_{\mathrm{FN}} + n_{\mathrm{TP}}}, \%$
Прогностичність позитивного результату	Positive predictive value, PPV	$\mathrm{PPV} = \frac{100 \cdot n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}}, \%$
Прогностичність хибного результату	Negative predictive value, NPV	$\mathrm{NPV} = \frac{100 \cdot n_{\mathrm{TN}}}{n_{\mathrm{TN}} + n_{\mathrm{FN}}}, \%$

Безперечним є той факт, що будь-який висновок про виявлення або ідентифікацію аналіту повинен супроводжуватися кількісною оцінкою його ненадійності (ймовірності хибного висновку, unreliability) [3]. Відповідно до [16–18], ненадійність результатів якісного аналізу можна охарактеризувати як кількісну оцінку ймовірності хибного логічного висновку (помилкової класифікації). Прикладами таких висновків можуть бути помилкова ідентифікація естеру як аміду, висновок про відсутність аналіту, який насправді наявний у пробі, неправомірне віднесення зразка мінеральної води до групи вод з високою мінералізацією тощо.

Для задач виявлення та ідентифікації

$$\text{Unreliability (\%)} = \text{\% false positives} + \text{\% false negatives}, \quad (1.1)$$

$$\text{Reliability (\%)} = 100\,\% - \text{\% false positives} - \text{\% false negatives}. \quad (1.2)$$

Але для задач, в яких результатів класифікації може бути не два, а більше, формули (1.1), (1.2) неприйнятні. Подібне твердження стосується і більшості метрологічних характеристик, представлених в табл. 1.1.

Оцінка ненадійності може бути апріорною [4, 19, 20] або статистичною [21, 22].

Панує статистичний підхід, що полягає у визначенні частки помилково класифікованих зразків у тестовій вибірці [23, 24]. Разом з тим, статистична («ad hoc») процедура, приваблюючи простотою і даючи розумну оцінку ненадійності розв'язання даної класифікаційної задачі, має і суттєвий недолік: важко передбачити поведінку алгоритму класифікації при переході до обробки нового масиву експериментальних даних, можливо, з іншою структурою і статистичними характеристиками.

1.3. Хемометричні засоби контролю автентичності продуктів харчування і напоїв, сільськогосподарської сировини, лікарських засобів, ідентифікації об'єктів довкілля

1.3.1. Проблема контролю автентичності

Ренесанс якісного аналізу обумовлений швидким зростанням практичних запитів на масовий аналіз проб складного складу в нових предметних областях (клінічний аналіз, промислова гігієна, контроль психотропних і наркотичних препаратів, виявлення допінгу, оцінка токсичності об'єктів довкілля, сигнальний контроль забруднювачів, токсикантів, диверсійних отрут тощо). Одним із найважливіших завдань є ідентифікація та забезпечення безпеки харчових продуктів, напоїв і харчової сировини.

Під ідентифікацією харчових продуктів розуміють встановлення відповідності продукту його заявленому найменуванню (виду, класу, категорії, сорту, географічному походженню тощо [25–36]). Продукти, що не відповідають заявленому найменуванню, відносять до розряду фальсифікованих. «Фальсифікацією» (порушенням справжності) вважають підробку чистого або справжнього товару чи заміну дорогого компонента компонентом із нижчою вартістю для отримання незаконної додаткової вигоди, що веде до зниження якості продукту. До порушення справжності продуктів можуть приводити такі дії [37, 38]:

- повна або часткова заміна компонентів;
- повна або часткова відсутність заявлених цінних компонентів;
- доповнення неоголошеною речовиною чи матеріалом із метою збільшення ваги, зниження якості або поліпшення зовнішньої привабливості продукту;
- недотримання норм технологічного процесу;

– недотримання законних вимог щодо стандарту продукту, його географічного походження, максимального або мінімального вмісту води тощо.

Отже, відносно продуктів харчування терміни «справжність» або «автентичність» позначають їх непідробленість, натуральність, відповідність указаним у сертифікатах сортовому, видовому і географічному походженню, а також технології їх переробки, відсутність нерегламентованих домішок і добавок [39].

Зіставлення показників досліджуваного продукту й автентичних зразків (еталонів), їх описів, опублікованих у відповідних документах, а також інформації, що міститься в супровідних документах і споживчих етикетках, із застосуванням аналітичних і органолептичних методів є актуальним завданням сучасного якісного хімічного аналізу [25, 40–51].

Якщо висновок про тотожність аналізованого продукту й автентичного зразка або про віднесення продукту до певного класу належить зробити на основі обробки багатовимірних масивів експериментальних даних, неминучим стає застосування хемометричних методів, серед яких особливе значення мають алгоритми класифікації, розпізнавання образів, дискримінантного аналізу та штучних нейронних мереж [52–58].

1.3.2. Експериментальні методи

Експресність і чутливість хроматографічного аналізу, можливість поєднання хроматографії з іншими фізико-хімічними методами [59–64] зробили його найбільш поширеною процедурою аналізу продуктів харчування, напоїв, об’єктів навколишнього середовища. На всіх стадіях виробництва продуктів і в контролі їх якості використовують різні види хроматографії: високоефективну рідинну (часто доповнену капілярним електрофорезом), іонну, газову, тонкошарову, міцелярну та інші види хроматографії.

Інтенсивно використовуються і різні спектроскопічні методи (видима спектроскопія [65], інфрачервона спектроскопія ближнього діапазону [65, 66], інфрачервона спектроскопія з Фур’є-перетворенням [67], Раманівська спектроскопія [68], ядерний магнітний резонанс [69, 70], мас-спектрометрія [70] (в тому числі ізотопна мас-спектрометрія [61, 71–74] та мас-спектрометрія з індуктивно зв’язаною плазмою [54]), атомно-абсорбційна спектрофотометрія [75–77], люмінесцентна спек-

троскопія [78]). Відносна простота підготовки зразків і виконання вимірювальних процедур, можливість досить просто одержувати великі масиви даних про властивості аналізованих об'єктів сприяли тому, що багатовимірні спектральні дані часто використовують як «відбитки пальців» об'єктів («fingerprints») [32, 79, 80].

Створення таких сенсорних систем, як «електронний ніс» (для аналізу газуватих зразків) та «електронний язик» (для аналізу рідких зразків) було викликано прагненням реалізувати аналітичні процедури, що моделюють нюхові і дегустаційні здібності людини. Системи «електронного язика» й «електронного носу» складаються з набору сенсорів різної селективності. З їх використанням одержують багатовимірні масиви характеристик аналізованих зразків, а потім за допомогою хемометричних методів отримують інформацію про досліджуваний зразок, яку можна співвіднести з людським сприйняттям, наприклад, смаку зразка (солоність, кислотність, гіркота), відомостями про його географічне походження, склад, інтенсивність аромату, ступінь свіжості тощо. Поширеними стали хімічні сенсори різних типів: вольтамперометричні, потенціометричні, спектрофотометричні, флуоресцентні, кондуктометричні, амперометричні та інші. Зростаюча популярність систем «електронного носу» й «електронного язика» пов'язана з можливостями їх мініатюризації, автоматизації роботи, простотою застосування й експресністю. В обробці масивів даних, що отримують на основі використання наборів сенсорів різної селективності, широке застосування знайшли хемометричні методи, зокрема методи розпізнавання образів і штучних нейронних мереж. З урахуванням досягнень теорії нейронних розрахунків ідентифікацію об'єктів аналізу за допомогою «електронного язика» та «електронного носу» можна розглядати як гілку розвитку штучного інтелекту та/або сферу застосування «електронного мозку» [81–90].

1.3.3. Хемометричні методи

Розвиток хемометрії швидко поповнює арсенал доступних хімікам алгоритмів обробки експериментальних даних. Багато підходів реалізовано як у спеціалізованих пакетах програм (Unscrambler, SIMCA, Eigenvector Research та інших), так і в статистичних пакетах загального призначення (Statistical Package for the Social Sciences, Statistica, Matlab).

Для попередньої обробки і вилучення інформації з багатовимірних масивів експериментальних даних використовують методи стиснення

даних, що дозволяє представити результати вимірювань у компактному виді, зручному для візуалізації та інтерпретації. Найбільш поширеною хемометричною технологією пониження розмірності масиву даних із мінімальною втратою інформації є метод головних компонент.

Метод головних компонент (Principal Component Analysis, PCA)

Метод головних компонент використовують в аналізі будь-яких складних хімічних даних, яким притаманна мультиколінеарність, тобто присутність внутрішніх, прихованих зв'язків між змінними. Основними передумовами для того, щоб ефективно понизити розмірність масиву даних, є «сильний» (аж до лінійного) зв'язок між початковими змінними, внаслідок чого інформація, що міститься в даних, дублюється; слабка інформативність деяких показників, що дозволяє виключити їх із набору даних; можливість об'єднання декількох показників в один [91, 92].

Алгоритм здійснюється ітераційно, його мета – розрахунок нових змінних, ортогональних і некорельованих головних компонент. Метод головних компонент застосовують до даних, записаних у виді матриці X розмірністю $I \times J$, рядки якої відповідають аналізованим зразкам $(i = 1,..., I)$, а стовпці – початковим змінним $x_j (j = 1,..., J)$, що характеризують ці зразки. Головні компоненти $t_a (a = 1,..., A)$ – це лінійна комбінація початкових змінних:

$$t_a = p_{a1} x_1 + ... + p_{aJ} x_J . \qquad (1.3)$$

За допомогою нових змінних початковий масив даних X розкладається на добуток двох матриць T і P:

$$X = TP^t + E = \sum_{a=1}^{A} t_a p_a^t + E . \qquad (1.4)$$

У цьому рівнянні T (розмірність $I \times A$) – матриця рахунків (scores), P (розмірність $J \times A$) – матриця навантажень (loadings), E (розмірність $I \times J$) – матриця залишків. Метод головних компонент можна розглядати як проектування даних на підпростір головних компонент із розмірністю, меншою, ніж початковий простір. Рядки $t_1,...,t_I$ матриці T – це координати зразків у новій системі координат, стовпці $t_1,...,t_A$ матриці T подають проекції всіх зразків на одну нову координатну вісь. Кожен рядок матриці P складається з коефіцієнтів, що зв'язують змінні t і x (1.3), кожен стовпець P – це проекція відповідної змінної x_j на нову систему координат [93].

У результаті застосування методу головних компонент здійснюється перехід до нового ортогонального базису, осі якого орієнтовані по напрямах максимальної дисперсії набору вхідних даних. Таке перетворення дозволяє стискувати інформацію шляхом відкидання координат, що відповідають напрямам із мінімальною дисперсією. Кожна головна компонента має дисперсію, максимально можливу з усіх комбінацій початкових змінних, за умови ортогональності попередній головній компоненті. Таким чином, задача зводиться до пошуку таких лінійних комбінацій початкових показників, які пояснювали б максимально можливу частку змінюваності (суму дисперсій) початкових показників. У більшості випадків багатопараметричний масив експериментальної інформації, що містить десятки змінних, можна представити у виді масиву, що містить дві-три змінні, оскільки перші дві або три головні компоненти відображають близько 80% початкових даних. Залишки (матриця *E*), що виникають при цьому, розглядають як шум, що не містить значущої хімічної інформації [94–96].

Як правило, застосуванню методу головних компонент передує центрування (віднімання з кожного стовпця x_j середнього по стовпцю значення m_j) або нормування (ділення кожного стовпця x_j на власне стандартне відхилення s_j) початкової матриці *X*, оскільки величини дисперсій показників істотно залежать від масштабу одиниць вимірювання [93]:

$$m_j = (x_{1j} + ... + x_{Ij}) / I ; \quad (1.5)$$

$$s_j = \sqrt{\sum_{i=1}^{I} (x_{ij} - m_j)^2 / I} . \quad (1.6)$$

За допомогою методу головних компонент прагнуть виявити взаємозв'язки даних, що складають великі масиви результатів первинних вимірювань. Отримання таких масивів експериментальних даних характерне для використання технології «електронного язика» [69, 97, 98], методів хроматографії [30, 99], інфрачервоної спектроскопії ближнього діапазону [100], ядерного магнітного резонансу [101] та інших.

Як приклад розглянемо застосування методу головних компонент при контролі якості різних видів квіткового меду. В роботі [99]

досліджено вміст сахарів у 280 зразках французького меду семи квіткових різновидів (акація, каштан, рапс, лаванда, ялина, липа, соняшник). За допомогою газової хроматографії з полум'яним іонізаційним детектором оцінювали вміст 17 ди- і трисахаридів, а рідинною хроматографією з імпульсним амперометричним детектором визначали вміст глюкози і фруктози (табл. 1.2). Справжність квіткового походження 280 зразків меду підтвердили на основі результатів пилкового, органолептичного і лабораторного аналізів. Автори поставили завдання визначити такі характеристики зразків меду, які забезпечували б їх надійну ідентифікацію. Для досягнення поставленої мети застосували метод головних компонент. Перед застосуванням методу головних компонент із отриманого масиву первинних експериментальних даних за допомогою дисперсійного аналізу (analysis of variance, ANOVA) відібрали сім таких характеристик зразків меду, як масові концентрації глюкози, рафінози, трехалози, фруктози і відношення концентрацій ерлоза/мальтулоза, мальтоза/трехалоза і фруктоза/глюкоза. Набір із цих характеристик забезпечував ідентифікацію 72% зразків меду.

Потім до масиву змінних, відібраних за допомогою дисперсійного аналізу, застосували метод головних компонент. У результаті роботи виділили дві головні компоненти, що є лінійною комбінацією вищеназваних характеристик. Таким чином, масив даних, що включав 20 характеристик зразків меду семи квіткових різновидів, було стиснуто до масиву розміром 2×7. Метод головних компонент дозволив визначити межі областей кожного квіткового різновиду меду, наприклад, область зразків ялинового меду визначилася чітко, але при цьому області інших квіткових різновидів меду перекрилися: акацієві і каштанові, лавандові і липові, рапсові і соняшникові. Фактично, незважаючи на відсутність повної диференціації, такий результат надав можливість контролювати якість меду відповідно до його квіткового різновиду.

Запропонований алгоритм застосували для оцінки справжності 47 комерційних зразків меду: 34 французьких (8 акацієвих, 8 каштанових, 7 лавандових, 6 ялинових, 5 липових), 7 угорських і 3 китайських (акацієвих), 1 іспанського і 1 марокканського (лавандових), 1 турецького (ялинового) і виявили, що 5 акацієвих і 4 ялинові комерційні зразки меду знаходяться за межами їх передбачуваної області, тобто не відповідають своєму заявленому виду.

Середні вмісти (M) і стандартні відхилення вмістів (SD) сахарів у 280 меду семи квіткових різновидів [99] *Таблиця 1.2*

Кількість зразків	Акація		Каштан		Рапс		Лаванда		Ялина		Липа		Соняшник	
	50		38		28		53		37		38		36	
	M[a]	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Fructose Фруктоза	438.9	12.9	407.2	22.8	379.2	15.2	384.6	13.0	314.9	19.9	381	26.3	399.8	22.6
Glucose Глюкоза	263.0	13.0	265.2	20.7	396.4	22.9	323.3	13.1	241.7	25.1	326	19.0	379.0	18.5
Ф/Г[b]	1.7	0.1	1.5	0.1	1.0	0.1	1.2	0.1	1.3	0.1	1.	0.1	1.1	0.1
Sucrose Сахароза	2.0	2.0	0.2	0.2	0.1	0.1	4.7	2.8	0.4	0.3	0.	0.6	0.1	0
Maltose Мальтоза	2.6	0.7	1.5	0.5	0.7	0.3	2.6	0.6	1.7	0.4	1.	0.4	1.1	0.4
Maltulose Мальтулоза	1.7	0.4	2.6	0.9	0.9	0.4	1.0	0.3	1.9	0.6	2.	1	0.5	0.2
Turanose Тураноза	2.9	0.6	2.8	0.8	1.3	0.6	1.8	0.5	2.5	0.4	3.	1.1	0.9	0.3
Trehalose Трехалоза	1.5	0.4	2.0	0.7	0.6	0.3	0.8	0.2	4.5	1.1	1	0.6	0.6	0.2
Palatinose Палатиноза (ізомальтулоза)	0.3	0.1	0.5	0.4	0.2	0.1	0.2	0.1	0.5	0.4	0.	0.2	0.1	0.0

Продовження табл. 1.2

Laminaribiose Ламінарібіоза	1.2	0.3	1.5	0.5	0.5	0.3	0.7	0.3	1.3	0.3	1.3	0.4	0.5	0.2
Melibiose Мелібіоза	0.1	0.1	0.3	0.3	НВ[c]		0.1	0.0	0.3	0.1	0.3	0.1	НВ	
Isomaltose Ізомальтоза	0.9	0.3	1.8	0.8	0.4	0.2	0.5	0.2	1.8	0.6	1.7	0.8	0.3	0.1
Gentiobiose Гентіобіоза	0.0	0.0	0.2	0.5	НВ		НВ		0.1	0.0	0.2	0.2	НВ	
Raffinose Рафіноза	0.0	0.1	0.0	0.1	НВ		0.1	0.0	2.1	0.6	НВ		НВ	
Neo-kestose Неокестоза	0.2	0.1	0.2	0.3	0.0	0.1	0.2	0.0	0.4	0.5	0.4	0.2	0.1	0.0
1-Kestose 1-Кестоза	0.1	0.0	0.1	0.1	НВ		0.1	0.0	НВ		0.3	0.1	НВ	
Erlose Ерлоза	1.9	1.2	0.2	0.2	НВ		1.4	0.6	2.1	1.3	1.0	0.6	НВ	
Melezitose Мелезітоза	0.1	0.1	0.2	0.4	НВ		0.1	5.7	4.2	0.2	0.1	НВ		
Maltotriose Мальтотріоза	0.4	0.2	0.2	0.1	НВ		0.2	0.1	0.8	0.3	0.5	0.2	0.1	0.0
Panose Паноза	0.2	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.3	0.2	0.4	0.2	0.1	0.1

[a] У процентних одиницях (окрім глюкози і фруктози – г / кг)
[b] Фруктоза/глюкоза
[c] Не виявлено

Формальне незалежне моделювання аналогій класів (Soft Independent Modeling of Class Analogy, SIMCA)

Запропоноване одним із засновників хемометрії, видатним шведським ученим Сванте Волдом формальне незалежне моделювання аналогій класів стало популярним і потужним засобом класифікації з навчанням [91, 102].

Цей метод володіє низкою переваг. По-перше, кожен клас моделюється окремо, незалежно від інших. По-друге, класифікація SIMCA є багатозначною (soft): кожен зразок може бути одночасно віднесений до декількох класів. По-третє, SIMCA надає можливість установити значення помилки першого роду і побудувати відповідний класифікатор.

Класифікація об'єктів за допомогою SIMCA відбувається в два етапи: етап навчання і етап тестування. На першому етапі формуються окремі моделі класів. Кожен клас із навчальної множини об'єктів незалежно моделюється методом головних компонент із різною кількістю головних компонент *A* [91, 103].

Під моделлю в даному випадку розуміють спосіб відтворення форми та/або інших характеристик складного первинного масиву в простішій формі, що включає також і набір даних, що характеризують властивості модельованої системи, і динаміку їх зміни.

Для порівняння створених моделей розраховують відстань між класами і оцінюють вплив змінних на розподіл зразків між класами (модельну потужність і дискримінаційну потужність змінних) [104, 105].

Значення модельної потужності змінної (M_j) показує, наскільки сильний вплив чинить дана змінна на моделювання даного класу (1 – сильний вплив, 0 – впливу немає),

$$M_j = 1 - \frac{S_{jresid}}{S_{jraw}}, \qquad (1.7)$$

де S_{jraw} – стандартне відхилення початкових значень змінної, S_{jresid} – стандартне відхилення залишків, що визначаються методом головних компонент для відповідної моделі даного класу.

Значення дискримінаційної потужності змінної (D_j) показує здатність змінної розділяти класи (здатність змінної моделювати клас не тягне за собою здатність до розділення на класи). Нехай наявні по

дві моделі для класів X_1 і X_2, тоді значення дискримінаційної потужності тієї змінної, що нас цікавить, відповідає такій формулі:

$$D_j = \sqrt{\frac{S^2_{jresid}(classX_1 \text{model}B) + S^2_{jresid}(classX_2 \text{model}A)}{S^2_{jresid}(classX_1 \text{model}A) + S^2_{jresid}(classX_2 \text{model}B)}}. \quad (1.8)$$

Чим більше значення дискримінаційної потужності змінної, тим більше її здатність до розділення класів.

Для розрахунку відстані між класами використовують відстань Махаланобіса; чим вище значення цього параметра, тим краще розрізняються моделі:

$$H = (\mu_1 - \mu_2)^T \cdot \text{cov}^{-1}(X) \cdot (\mu_1 - \mu_2), \quad (1.9)$$

де μ_1, μ_2 – вектори середніх значень змінних для класів X_1, X_2, $\text{cov}(X)$ – об'єднана коваріаційна матриця для класів X_1, X_2:

$$\text{cov}(X) = \frac{\text{cov}(X_1) + \text{cov}(X_2)}{(n_1 + n_2 - 2)}, \quad (1.10)$$

де n_1, n_2 – довжини векторів μ_1, μ_2.

На етапі тестування нові об'єкти, не використані на стадії навчання, відносять до класів, сформованих на першому етапі. Етап класифікації нових зразків включає розрахунок відстаней від зразка до центру класу (H_1) і від зразка до класу (H_2) [93, 104–108]:

$$H_1 = \frac{1}{I} + \sum_{a=1}^{A} \frac{\tau_a^2}{t_a^T t_a}, \quad (1.11)$$

$$H_2 = \sqrt{\frac{1}{J - A} \sum_{j=1} E_j^2}, \quad (1.12)$$

де I, J – кількості зразків і змінних, відповідно, τ_a – проекція нового зразка на головну компоненту a, t_a – вектор, що містить рахунки всіх навчальних зразків у класі.

Відстань від нового зразка до класу показує, наскільки далеко зразок знаходиться від даного класу. Відстань від зразка до центру класу розраховується як розмах і показує, наскільки проекція зразка на даний клас далека від його центроїда, тобто наскільки він відрізняється від інших зразків даного класу.

Такі моделюючі властивості і можливість віднесення зразка до одного або декількох класів або до жодного із змодельованих класів (на відміну від класичних методів класифікації, які дозволяють віднести зразок лише до одного класу) забезпечили методу формального незалежного моделювання аналогій класів популярність при визначенні справжності продуктів харчування, особливо в комбінації зі спектроскопічними методами [107–114].

Прикладом успішного застосування алгоритму SIMCA є робота [79], в якій його використали для класифікації вин за їх типом, походженням винограду і технологією витримки. Для цього отримали «відбитки пальців» (нормалізовані спектри в середній ІЧ-області) 60 зразків червоних вин і 60 зразків білих вин. Потім розрахували матриці подібності спектрів на основі індексу подібності Танімото:

$$T_{A,B} = \frac{c}{a+b-c}, \qquad (1.13)$$

де c – множина фрагментів (частинок) «відбитків пальців», загальних для сполук A і B; a – множина фрагментів «відбитків пальців» сполуки A, відсутніх у сполуки B; b – множина фрагментів «відбитків пальців» сполуки B, відсутніх у сполуки A.

Помилки класифікації за типом вина (біле або червоне) в результаті застосування SIMCA до спектрів зразків і до матриць подібності спектрів склали 20% і 10%, відповідно. Відношення кількості помилково класифікованих об'єктів до загальної кількості об'єктів слугує простою статистичною (евристичною) оцінкою ненадійності ідентифікації. При застосуванні SIMCA до матриць спектральної подібності для розрізнення вин за видом винограду і способом витримки помилки склали 9% і 5%, відповідно. Підсумком роботи стала демонстрація можливості створити бібліотеку «відбитків пальців» різних зразків вин.

Дискримінантний аналіз

Дискримінантний аналіз (Discriminant Analysis, DA) є розділом багатовимірного статистичного аналізу, який включає методи класифікації багатовимірних спостережень за принципом максимальної схожості за наявності навчальних ознак. Основними завданнями дискримінантного аналізу є дослідження міжкласової відмінності (дискримінація) в апріорі за декількома змінними одночасно, тобто

визначення «внеску» кожної зі змінних у розрізнення класів, а також класифікація об'єктів, що не входили до вибірки, яка навчається. Змінні, якими оперують для пошуку відмінностей між класами, називають дискримінантними змінними. В дискримінантному аналізі знаходять таку комбінацію дискримінантних змінних і дискримінантну функцію, яка б оптимально розділяла класи, що розглядаються [103–115].

Здійснення дискримінантного аналізу передбачає наявність мінімум двох класів і мінімум двох об'єктів у кожному класі, лінійну незалежність дискримінантних змінних, багатовимірну нормальність розподілу дискримінантних змінних для кожного класу, приблизну рівність коваріаційних матриць для усіх класів.

Основними проблемами є відбір дискримінантних змінних і вибір виду дискримінантної функції (лінійна або нелінійна). Найчастіше використовується лінійна форма дискримінантної функції такого виду [116]:

$$d_{km} = \beta_0 + \beta_1 x_{1km} + ... + \beta_p x_{pkm};\ m = 1,...,n;\ k = 1,...,g\ , \qquad (1.14)$$

де d_{km} – значення дискримінантної функції для m-го об'єкта в класі k, x_{ikm} – значення дискримінантної змінної x_i для m-го об'єкта в класі k, β_i – дискримінантні множники.

Опишемо коротко процедуру дискримінантного аналізу при використанні лінійної дискримінантної функції (Linear Discriminant Analysis, LDA) [116–120].

Вибір дискримінантних змінних можна проводити методом їх покрокового включення або покрокового виключення.

Процедура покрокового включення починається з вибору змінної, що забезпечує найкраще одновимірне розрізнення. Потім аналізують пари, утворені відібраною змінною і однією з тих, що залишилися. В результаті знаходять пару, що дає найкраще розрізнення. На кожному наступному кроці процедури відбирають змінні, які в поєднанні з відібраними раніше дають найкраще розрізнення.

Процедура покрокового виключення виходить із припущення, що всі змінні входять до системи, а потім на кожному кроці відкидається одна зі змінних – та, що в поєднанні зі змінними, що залишилися, дає найгірше розрізнення.

Процедура покрокового дискримінантного аналізу передбачає перевірку (на початку кожного кроку) всіх дискримінантних змінних на відповідність двом умовам: необхідній точності розрахунку (толерантності) і перевищення заданого рівня розрізнення (за *F*-критерієм Фішера). Статистика *F*-введення оцінює поліпшення розрізнення завдяки використанню даної змінної порівняно з розрізненням, досягнутим за допомогою вже відібраних змінних. Статистика *F*-виключення оцінює значущість погіршення розрізнення після видалення змінної зі списку вже відібраних змінних. Як відомо, критерій Фішера застосовують для перевірки рівності дисперсій двох вибірок. Якщо розраховане значення *F*-критерію більше критичного для певного рівня значущості і відповідної кількості ступенів свободи для чисельника і знаменника, то дисперсії визнають відмінними:

$$F = \frac{\sigma_1^2}{\sigma_2^2}, \tag{1.15}$$

де σ_1^2 – більша дисперсія, σ_2^2 – менша дисперсія.

Значення толерантності змінної визначається як $(1 - R_i^2)$, де R_i^2 – квадрат коефіцієнта множинної кореляції *i*-ї змінної з усіма іншими змінними (припустимо, що ми розраховуємо толерантність для змінної y, причому змінні x, z вже відібрані для розрізнення класів) [121, 122]:

$$L = 1 - \frac{r_{yx}^2 + r_{yz}^2 - 2 r_{yx} r_{yz} r_{xz}}{1 - r_{xz}^2}, \tag{1.16}$$

де r_{yx}, r_{yz}, r_{xz} – коефіцієнти лінійної парної кореляції;

$$r_{yx} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}. \tag{1.17}$$

Значення толерантності слугує мірою надлишковості змінної. Наприклад, якщо змінна, призначена для включення до дискримінантної функції, має значення толерантності $L = 0{,}01$, то вона може розглядатися як на 99% надлишкова порівняно з уже включеними змінними.

Коефіцієнти β_i першої дискримінантної функції вибираються так, щоб центроїди різних класів якомога більше відрізнялися один від одного. Коефіцієнти другої функції вибираються так само, але при цьому значення другої функції мають бути некорельованими зі значеннями першої. Аналогічно визначаються й інші функції.

Для отримання коефіцієнтів β_i дискримінантної функції потрібний статистичний критерій розрізнення класів. Класифікація змінних здійснюватиметься тим краще, чим менше розсіяння точок відносно центроїда усередині групи і чим більша відстань між центрами груп. Таким чином, метод пошуку найкращої дискримінації даних полягає в знаходженні такої дискримінантної функції, яка б максимізувала відношення міжгрупової варіації (матриця B) до внутрішньогрупової (матриця W). Елементи цих матриць визначаються виразами:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk}), \tag{1.18}$$

$$b_{ij} = \sum_{k=1}^{g} n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j), i, j = 1,..., p, \tag{1.19}$$

де g – кількість класів; p – кількість дискримінантних змінних; n_k – кількість спостережень в k-й групі; x_{ikm}, x_{jkm} – величина змінної i (j) для m-го спостереження в k-й групі; $\bar{x}_{ik}$, $\bar{x}_{jk}$ – середнє значення змінної i (j) в k-й групі; $\bar{x}_i$, $\bar{x}_j$ – середнє значення змінної i (j) по всіх групах.

Матриці W і B містять всю основну інформацію про залежності усередині груп і міжгрупові залежності. Знаходження коефіцієнтів дискримінантних функцій зводиться до розв'язання задачі на власні значення і вектори:

$$\sum b_{1i} v_i = \lambda \sum w_{1i} v_i, \tag{1.20}$$

$$\sum b_{2i} v_i = \lambda \sum w_{2i} v_i, \tag{1.21}$$

$$\sum b_{pi} v_i = \lambda \sum w_{pi} v_i, \tag{1.22}$$

де λ – власне число, v_i – власний вектор.

Кожний розв'язок, що має своє власне значення λ і свою послідовність ν_i, відповідає одній дискримінантній функції. Компоненти власного вектора ν_i можна використовувати як коефіцієнти дискримінантної функції

$$\beta_i = \nu_i \sqrt{n-g}, \beta_0 = -\sum_{i=1}^{p} \beta_i \bar{x}_i , \tag{1.23}$$

де n – загальна кількість спостережень в усіх групах.

Загальна кількість дискримінантних функцій не перевищує кількості дискримінантних змінних і, принаймні, на одиницю менше кількості груп. Існує декілька характеристик, що дозволяють оцінити корисність дискримінантної функції: коефіцієнт кореляції r_i

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}} , \tag{1.24}$$

Λ – статистика Уїлкса

$$\Lambda = \prod_{i=k+1}^{g} \left(\frac{1}{1+\lambda_i} \right), \tag{1.25}$$

де k – кількість розрахованих функцій.

Чим більше значення r_i, тим краще розділяюча здатність дискримінантної функції. Чим менше значення Λ, тим вище значущість відповідної дискримінантної функції.

Новий об'єкт відноситься до класу k, для якого значення дискримінантної функції d_{km} є максимальним.

При невиконанні вказаних вище вимог використання даного алгоритму може викликати проблеми. Проте при практичному застосуванні алгоритму дискримінантного аналізу для розв'язання різноманітних задач класифікації і перевірки справжності встановлено його достатньо високу стійкість [123–125].

Так, LDA успішно застосовували для класифікації 165 зразків французького і баскського сидру, кожен з яких характеризували набором із 27 ознак, визначених високоефективною рідинною хроматографією і вказуючих на вміст різних поліфенолів у зразках [29]. Випадковим чином масив даних розбили на навчальну (75% зразків) і тестову (25% зразків) вибірки. Результати вико-

ристання методу лінійного дискримінантного аналізу такі: на етапі навчання частки правильно віднесених зразків французького і баскського сидру склали 99% і 100%, а на етапі класифікації тестових зразків – 96% і 100%, відповідно.

Разом з тим, відомі і задачі, при розв'язанні яких лінійний дискримінантний аналіз приводив до незадовільних результатів (класифікація зразків свинини з помилкою 32% [103], класифікація зразків сиру з помилкою 44% [126]).

Слід зазначити, що LDA недоцільно застосовувати при перекриванні класів. Для розв'язання таких задач більш перспективними є алгоритми, засновані на теорії нечітких множин. Теорію нечітких множин запропонував понад 40 років тому американський математик Лотфі Заде. Методи на основі нечіткої логіки дозволяють працювати з даними, яким властиві невизначеність і неточність, і допускається можливість часткової приналежності об'єктів до декількох класів (один і той же об'єкт може одночасно відноситися до різних класів, але з різним ступенем приналежності).

Нечіткий лінійний дискримінантний аналіз (Fuzzy Linear Discriminant Analysis, FLDA) є вдосконаленим алгоритмом лінійного дискримінантного аналізу. Він здатний надати більше інформації про структуру наборів даних, що вивчаються [115, 127–129].

FLDA складається з таких кроків [130]:

1. Розрахунок нечітких приналежностей кожного зразка до всіх класів за допомогою нечіткого алгоритму *k*-середніх (Fuzzy C-means, FCM).

FCM-алгоритм припускає, що об'єкти належать усім кластерам із певним ступенем приналежності. Ступінь приналежності визначається відстанню від об'єкта до відповідних центрів кластерів (найчастіше використовують Евклідову метрику). Цей алгоритм ітераційно розраховує координати центрів кластерів і нові ступені приналежності об'єктів.

Алгоритм заснований на мінімізації цільової функції

$$J = \sum_{k=1}^{g} \sum_{m=i}^{n_k} u_{mk}^{l} \sqrt{(x_{ikm} - c_k)^2}\ , \qquad (1.26)$$

де c_k – центр k-го класу, u_{mk} – ступінь приналежності об'єкта m k-му кластеру, l – експоненціальна вага – дійсне число, більше за 1.

Завданням FCM-алгоритму є розбиття набору об'єктів на задану кількість класів. Первинна матриця приналежностей об'єктів до класів генерується випадковим чином. Потім розраховуються координати центрів класів, Евклідові відстані і нові значення приналежностей об'єктів до класів:

$$c_k = \frac{\sum_{m=1}^{n_k} u_{mk}^{l} x_{ikm}}{\sum_{m=1}^{n_k} u_{mk}^{l}}; \tag{1.27}$$

$$u_{mk} = \frac{1}{[(x_{ikm} - c_k)^2 \sum_{j=1}^{c} \frac{1}{(x_{imk} - c_k)^2}]^{\frac{2}{l-1}}}. \tag{1.28}$$

Робота алгоритму припиняється, коли квадрат різниці між набутими значеннями приналежностей і значеннями приналежностей, отриманих на попередній ітерації, стає менше заданого параметра точності $\left\| u - u^* \right\| < \varepsilon$ [127, 131, 132].

2. Розрахунок матриць W і B внутрішньокласових і міжкласових варіацій. Показники, що входять до формул розрахунку внутрішньокласових і міжкласових варіацій (1.18), (1.19), розраховують так:

$$n_k = n_k u_{mk}, \quad x_{ikm} = u_{km} x_{ikm}, \tag{1.29}$$

$$\bar{x}_{ik} = \frac{\sum_{m=1}^{n_k} u_{mk} x_{ikm}}{n_k}, \tag{1.30}$$

$$\bar{x}_i = \frac{\sum_{k=1}^{g} \sum_{m=1}^{n_k} u_{mk} x_{ikm}}{\sum_{k=1}^{g} n_k}. \tag{1.31}$$

Аналогічно розраховують x_{jkm}, $\bar{x}_{jk}$, $\bar{x}_j$.

3. Відбір дискримінантних змінних і розрахунок дискримінантних функцій.

Застосування алгоритму FLDA дещо підвищує ефективність класифікації порівняно з методом LDA [133]. Так, показано, що при розбитті набору зразків арахісу за даними щодо їх спектрів у ближній ІЧ-області на 3, 5 і 6 груп методом LDA точність класифікації склала 42%, 56% і 70%, відповідно, а при використанні алгоритму FLDA точність розділення підвищилася до 45%, 63% і 73%.

Дерева класифікації і регресії (Classification and Regression Trees, CART)

Дерева регресії і класифікації, відомі також під загальною назвою як дерева прийняття рішень (Decision Tree, DT), є популярним методом розв'язання задач класифікації і прогнозування з навчанням [134−139]. Після навчання дерева прийняття рішень спостереження, що класифікуються, представляються як послідовності конструкцій «If-Then» («Якщо-То»), організованих у вигляді дерева.

Будь-яке дерево рішень, по суті, є деревовидним графом. Ця структура даних складається з вузлів, сполучених один з одним ребрами (гілками). При цьому не допускається, щоб ребра утворювали цикл, оскільки тоді дерево перетворюється на граф, відмінний від деревовидного. Дерево має один особливий вузол – кореневий (корінь). Кореневий вузол є основою дерева, оскільки від нього можна перейти по дереву до будь-якого іншого вузла. В кінці ланцюжка підряд ідучих ребер знаходяться листові вузли (листя, термінальні вузли). Корінь дерева містить усі дані, що класифікуються, а листя – певні класи, отримані в результаті класифікації. Проміжні вузли дерева є пунктами прийняття рішення про вибір або виконання процедур тестування, а гілки, що виходять із вузлів, відповідають кількості можливих результатів процедури тестування. Вузол, який можна розбити на два нові, називається батьківським (вузол-батько), а нові вузли – вузлами-нащадками (підлеглими).

Для побудови дерева рішень алгоритм спочатку створює кореневий вузол. В алгоритмі CART кожен вузол дерева рішень має двох нащадків. На кожному кроці побудови дерева рішень правило, сформоване у вузлі, ділить навчальну вибірку на дві частини – частину, в якій виконується правило (нащадок-right), і частину, в якій правило не виконується (нащадок-left). У кожному вузлі розбиття йде тільки по одній змінній. Загальне правило для вибору змінної можна

сформулювати так: вибрана змінна повинна розбити вибірку даних так, щоб отримані в результаті підмножини складалися з об'єктів, що належать до одного класу (або були максимально наближені до цього, тобто кількість об'єктів з інших класів у кожній з цих підмножин має бути якомога меншою). На кожному кроці побудови дерева алгоритм послідовно порівнює всі можливі розбиття для всіх змінних і вибирає найкраще розбиття.

Побудова дерева прийняття рішень відноситься до задач навчання з учителем. Якість розбиття оцінюють за такою формулою:

$$G = \frac{1}{N}\left[L\left(1 - \frac{1}{L^2}\sum_{i=1}^{n} l_i^2\right) + R\left(1 - \frac{1}{R^2}\sum_{i=1}^{n} r_i^2\right)\right], \qquad (1.32)$$

де N – кількість зразків у вузлі-батьку; L і R – кількість зразків у лівому і правому вузлах-нащадках, відповідно; l_i та r_i – кількість об'єктів i-го класу в лівому / правому вузлі-нащадку.

Кращим буде те розбиття, для якого значення функції G максимальне.

Недоліком методу є те, що дерево може виявитися перенавченим. Тому наступний етап алгоритму CART – відсікання (скорочення, зменшення) дерева (minimal cost-complexity tree pruning). Основна проблема відсікання – велика кількість можливих відсічених піддерев для одного дерева. Позначимо через T кількість термінальних вузлів дерева, а через $R(T)$ – помилку класифікації дерева (відношення кількості неправильно класифікованих об'єктів до кількості об'єктів у навчальній вибірці). Визначимо $C_\alpha(T)$ – повну вартість (показник витрати-складність) дерева T – як

$$C_\alpha(T) = R(T) + \alpha T, \qquad (1.33)$$

де α – параметр, що змінюється від 0 до $+\infty$.

Повна вартість дерева включає дві компоненти – помилку класифікації і штраф за його складність. Для скорочення необхідно перевірити пари вузлів, що мають загального батька, і з'ясувати, наскільки збільшиться $C_\alpha(T)$ при їх об'єднанні. Якщо збільшення опиниться менше заданого порогу, то обидва листові вузли зливаються в один, для якого множину результатів отримують об'єд-

нанням результатів у початкових листях. Ідея полягає в тому, щоб знайти значення α для кожного вузла t в дереві T і вибрати «слабкі зв'язки», тобто вузли t, для яких значення

$$\alpha(t) = \frac{R(t) - R(T)}{T - 1} \tag{1.34}$$

є найменшим. Так уникають феномену перенавчання.

Побудова дерева регресії багато в чому схожа з побудовою дерева класифікації. Спочатку будують дерево максимального розміру, яке потім обрізують до оптимального розміру. Процес побудови дерева відбувається послідовно. На першому кроці отримуємо регресійну оцінку як константу по всьому простору об'єктів. Константа розраховується як середнє арифметичне вихідної змінної в навчальній вибірці. Якщо позначити всі значення вихідної змінної як $Y_1, Y_2, ..., Y_n$, то регресійну оцінку знаходять як

$$\hat{f}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) I_R(x), \tag{1.35}$$

де R – простір навчальних об'єктів, n – кількість об'єктів, $I_R(x)$ – індикаторна функція простору (набір правил, що описують попадання змінної x до простору).

На другому кроці простір ділять на дві частини. Регресійна оцінка набуває виду

$$\hat{f}(x) = \left(\frac{1}{I_1} \sum_{I_1} Y_i \right) I_{R1}(x) + \left(\frac{1}{I_2} \sum_{I_2} Y_i \right) I_{R2}(x). \tag{1.36}$$

Оцінкою кращого розбиття слугує сума квадратів різниць

$$E = \sum_{i=1}^{n} (Y_i - \hat{f}(x_i))^2, \tag{1.37}$$

яку потрібно мінімізувати.

Процес розбиття продовжується до тих пір, поки сума квадратів різниць не стане менше певного наперед заданого порогу.

Відсікання і вибір фінального дерева проводиться за процедурами, аналогічними використаним для дерева класифікації. Єдина

відмінність полягає в оцінюванні помилки відповіді дерева, яку розраховують за формулою

$$R(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2 . \quad (1.38)$$

Вартість дерева дорівнює

$$C_\alpha(\hat{f}) = R(\hat{f}) + \alpha\hat{f} . \quad (1.39)$$

Після створення дерева рішень з'являється можливість класифікувати невідомі вибірки даних. Визначення класу здійснюється шляхом переходу по дереву, починаючи від кореневого вузла, у напрямі до листових вузлів [134–139].

До переваг методу побудови дерев регресії і класифікації відносять їх ієрархічну будову та можливість послідовно вивчати ефект впливу окремих змінних.

Таблиця 1.3

Результати застосування алгоритму CART для класифікації зразків пшениці, вершків і зеленого чаю за даними ІЧ-спектроскопії

Об′єкт	Навчальна вибірка	Тестова вибірка	Ненадійність класифікації, %
	(кількість зразків)		
Пшениця	70	30	13
Вершки	450	537	5
Зелений чай	70	50	18

Розглянемо приклад використання алгоритму. Для 100 зразків пшениці, 987 зразків вершків і 120 зразків зеленого чаю отримали ІЧ-спектри і розділили, застосовуючи алгоритм CART, кожен із отриманих масивів даних на два класи [136]. У табл. 1.3 представлено оцінки ненадійності роботи алгоритму. Оцінкою ненадійності слугує помилка класифікації зразків тестової вибірки (відношення кількості неправильно класифікованих зразків тестової вибірки до загальної кількості зразків у цій вибірці).

Метод опорних векторів (Support Vector Machines, SVM)

Метод опорних векторів – це група алгоритмів, заснованих на навчанні з учителем. Його використовують для розв'язання задач

бінарної класифікації та регресійного аналізу. В основі методу лежить поняття площини розв'язків, що визначають границі прийняття рішень, тобто розділення об'єктів у просторі змінних здійснюють за допомогою побудови гіперплощини [95, 140–142].

Гіперплощину будують на основі навчальної вибірки – множини об'єктів x_i, заданих у виді векторів змінних з явним позначенням приналежності до одного з класів c_i (+1, –1). Гіперплощину будують так, щоб максимізувати ширину границі (ширину смуги) між позитивною і негативною частинами навчальної вибірки.

Таким чином, у побудові гіперплощини беруть участь лише опорні вектори (support vectors), тобто об'єкти на границі між частинами навчальної вибірки [143, 144].

Гіперплощину можна представити у виді

$$w \cdot x = b. \tag{1.40}$$

Необхідно знайти вектор w такий, щоб для певного граничного значення b і нової точки x_i виконувалася умова

$$\begin{aligned} w \cdot x_i > b &\Rightarrow c_i = 1; \\ w \cdot x_i < b &\Rightarrow c_i = -1. \end{aligned} \tag{1.41}$$

Якщо скалярний добуток вектора w на x_i більше певного порогового значення b, то нову точку відносять до першої категорії, якщо менше – до другої. Вектор w перпендикулярний шуканій розділяючій прямій, а значення b залежить від найкоротшої відстані між розділяючою прямою і початком координат [145, 146].

Якщо навчальна вибірка є лінійно розділимою, то обрати гіперплощини можна так, щоб між ними не лежала жодна точка навчальної вибірки, а потім максимізувати відстань між гіперплощинами. Ширина смуги між ними

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+ \rangle - \langle w, x_- \rangle}{\|w\|} = \frac{(w_0 + 1) - (w_0 - 1)}{\|w\|} = \frac{2}{\|w\|}. \tag{1.42}$$

Ширина смуги максимальна, коли норма вектора w мінімальна.

Для виключення всіх точок зі смуги необхідно переконатися, що для всіх i $(1 \le i \le n)$ виконується умова

$$c_i(w \cdot x_i - b) \ge 1. \qquad (1.43)$$

Оптимальну розділяючу гіперплощину будують, мінімізуючи $\|w\|$ з урахуванням умови (1.43) [146, 147].

При роботі з багатопараметричними даними рекомендують визначити найбільш інформативні змінні. Їх вибір ґрунтується на розрахунку значення *F* для кожної змінної:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,j}^{(-)} - \bar{x}_i^{(-)})^2}, \qquad (1.44)$$

де n_+ – кількість позитивних зразків, n_- – кількість негативних зразків, $\bar{x}_i^{(+)}, \bar{x}_i^{(-)}, \bar{x}_i$ – середні значення i-ї ознаки в позитивній, негативній частинах і в усій вибірці, відповідно, $x_{k,i}^{(+)}, x_{k,i}^{(-)}$ – значення i-ї ознаки k-го позитивного і негативного об'єктів, відповідно.

Змінні сортують за значеннями *F*, причому змінні з найменшими значеннями *F* відкидають [148].

У випадку лінійно роздільної вибірки розв'язок задачі квадратичної оптимізації має вид

$$w = \sum_{i=1}^{n} \lambda_i c_i x_i , \qquad (1.45)$$

де λ_i – вектор подвійних змінних. Якщо $\lambda_i > 0$, то об'єкт навчальної вибірки x_i має назву опорного вектора [144, 149].

Але в більшості випадків задачі класифікації не такі прості (наприклад, за умови перекривання класів), і часто необхідно будувати оптимальні розбиття множин об'єктів за допомогою роздільників набагато складнішої структури, ніж лінійний роздільник. Для цього початкові об'єкти переупорядковують за допомогою спеціального класу математичних функцій, званих ядрами. Цей процес переупорядковування називають ще перетворенням об'єктів (перегрупуванням). Тоді класифікуюча функція має вид [148, 150]

$$g(x) = \sum_{i=1}^{n} \lambda_i c_i K(x_i, x) + b, \qquad (1.46)$$

де λ_i – коефіцієнти, що визначаються в ході оптимізації, x_i – опорні вектори, K – ядерна функція (kernel function), яку використовують для переходу до нелінійної розділяючої поверхні.

Найбільш поширеними є такі ядра [147, 151]:

– поліноміальне $\langle \quad \rangle$

Таблиця 1.4

Результати класифікації зразків сиру методом опорних векторів

Ненадійність класифікації, %						
Вид 1	Вид 2	Вид 3	Вид 4	Вид 5	Вид 6	Середнє значення
4	24	34	26	0	14	17

Проекція на латентні (приховані) структури (Projection on Latent Structures, PLS) і PLS-регресія

Існують два тлумачення фундаментального методу PLS: «Partial Least Squares» – «часткові» або «окремі найменші квадрати» і «Projection on Latent Structures», що дослівно перекладається як «проекція на латентні структури». Рекомендують використовувати друге тлумачення, оскільки воно відображає суть методу [91, 106].

PLS-аналіз – популярний метод багатовимірного градуювання, він являє собою потужну альтернативу методу головних компонент, оскільки надає можливість тлумачити результати із використанням меншої кількості головних компонент (у даному випадку – PLS-компонент) [101, 106, 153].

На відміну від методу головних компонент, у методі PLS проводиться одночасна декомпозиція (розкладання, розбиття) двох матриць: X (предиктори, незалежні змінні, наприклад, спектральні смуги) та Y (відгуки, спостереження, залежні змінні, наприклад, концентрації компонента):

$$X = TP^t + E = t_1 p_1^{'} + t_2 p_2^{'} + ... + t_n p_n^{'} + E , \qquad (1.47)$$

$$Y = UQ^t + F = u_1 q_1^{'} + u_2 q_2^{'} + ... + u_n q_n^{'} + F , \qquad (1.48)$$

де T і U – матриці рахунків, P і Q – матриці навантажень, E та F – матриці залишків.

Існують два популярні різновиди методу проекцій на латентні структури: PLS1 і PLS2. Модель PLS1 будують для єдиної змінної y, наприклад, для концентрації одного компонента в суміші. Модель PLS2 розраховують для декількох змінних Y одночасно, що дозволяє моделювати будь-яку комбінацію змінних спільно. Відповідним чином відрізняються і розрахункові алгоритми цих методів (табл. 1.5, 1.6), на кожній ітерації розраховується одна PLS-компонента. Як правило, методу PLS передує центрування або нормування первинних матриць X і Y [154].

Таблиця 1.5

Алгоритм PLS1

№ кроку	Процедура	Пояснення
1	$w_i = \frac{X_i^t y_i}{\left\|X_i^t y_i\right\|}$	Розрахунок нормалізованого вектора зважених навантажень w
2	$t_i = X_i^t w_i$	Розрахунок вектора ваг t
3	$q_i = \frac{y_i^t t_i}{\left\|t_i^t t_i\right\|}$	Розрахунок навантаження q змінної i
4	$p_i = \frac{X_i^t t_i}{t_i^t t_i}$	Розрахунок вектора навантажень p
5	$E_i = X_i - T_i^t P_i$ $F_i = y_i - b_i T_i Q_i^T$	Розрахунок залишків
6	$i = i + 1$	Перехід до наступної PLS-компоненти

Таблиця 1.6

Алгоритм PLS2

№ кроку	Процедура	Пояснення
1	u_i	Вибір початкового наближення u
2	$w_i = \frac{X_i^t u_i}{\left\|X_i^t u_i\right\|}$	Розрахунок нормалізованого вектора зважених навантажень w
3	$t_i = X_i^t w_i$	Розрахунок вектора рахунків t
4	$q_i = \frac{Y_i^t t_i}{\left\|Y_i^t t_i\right\|}$	Розрахунок нормалізованого вектора навантажень q
5	$u_i = Y_i^t q_i$	Розрахунок вектора рахунків u
6	$\left\|t_{i,new} - t_{i,old}\right\| < \lim$	Перевірка збіжності алгоритму: якщо умова не виконується – перехід до кроку 2
7	$p_i = \frac{X_i^t t_i}{t_i^t t_i}$	Розрахунок вектора навантажень p
8	$b_i = \frac{u_i^t t_i}{t_i^t t_i}$	Розрахунок внутрішнього коефіцієнта регресії

Продовження табл. 1.6

9	$E_i = X_i - T_i^t P_i$ $F_i = Y_i - b_i T_i Q_i^T$	Розрахунок залишків
10	$i = i + 1$	Перехід до наступної PLS-компоненти

Розкладання матриць X і Y методом PLS2 тісно зв'язані одне з одним. Перший набір значень векторів-навантажень p_1 і q_1 отримують з умови максимізації коваріації (міри лінійної залежності) між відповідними векторами X та Y. Проектуванням даних Y на q_1 отримують перший набір значень вектора рахунків u_1. Матриці X і Y опосередковано зв'язані через вектори-рахунки: вектор рахунків u_1 є відправним пунктом для розрахунку вектора t_1-рахунків. При цьому в методі PLS2 початковий вектор t_1 замінюється на u_1, і, таким чином, структура даних Y безпосередньо впливає на декомпозицію матриці X. Вектор u_1 (відображає структуру Y), за допомогою якого вибирається перший напрям у декомпозиції матриці X, дає нові навантаження для матриці X, що позначаються символом w (loadings-weights – зважені навантаження).

Після цього розраховуються нові t-вектори простору X із використанням векторів навантажень w. Потім ці t-вектори застосовують як стартові, замінюючи u_1. Таким чином, структура даних X також впливає на декомпозицію Y. Ітераційна заміна незалежних векторів $u_i \to t_i$ та $t_i \to u_i$ (обмін векторами рахунків) здійснюється до досягнення збіжності. Вектор w_1 визначає напрям першої PLS-компоненти в просторі X, на який проектуються всі зразки. Звичайно цей напрям не збігається з напрямом p_1, оскільки в PLS-алгоритмі одночасно проводиться і максимізація коваріації (t,u). Відмінність між напрямами цих альтернативних компонент показує, наскільки сильно матриця Y вплинула на розбиття матриці X [153, 155, 156].

Значущість обох матриць P та W стає очевидною, якщо записати формулу розрахунку коефіцієнта B формального регресійного рівняння $Y = XB$ [153]:

$$B = W(P^t W)^{-1} Q^t . \qquad (1.49)$$

Значення B часто використовують у практичних цілях при класифікації нових об'єктів:

$$\hat{Y} = X_{new} B . \qquad (1.50)$$

Алгоритм PLS1 простіший, ніж PLS2: відсутня ітераційна підміна $u_i \to t_i$ і $t_i \to u_i$.

Інтерпретація PLS-моделей є важливою для вивчення внутрішньої структури даних: визначення груп об'єктів, викидів (графік $t-u$), взаємозв'язків між змінними (графіки $w-q, p-w$). Методи багатовимірного градуювання також дозволяють замінити пряме вимірювання певної властивості вимірюванням іншої властивості, корельованої з першою [157].

За допомогою алгоритму PLS (як і за допомогою методу головних компонент) часто виявляють взаємозв'язки даних у великих масивах результатів вимірювань, отриманих за допомогою ІЧ-спектроскопії з Фур'є-перетворенням [66, 158, 159], спектроскопії в ближньому ІЧ-діапазоні [160, 161], Раманівської спектроскопії [162].

Як приклад розглянемо застосування алгоритму проекцій на латентні структури для ідентифікації яблучних соків [163]. Було створено лабораторні стандарти з різним вмістом яблучного соку: 2%, 4%, 6%, 8%, 10%, 16%, 20%, 25%, 50%, 70% і 100%. Набір стандартних зразків розділили на дві групи, для кожної групи визначили навчальну і тестову вибірки. Перша група включала зразки з вмістом яблучного соку від 2 до 20%; загальна кількість зразків 173, 134 з яких склали навчальну вибірку, а 39 зразків використали для перевірки алгоритму. Друга група включала зразки з вмістом яблучного соку від 25 до 100%; загальна кількість зразків 130, 86 з яких утворили навчальну вибірку, 44 сформували тестову вибірку. Для кожного стандартного зразка виміряли інфрачервоні спектри в середньому ІЧ-діапазоні.

До отриманого масиву результатів вимірювань приклали метод PLS. При цьому матриця *X* подає набір отриманих спектральних смуг, матриця *Y* – стандартні зразки з різним вмістом соку (матриця і є двійковим кодуванням цих зразків). У результаті роботи алгоритму виділили дві латентні компоненти.

Отриману модель застосували для класифікації 23 комерційних зразків соку: 2 зразки належали до першої групи об'єктів, 21 зразок –

до другої групи. В табл. 1.7 наведено два параметри, що характеризують роботу алгоритму PLS: ефективність (відношення кількості зразків, правильно віднесених до відповідного класу, до загальної кількості зразків, що належать цьому класу, %) і чутливість (відношення кількості зразків, правильно віднесених до відповідного класу, до загальної кількості віднесених до цього класу в результаті роботи алгоритму, %).

В результаті застосування алгоритму PLS до класифікації промислових зразків соку лише один зразок другої групи був ідентифікований неправильно.

Таблиця 1.7

Значення ефективності й чутливості методу PLS при його застосуванні для ідентифікації зразків із різним вмістом яблучного соку

Вміст яблучного соку у зразках, %	2	4	6	8	10	16	20	25	50	70	100
Ефективність, %	100	100	100	50	56	60	63	81	92	100	100
Чутливість, %	100	100	85	33	83	50	83	94	100	100	100

Потужним інструментом виявлення фальсифікації є метод дискримінантного аналізу за допомогою регресії на латентні структури (PLS Discriminant Analysis, PLS-DA). Його використовують для роботи з великими масивами даних, отриманих в результаті застосування, наприклад, мас-спектрометрії [164, 165], спектроскопії в ближньому ІЧ-діапазоні та видимої спектроскопії [128, 166], Раманівської спектроскопії [68].

Ідея цього підходу полягає в тому, що правила дискримінації для K класів задаються лінійними регресійними рівняннями виду

$$XB = D, \qquad (1.51)$$

де X – повна матриця всіх початкових даних $(I \times J)$, B – матриця невідомих коефіцієнтів $(J \times K)$, а D – це спеціальна матриця $(I \times K)$, яка складається з нулів і одиниць. При побудові матриці D одиниці ставлять лише в ті рядки (рядки відповідають досліджуваним об'єктам), які належать класу, що відповідає номеру стовпця. Регресійну задачу розв'язують методом PLS, що дозволяє надалі застосовувати побудовану регресію для передбачення приналежності нових об'єктів. Для цього будують прогноз відгуку нового об'єкта і порівнюють результат із нулем або одиницею [167].

На прикладі класифікації географічного походження зразків женьшеню [68] покажемо ефективність використання алгоритму PLS-DA. Було відібрано 50 китайських і корейських зразків женьшеню з різних регіонів цих країн. Всі зразки висушили та подрібнили до порошкоподібного стану і для кожного зразка виміряли ІЧ-спектри в ближній ІЧ-області та Раманівські спектри.

Весь отриманий масив даних випадковим чином розділили на навчальну вибірку для знаходження оптимальних параметрів класифікації і тестову вибірку для перевірки створеної моделі.

Перед застосуванням методу PLS-DA для того, щоб поліпшити розділення спектрів (більш точно визначити положення спектральних смуг, що перекриваються), для ІЧ-спектрів розрахували другі похідні. У табл. 1.8 представлено інформацію про використаний масив даних і результати роботи методу PLS-DA.

Таблиця 1.8

Відомості про спектральні характеристики зразків женьшеню і результати PLS-DA-аналізу

Параметр	ІЧ-спектроскопія	Раманівська спектроскопія
Кількість зразків у навчальній вибірці (корейські/китайські)	70 (35/35)	70 (35/35)
Кількість зразків у тестовій вибірці (корейські/китайські)	30 (15/15)	30 (15/15)
Діапазон вимірювань	400–2500 нм	250–1700 $см^{-1}$
Кількість PLS-компонент	3	6
Кількість неправильно класифікованих зразків у навчальній/тестовій вибірках	0/0	5/3
Ненадійність класифікації, %	0	9

***

Розширення кола завдань і розвиток методології якісного хімічного аналізу привели до нового розуміння якісного аналізу як сукупності експериментальних і розрахункових процедур, що забезпечують класифікацію об’єктів за їх хімічними, фізико-хімічними та іншими характеристиками. Відповідним чином зросла роль хемометричних методів, без яких обробка великих масивів багатовимірних даних є неможливою.

У зв'язку з цим, добиваючись високої надійності висновків якісного аналізу, хімік повинен уважно вибирати не лише експериментальні методи, але і хемометричні засоби обробки даних. Сьогодні для розв'язання задач якісного аналізу все ширше застосовують не лише традиційні, але і достатньо нові хемометричні методи.

Метод головних компонент слугує засобом попередньої обробки і вилучення інформації з багатовимірних масивів даних. Він дозволяє суттєво знижувати розмірність масиву даних при мінімальній втраті інформації.

Метод формального незалежного моделювання аналогій класів дозволяє відносити зразки до одного чи декількох класів або до жодного із змодельованих класів. Особливо успішно його використовують для обробки спектральних даних.

Популярним є метод дискримінантного аналізу, але він не завжди забезпечує високу надійність аналітичних висновків.

Використання в розв'язанні задач класифікації дерев класифікації і регресії приваблює тим, що цей метод дозволяє виявляти вплив кожної зі змінних на результат класифікації.

Метод опорних векторів, популярність якого зростає особливо швидко, є ефективним засобом розв'язання задач бінарної класифікації.

Корисним є метод проекції на латентні структури. Цей метод забезпечує виявлення внутрішньої структури даних і успішно застосовувався для обробки масивів спектральних даних.

Однією з головних проблем розв'язання задач класифікації є оцінка надійності/ненадійності логічних висновків, до яких приходить хімік за результатами якісного аналізу. До теперішнього часу панує статистичний (евристичний) підхід, що зводиться до оцінки ненадійності класифікації за часткою правильно/помилково класифікованих зразків у тестовій вибірці. Ненадійність оцінюють за допомогою або стандартних алгоритмів, або алгоритмів, параметри яких були відрегульовані при обробці навчальної вибірки. Тому, вибираючи метод обробки масивів експериментальних даних, перевагу, ймовірно, слід віддавати класифікації з навчанням. Підвищити упевненість у надійності класифікації дозволить і ширше використання перехресної оцінки достовірності (крос-валідації). Разом із тим, статистична («ad hoc») процедура, приваблюючи простотою і даючи розумну оцінку ненадійності розв'язання певної класифікаційної задачі, має і суттєвий недолік: важко передбачати, як буде

працювати той чи інший хемометричний алгоритм при переході до обробки нового масиву експериментальних даних з, можливо, іншою структурою і статистичними характеристиками. Нагальною є потреба виявлення серед хемометричних алгоритмів найбільш ефективних для розв'язання конкретних завдань обробки даних хімічного експерименту, розробка нових і така модернізація існуючих алгоритмів, що дозволять знизити вимоги до об'єму і точності початкових експериментальних даних (потрібні алгоритми, стійкі до наявності в даних шуму, пропусків, промахів).

Література до глави 1

1. Основы аналитической химии. Кн. 1. Общие вопросы. Методы разделения : учебник / Ю. А. Золотов, Е. Н. Дорохова, В. И. Фадеева, Т. А. Большова, Г. Д. Брыкина, А. В. Грамаш, И. Ф. Долманова, В. М. Иванов, О.А. Шпигун; под ред. Ю.А. Золотова. – 2-е изд., перераб. и доп. – М. : Высшая школа, 2002. – 351 с.
2. TrAC Trends in Analytical Chemistry. – 2005. – Vol. 24, No 6. – P. 461-556. Modern qualitative analysis / Editors M. Valcárcel and S. Cárdenas.
3. Cárdenas S. Analytical features in qualitative analysis / S. Cárdenas, M. Valcárcel // Trends Anal. Chem. – 2005. – Vol. 24, No 6. – P. 477-487.
4. Вершинин В. И. Компьютерная идентификация органических соединений : монография / В. И. Вершинин, Б. Г. Дерендяев, К. С. Лебедев. – М. : Наука, 2002. – 192 с.
5. Valcárcel M. Principles of qualitative analysis in the chromatographic context / M. Valcárcel, S. Cárdenas, B. M. Simonet, C. Carrillo-Carrion // J. Chromatogr. A. – 2007. – Vol. 1158, No 1-2. – P. 234-240.
6. Мильман Б. Л. Введение в химическую идентификацию / Б. Л. Мильман. – СПб. : ВВМ, 2008. – 180 с.
7. Vlasov Yu. Nonspecific sensor arrays («electronic tongue») for chemical analysis of liquids / Yu. Vlasov, A. Legin, A. Rudnitskaya, C. Di Natale, A. D'Amico // Pure Appl. Chem. – 2005. – Vol. 77, No 11. – P. 1965-1983.
8. Vershinin V. I. Chemometrics in the works of Russian analysts / V. I. Vershinin // J. Analyt. Chem. – 2011. – Vol. 66, No 11. – P. 1010-1019.
9. Qualitative analysis: a guide to best practice [Electronic Resource] / Editor W. A. Hardcastle. – Cambridge : Royal Society of Chemistry, 1998. – 24 p. – Way of access : http://www.rsc.org/ publishing/ebooks/ 1998/9780854044627.asp
10. Milman B. L. Identification of chemical compounds / B. L. Milman // Trends Anal. Chem. – 2005. – Vol. 24, No 6. – P. 493-508.
11. Родионова О. Е. Хемометрика: достижения и перспективы / О. Е. Родионова // Успехи химии. – 2006. – Т. 75, No 4. – С. 302-321.

12. Valcárcel M. Analytical chemistry at the interface between metrology and problem solving / M. Valcárcel // Trends Anal. Chem. – 2004. – Vol. 23, No 8. – P. 527-534.
13. Munoz-Olivas R. Screening analysis: an overview of methods applied to environmental, clinical and food analyses / R. Munoz-Olivas // Trends Anal. Chem. – 2004. – Vol. 23, No 3. – P. 203-216.
14. Mil'man B. L. Uncertainty of qualitative chemical analysis: general methodology and binary test methods / B. L. Mil'man, L. A. Konopel'ko // J. Anal. Chem. – 2004. – Vol. 59, No 12. – P. 1244-1258.
15. Холин Ю. В. Метрологические характеристики методик обнаружения с бинарным откликом / Ю. В. Холин, Н. А. Никитина, А. В. Пантелеймонов, Е. А. Решетняк, А. А. Бугаевский, Л. П. Логинова. – Харьков : Тимченко, 2008. – 128 с.
16. Rios A. Quality assurance of qualitative analysis in the framework of the European project «MEQUALAN» / A. Rios, D. Barcelo, L. Buydens, A. Ríos, S. Cárdenas, K. Heydorn, B. Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, R. W. Stephany, A. Townshend, A. Zschunkem M. Valcárcel // Accred. Qual. Assur. – 2003. – Vol. 8. – P. 68-77.
17. Valcárcel M. Metrology of qualitative chemical analysis [Electronic Resource] / M. Valcárcel, S. Cárdenas, D. Barceló, L. Buydens, K. Heydorn, B.Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, A. Rios, R. Stephany, A.Townshend, A. Zschunke. – Luxembourg: Office for Official Publications of the European Communities, 2002. – 166 p. – Way of access : http://bookshop. europa. eu/en/metrology-of-qualitative-chemical-analysis-pbKINA20605/
18. Encyclopedia of analytical science : ten-volume set / eds. P. Worsfold, A. Townshend, C. Poole. – 2nd edn. – Elsevier, 2005 – P. 405-411.
19. Вершинин В. И. Методология компьютерной идентификации веществ с применением информационно-поисковых систем / В. И. Вершинин // Журн. аналит. химии. – 2000. – Т. 55, No 5. – С. 468-476.
20. Соколова О. В. Достоверность компьютерной идентификации углеводородов при хроматографическом анализе бензинов / О. В. Соколова, Н. Б. Ильичева, В. И. Вершинин // Аналитика и контроль. – 2000. – Т. 4, No 4. – С. 363-369.
21. Matson J. L. True versus false positives and negatives on the modified checklist for autism in toddlers / J. L. Matson, A. M. Kozlowski, M. E. Fitzgerald, M. Sipes // Res. Autism Spect. Dis. – 2013. – Vol. 7. – P. 17-22.
22. Kumar R. False-negative and false-positive results in FDG-PET and PET/CT in breast cancer / R. Kumar, N. Rani, C. Patel, S. Basu, A. Alavi // PET Clinics. – 2009. – Vol. 4, No 3. – P. 289-298.
23. Balabin R. M. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques / R. M. Balabin, R. Z. Safieva, E. I. Lomakina // Anal. Chim. Acta. – 2010. – Vol. 671. – P. 27-35.

24. Balabin R. M. Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data / R. M. Balabin, R. Z. Safieva // Fuel. – 2008. – Vol. 87. – P. 2745-2752.
25. Методические указания по оценке подлинности и выявлению фальсификации молочной продукции. Методические указания МУ 4.1./4.2. 2484-09. – М. : Федеральный Центр гигиены и эпидемиологии Роспотребнадзора, 2009. – 26 с.
26. Kuchmenko T. A. Detection of synthetic components in food matrices using piezoelectric resonators / T. A. Kuchmenko, R. P. Lisitskaya, V. A. Golovanova, M. S. Arsenova // Anal. Chem. – 2009. – Vol. 64, No 4. – P. 338-345.
27. Ivleva N. P. Characterisation and discriminant of pollen by Raman microscopy / N. P. Ivleva, R. Niessner, U. Panne // Anal. Bioanal. Chem. – 2005. – Vol. 381. – P. 261-267.
28. Coetzee P. P. Classifying wine according to geographical origin via quadrupole-based ICP-mass spectrometry measurements of boron isotope ratios / P. P. Coetzee, F. Vanhaecke // Anal. Bioanal. Chem. – 2005. – Vol. 383. – P. 977-984.
29. Alonso-Salces R. M. Chemometrics characterisation of Basque and French ciders according to their polyphenolic profiles / R. M. Alonso-Salces, S. Guyot, C. Herrero, L. A. Berrueta, J. F. Drilleau, B. Gallo, F. Vicente // Anal. Bioanal. Chem. – 2004. – Vol. 379. – P. 464-475.
30. Bellorini S. Discriminating animal fats and their origins: assessing the potentials of Fourier transform infrared spectroscopy, gas chromatography, immunoassay and polymerase chain reaction techniques / S. Bellorini, S. Strathmann, V. Baeten, O. Fumiere, G. Berben, S. Tirendi, C. von Holst // Anal. Bioanal. Chem. – 2005. – Vol. 382. – P. 1073-1083.
31. Poulli K. I. Synchronous fluorescence spectroscopy for quantitative determination of virgin olive oil adulteration with sunflower oil / K. I. Poulli, G. A. Mousdis, C. A. Georgiou // Anal. Bioanal. Chem. – 2006. – Vol. 386. – P. 1571-575.
32. Baeten V. Detection of banned meat and bone meal in feedstuffs by near-infrared microscopic analysis of the dense sediment fraction / V. Baeten, C. von Holst, A. Garrido, J. Vancutsem, A. M. Renier, P. Dardenne // Anal. Bioanal. Chem. – 2005. – Vol. 382. – P. 149-157.
33. Husted S. Elemental fingerprints analysis of barley (Hordeum vulgare) using inductively coupled plasma mass spectrometry, isotope-ratio mass spectrometry, and multivariate statistics / S. Husted, B. F. Mikkelsen, J. Jensen, N.E. Nielsen // Anal. Bioanal. Chem. – 2004. – Vol. 378. – P. 171-182.
34. Nalda N. M. J. Classifying honeys from the Soria Province of Spain via multivariate analysis / N. M. J. Nalda, B. J. L. Yagüe, D. J. C. Calva, M. M. T. Gómez // Anal. Bioanal. Chem. – 2005. – Vol. 382. – P. 311-319.

35. Jurado J. M. Differentiation of certified brands of origins of Spanish white wines by HS-SPME-GC and chemometrics / J. M. Jurado, O. Ballesteros, A. Alcázar, F. Pablos, M. J. Martín, J. L. Vilchez, A. Navalón // Anal. Bioanal. Chem. – 2008. – Vol. 390. – P. 961-970.
36. Cuny M. Fruit juice authentication by ^{1}H NMR spectroscopy in combination with different chemometrics tools / M. Cuny, E. Vigneau, G. Le Gall, I. Colquhoun, M. Less, D. N. Rutledge // Anal. Bioanal. Chem. – 2008. – Vol. 390. – P. 419-427.
37. Sumar S. Adulteration of foods – past and present / S. Sumar, H. Izmail // Nutrition & Food Sci. – 1995. – Vol. 95, No 4. – P. 11-15.
38. Arvanitoyannis I. S. A review on tomato authenticity: quality control methods in conjunction with multivariate analysis / I. S. Arvanitoyannis, O. B. Vaitsi // Crit. Rev. in Food Sci. and Nutrition. – 2007. – Vol. 47. – P. 675-699.
39. Рыков С. В. Принципы контроля экологической и санитарно-гигиенической безопасности плодовых и ягодных соков / С. В. Рыков, И. С. Корягина, Е. Д. Скаковский // Экологические нормы. Правила. Информация. – 2009. – No 4. – С. 34-39.
40. Knödler M. A novel approach to authenticity control of whole grain durum wheat (Triticum durum Desf.) flour and pasta, based on analysis of alkylresorcinol composition / M. Knödler, M. Most, A. Schieber, R. Carle // Food Chem. – 2010. – Vol. 118, No 1. – P. 177-181.
41. Socaciu C. Complementary advanced techniques applied for plant and food authentication / C. Socaciu, F. Ranga, F. Fetea, L. Leopold, F. Dulf, R. Parlog // Czech J. of Food Sci. – 2009. – Vol. 27. – P. 70-75.
42. Carter R. M. Digital imaging based classification and authentication of granular food products / R. M. Carter, Y. Yan, K. Tomlins // Management Sci. and Technol. – 2006. – Vol. 17, No 2. – P. 235-240.
43. Díaz-Maroto M. C. Authenticity evaluation of different mints based on their volatile composition and olfactory profile / M. C. Díaz-Maroto, N. Castillo, L. Castro-Vázquez, C. De Torres, M. S. Pérez-Coello // J. of Essential Oil-Bearing Plants. – 2008. – Vol. 11, No 1. – P. 1-16.
44. Winterhalter P. Authentication of food and wine / P. Winterhalter // ACS Symposium Series. – 2007. – Vol. 952. – P. 2-12.
45. Martin M. Frauds in food products: A challenge for analytical chemistry / M. Martin, G. Martin // Actualite Chimique. – 2000. – Vol. 11. – P. 18-20.
46. Destaillats F. Authenticity of milk fat by fast analysis of triacylglycerols. Application to the detection of partially hydrogenated vegetable oils / F. Destaillats, M. de Wispelaere, F. Joffre, P.-A. Golay, B. Hug, F. Giuffrida, L. Fauconnot, F. Dionisi // J. of Chromatography A. – 2006. – Vol. 1131, No 1-2. – P. 227-234.
47. Fügel R. Quality and authenticity control of fruit purées, fruit preparations and jams – A review / R. Fügel, R. Carle, A. Schieber // Trends in Food Sci. and Technol. – 2005. – Vol. 16, No 10. – P. 433-441.

48. Jee M. Oils and fats – authenticity and adulteration / M. Jee // Food Sci. and Technol. – 2004. – Vol. 18, No 1. – P. 28-29.
49. Zhang Y. International multidimensional authenticity specification (IMAS) algorithm for detection of commercial pomegranate juice adulteration / Y. Zhang, D. Krueger, R. Durst, R. Lee, D. Wang, N. Seeram, D. Heber // J. of Agricult. and Food Chem. – 2009. – Vol. 57, No 6. – P. 2550-2557.
50. Cotte J. F. Application of carbohydrate analysis to verify honey authenticity / J. F. Cotte, H. Casabianca, S. Chardon, J. Lheritier, M. F. Grenier-Loustalot // J. of Chromatography A. – 2003. – Vol. 1021, No 1-2. – P. 145-155.
51. Lohumi S. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration / S. Lohumi, S. Lee, H. Lee, B.-K. Cho // Trends in Food Sci. and Technol. – 2015. – Vol. 46, No 1. – P. 85-98.
52. Kelly J. F. D. Initial study of honey adulteration by sugar solutions using midinfrared (MIR) spectroscopy and chemometrics / J. F. D. Kelly, G. Downey, V. Fouratier // J. of Agricult. and Food Chem. – 2004. – Vol. 52, No 1. – P. 33-39.
53. Marini F. Supervised pattern recognition to discriminate the geographical origin of rice bran oils: a first study / F. Marini, F. Balestrieri, R. Bucci, A.L. Magri, D. Marini // Microchem. J. – 2003. – Vol. 74. – P. 239-248.
54. Laursen K. H. Multi-elemental fingerprinting of plant tissue by semi-quantitative ICP-MS and Chemometrics / K. H. Laursen, T. H. Hansen, D. P. Persson, J. K. Schjoerring, S. Husted // J. of Anal. At. Spectrom. – 2009. – Vol. 24, No 9. – P. 1198-1207.
55. Arvanitoyannis I. S. Application of quality control methods for assessing wine authenticity: use of multivariate analysis (chemometrics) / I. S. Arvanitoyannis, M. N. Katsota, E. P. Psarra, E. H. Soufleros, S. Kallithraka // Trends in Food Sci. and Technol. – Vol. 10, No 10. – P. 321-336.
56. Arvanitoyannis I. S. Implementation of quality control methods (physicochemical, microbiological and sensory) in conjunction with multivariate analysis towards fish authenticity / I. S. Arvanitoyannis, E. V. Tsitsika, P. Panagiotaki // Int. J. of Food Sci. and Technol. – 2005. – Vol. 40, No 3. – P. 237-263.
57. Marini F. Artificial neural networks in food analysis: trends and perspectives. A review / F. Marini // Anal. Chim. Acta. – 2009. – Vol. 635. – P. 121-131.
58. Marini F. Authentication of Italian CDO wines by class-modeling techniques / F. Marini, R. Bucci, A. D. Magri, A. L. Magri // Chemometr. Intell. Lab. – 2006. – Vol. 84. – P. 164-171.
59. De León-Rodríguez A. Characterization of volatile compounds from ethnic Agave alcoholic beverages by gas chromatography-mass spectrometry / A. De León-Rodríguez, P. Escalante-Minakata, M. I. Jiménez-García,

L. G. Ordoñez-Acevedo, J. L. F. Flores, A. P. B. De La Rosa // Food Technol. and Biotechnol. – 2008. – Vol. 46, No 4. – P. 448-455.

60. Cabañero A. I. Liquid chromatography coupled to isotope ratio mass spectrometry: a new perspective on honey adulteration detection / A. I. Cabañero, J. L. Recio, M. Rupérez // J. of Agricult. and Food Chem. – 2006. – Vol. 54, No 26. – P. 9719-9727.
61. Calderone G. Helping to authenticate sparkling drinks with $^{13}C/^{12}C$ of CO_2 by gas chromatography-isotope ratio mass spectrometry / G. Calderone, C. Guillou, F. Reniero, N. Naulet // Food Research Int. – 2007. – Vol. 40, No 3. – P. 324-331.
62. Gómez-Ariza J. L. Characterization and analysis of amino acids in orange juice by HPLC-MS/MS for authenticity assessment / J. L. Gómez-Ariza, M. J. Villegas-Portero, V. Bernal-Daza // Anal. Chim. Acta. – 2005. – Vol. 540, No 1. – P. 221-230.
63. Nhu-Trang T.-T. Authenticity control of essential oils containing citronellal and citral by chiral and stable-isotope gas-chromatographic analysis / T.-T. Nhu-Trang, H. Casabianca, M.-F. Grenier-Loustalot // Anal. and Bioanal. Chem. – 2006. – Vol. 386, No 7-8. – P. 2141-2152.
64. Alonso-Salces R.M. On-line characterisation of apple polyphenols by liquid chromatography coupled with mass spectrometry and ultraviolet absorbance detection / R. M. Alonso-Salces, K. Ndjoko, E. F. Queiroz, J. R. Ioset, K. Hostettmann, L. A. Berrueta, B. Gallo, F. Vicente // J. of Chromatography A. – 2004. – Vol. 1046, No 1-2. – P. 89-100.
65. Liang L. Discrimination of variety and authenticity for rice based on visual/near infrared reflection spectra / L. Liang, Z.-X. Liu, M.-H. Yang, Y.-X. Zhang, C.-H. Wang // J. of Infrared and Millimeter Waves. – 2009. – Vol. 28, No 5. – P. 353-356.
66. Chen L.-Z. Determination of adulteration in honey using near-infrared spectroscopy / L.-Z. Chen, J. Zhao, Z.-H. Ye, Y.-P. Zhong // Spectroscopy and Spectral Analysis. – 2008. – Vol. 28, No 11. – P. 2565-2568.
67. Nicolaou N. Rapid and quantitative detection of the microbial spoilage in milk using Fourier transform infrared spectroscopy and chemometrics / N. Nicolaou, R. Goodacre // Analyst. – 2008. – Vol. 133. – P. 1424-1431.
68. Feng X. Preliminary study on classification of rice and detection of paraffin in the adultered samples by Raman spectroscopy combined with multivariate analysis / X. Feng, O. Zhang, P. Cong, Z. Zhu // Talanta. – 2013. – Vol. 115. – P. 548-555.
69. Boffo E. F. Classification of Brazilian vinegars according to their ^{1}H NMR spectra by pattern recognition analysis / E. F. Boffo, L. A. Tavares, M. M. C. Ferreira, A. G. Ferreira // LWT – Food Sci. and Technol. – 2009. – Vol. 42, No 9. – P. 1455-1460.
70. Viggiani L. Characterization of wines by nuclear magnetic resonance: a work study on wines from the Basilicata region in Italy / L. Viggiani,

M. A. C. Morelli // J. of Agricult. and Food Chem. – 2008. – Vol. 56, No 18. – P. 8273-8279.
71. Shintu L. Pre-selection of potential molecular markers for the geographic origin of dried beef by HR-MAS NMR spectroscopy / L. Shintu, S. Caldarelli, B. M. Franke // Meat Sci. – 2007. – Vol. 76, No 4. – P. 700-707.
72. Cabañero A. I. Isotope ratio mass spectrometry coupled to liquid and gas chromatography for wine ethanol characterization / A. I. Cabañero, J. L. Recio, M. Rupérez // Rapid Commun. in Mass Spectrom. – 2008. – Vol. 22, No 20. – P. 3111-3118.
73. Greule M. Feed additives: authenticity assessment using multicomponent-/ multielement-isotope ratio mass spectrometry / M. Greule, C. Hänsel, U. Bauermann, A. Mosandl // Eur. Food Res. and Technol. – 2008. – Vol. 227, No 3. – P. 767-776.
74. Richling E. Flavor authenticity studies by isotope ratio mass spectrometry: perspectives and limits / E. Richling, M. Appel, F. Heckel, K. Kahle, M. Kraus, C. Preston, W. Hummer, P. Schreier // ACS Symposium Series. – 2007. – Vol. 952. – P. 75-86.
75. Zeynali F. Determination of copper, zinc and iron levels in edible muscle of three commercial fish species from iranian coastal waters of the Caspian sea / F. Zeynali, H. Tajik, S. Asri-Rezaie, S. Meshkini, A. A. Fallah, M. Rahnama // J. of Animal and Veterinary Advances. – 2009. – Vol. 8, No 7. – P. 1285-1288.
76. Khan S. Separation and preconcentration of trace amounts of aluminum ions in surface water samples using different analytical techniques / S. Khan, T. G. Kazi, J. A. Baig, N. F. Kolachi, H. I. Afridi, A. Q. Shah, G. A. Kandhro, S. Kumar // Talanta. – 2009. – Vol. 80, No 1. – P. 158-162.
77. Rizzon L. A. Analytical characteristics of Merlot wines from the Serra Gaúcha region / L. A. Rizzon, A. Miele // Ciencia Rural. – 2009. – Vol. 39, No 6. – P. 1913-1916.
78. Scott S. M. Total luminescence spectroscopy with pattern recognition for classification of edible oils / S. M. Scott, D. James, Z. Ali, W. T. O'Hare, F. J. Rowell // Analyst. – 2003. – Vol. 128. – P. 966-973.
79. Cuadrado U. M. Study of spectral analytical data using fingerprints and scaled similarity measurements / U. M. Cuadrado, L. M. D. de Castro, M. A. Gomez-Nieto // Anal. Bioanal. Chem. – 2005. – Vol. 381. – P. 953-963.
80. Woodcock T. Near infrared spectral fingerprinting for confirmation of claimed PDO provenance of honey / T. Woodcock, G. Downey, C. P. O'Donnell // Food Chem. – 2009. – Vol. 114, No 2. – P. 742-746.
81. Vlasov Y. Non-selective chemical sensors in analytical chemistry: from «electronic nose» to «electronic tongue» / Y. Vlasov, A. Legin // Fresenius J. Anal. Chem. – 1998. – Vol. 361. – P. 255-260.

82. Mielle P. From human to artificial mouth, from basics to results / P. Mielle, A. Tarrega, P. Gorria, J. J. Liodenot, J. Liaboeuf, J.-L. Andrejewski, C. Salles // AIP Conference Proceedings. – 2009. – Vol. 1137. – P. 144-147.
83. Leake L. L. Electronic noses and tongues / L. L. Leake // Food Technol. – 2006. – Vol. 60, No 6. – P. 96-102.
84. Lvova L. All-solid-state electronic tongue and its application for beverage analysis / L. Lvova, S. S. Kim, A. Legin, Y. Vlasov, J. S. Yang, G. S. Cha, H. Nam // Anal. Chim. Acta. – 2002. – Vol. 468. – P. 303-314.
85. Tiwari K. Identification of monofloral honey using voltammetic electronic tongue / K. Tiwari, B. Tudu, R. Bandyopadhyay, A. Chatterjee // J. Food Engineering. – 2013. – Vol. 117. – P. 205-210.
86. Winquist F. Electronic tongues / F. Winquist, C. Krantz-Rülcker, I. Lundström // MRS Bulletin. – 2004. – Vol. 29, No 10. – P. 726-731.
87. Ciosek P. Classification of beverages using a reduced sensor array / P. Ciosek, Z. Brzózka, W. Wróblewski // Sensors and Actuators B: Chemical. – 2004. – Vol. 103, No 1-2. – P. 76-83.
88. Buratti S. Characterization and classification of Italian Barbera wines by using an electronic nose and an amperometric electronic tongue / S. Buratti, S. Benedetti, M. Scampicchio, E. C. Pangerod // Anal. Chim. Acta. – 2004. – Vol. 525, No 1. – P. 133-139.
89. Tan T. Electronic tongue: a new dimension in sensory analysis: instrument complements use of the electronic nose for sensory analysis of food flavor by measuring nonvolatile flavor components / T. Tan, V. Schmitt, S. Isz // Food Technol. – 2001. – Vol. 55, No 10. – P. 44-50.
90. Bleibaum R. N. Comparison of sensory and consumer results with electronic nose and tongue sensors for apple juices / R. N. Bleibaum, H. Stone, T. Tan, S. Labreche, E. Saint-Martin, S. Isz // Food Quality and Preference. – 2002. – Vol. 13, No 6. – P. 409-422.
91. Родионова О. Е. Хемометрика в аналитической химии / О. Е. Родионова, А. Л. Померанцев, 2006. – 61 с. [Электронный ресурс]. – Режим доступа : http://www.chemometrics.ru/materials/articles/chemometrics_review.pdf
92. Дронов С. В. Многомерный статистический анализ : учебное пособие / С. В. Дронов. – Барнаул : Изд-во Алт. гос. ун-та, 2003. – 213 с.
93. Rodionova O. Ye. Quantitative risk assessment in classification of drugs with identical API content / O. Ye. Rodionova, K. S. Balyklova, A. V. Titova, A. L. Pomerantsev // J. Pharmaceutical and Biomedical Analysis. – 2014. – Vol. 98. – P. 186-192.
94. Coussement A. Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches / A. Coussement, B. J. Isaac, O. Sicquel, A. Parente // Combustion and Flame. – 2016. – Vol. 168. – P. 83-97.
95. Hopke P. K. The evolution of chemometrics / P. K. Hopke // Anal. Chim. Acta. – 2003. – Vol. 500. – P. 365-377.

96. Legin A. Electronic tongue for pharmaceutical analytics: quantification of tastes and masking effects / A. Legin, A. Rudnitskaya, D. Clapham, B. Seleznev, Y. Vlasov // Anal. Bioanal. Chem. – 2004. – Vol. 380. – P. 36-45.
97. Hruškar M. Evaluation of milk and dairy products by electronic tongue / M. Hruškar, N. Major, M. Krpan, I. P. Krbavčić, G. Šarić, K. Marković, N. Vahčić // Mljekarstvo. – 2009. – Vol. 59, No 3. – P. 193-200.
98. Oliveri P. Development of a voltammetric electronic tongue for discrimination of edible oils / P. Oliveri, M. A. Baldo, S. Daniele, M. Forina // Anal. Bioanal. Chem. – 2009. – Vol. 395. – P. 1135-1143.
99. Cotte J. F. Chromatographic analysis of sugars applied to the characterization of monofloral honey / J. F. Cotte, H. Casabianca, S. Chardon, J. Lheritier, M. F. Grenier-Loustalot // Anal. Bioanal. Chem. – 2004. – Vol. 380. – P. 698-705.
100. Li H. A chemometrics approach for distinguishing between beers using near infrared spectroscopy / H. Li, Y. Takahashi, M. Kumagai, K. Fujiwara, R. Kikuchi, N. Yoshimura, T. Amano, N. Ogawa // J. of Near Infrared Spectrosc. – 2009. – Vol. 17, No 2. – P. 69-76.
101. Shin Y.-S. Fingerprinting analysis of fresh ginseng roots of different ages using ^{1}H-NMR spectroscopy and principal components analysis / Y. S. Shin, K.-H. Bang, D.-S. In, O.-T. Kim, D.-Y. Hyun, I.-O. Ahn, C. K. Bon, H.-K. Choi // Archives of Pharm. Res. – 2007. – Vol. 30, No 12. – P. 1625-1628.
102. Шараф М. А. Хемометрика / М. А. Шараф, Д. Л. Иллмэн, Б. Р. Ковальски; пер. с англ. – Л. : Химия, 1989. – 272 с.
103. Cozzolino D. Combining visible and near-infrared spectroscopy with Chemometrics to trace muscles from an autochthonous breed of pig produced in Uruguay: a feasibility study / D. Cozzolino, A. Vadell, F. Ballesteros, G. Galietta, N. Barlocco // Anal. Bioanal. Chem. – 2006. – Vol. 385. – P. 931-936.
104. Родионова О. Е. Задачи классификации и дискриминации. Презентация лекции / Родионова О. Е. – Омск : Омский гос. ун-т им. Ф. М. Достоевского, 2007. [Электронный ресурс]. – Режим доступа : http:// www.chemometrics.ru/materials/presentations/omsk-2007/Classifiation.ppt
105. Durante C. A classification tool for N-way array based on SIMCA methodology / C. Durante, R. Bro, M. Cocchi // Chemometr. and Intell. Lab. Systems. – 2011. – Vol. 106. – P. 73-85.
106. Родионова О. Е. Хемометрический подход к исследованию больших массивов химических данных / О. Е. Родионова // Рос. хим. журн. (Журн. Рос. хим. об-ва им. Д. И. Менделеева). – 2006. – Т. L, No 2. – С. 128-144.
107. Stanimirova I. Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques /

I. Stanimirova, B. Ustun, T. Cajka, K. Riddelova, J. Hajslova, L. M. C. Buydens, B. Walczak // Food Chem. – 2010. – Vol. 118, No 1. – P. 171-176.

108. Shiroma C. Application of NIR and MIR spectroscopy in quality control of potato chips / C. Shiroma, L. Rodriguez-Saona // J. of Food Composition and Analysis. – 2009. – Vol. 22, No 6. – P. 596-605.
109. Jaganathan J. Geographic origin of wine via trace and ultra-trace elemental analysis using inductively coupled plasma mass spectrometry and chemometrics / J. Jaganathan, Md. A. Mabud, S. Dugar // ACS Symposium Series. – 2007. – Vol. 952. – P. 200-206.
110. Xie L.-J. Use of near-infrared spectroscopy and least-squares support vector machine to determine quality change of tomato juice / L.-J. Xie, Y. B. Ying // J. of Zhejiang University: Science B. – 2009. – Vol. 10, No 6. – P. 465-471.
111. Woodcock T. Geographical classification of honey samples by near-infrared spectroscopy: a feasibility study / T. Woodcock, G. Downey, J. D. Kelly, C. O'Donnell // J. of Agricult. and Food Chem. – 2007. – Vol. 55, No 22. – P. 9128-9134.
112. Chen Q. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration / Q. Chen, J. Zhao, H. Zhang, X. Wang // Anal. Chim. Acta. – 2006. – Vol. 572, No 1. – P. 77-84.
113. Tian S.-Y. Discrimination of red wine age using voltammetric electronic tongue based on multifrequency large-amplitude voltammetry and pattern recognition method / S.-Y. Tian, S.-P. Deng, C.-H. Ding, C.-L. Yin, H. Li // Sensors and Materials. – 2007. – Vol. 19, No. 5. – P. 287-298.
114. He J. Midinfrared spectroscopy for juice authentication – rapid differentiation of commercial juices / J. He, L. E. Rodriguez-Saona, M. M. Giusti // J. of Agricult. and Food Chem. – 2007. – Vol. 55, No 11. – P. 4443-4452.
115. Chen Z.-P. Fuzzy linear discriminant analysis for chemical data sets / Z.-P. Chen, J.-H. Jiang, Y. Li, Yi-Z. Liang, Yu. Ru-Qin // Chemometr. and Intell. Lab. Systems. – 1999. – Vol. 45. – P. 295-302.
116. Ким Дж. О. Факторный, дискриминантный и кластерный анализ / Дж. О. Ким, Ч. У. Мьюллер, У. Р. Клекка, М. С. Олдендерфер, Р. К. Блэшфилд; пер. с англ. под ред. И.С. Енюкова. – М. : Финансы и статистика, 1989. – 215 с.
117. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с.
118. Кендалл М. Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт, Э. Л. Пресман, В. И. Ротарь, А. Н. Колмогоров, Ю. В. Прохоров. – М. : Наука, 1976. – 736 с.

119. Дубров А. М. Многомерные статистические методы / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 2000. – 350 с.
120. Афифи А. Статистический анализ: Подход с использованием ЭВМ / А. Афифи, С. Эйзен, И. С. Енюков, И. Д. Новиков, Г. П. Башарин; пер. с англ. – М. : Мир, 1982. – 488 с.
121. Михальчук В. М. Линейный регрессионный анализ результатов химического эксперимента в системе Statistica : учебно-методическое пособие / В. М. Михальчук, А. В. Михальчук. – Донецк : ДонНУ, 2002. – 66 с. [Электронный ресурс]. – Режим доступа : http://www.donnu.edu.ua/chem/student/methodic/line_analysis.pdf
122. Налимов В. В. Применение математической статистики при анализе вещества / В. В. Налимов. – М. : Гос. изд-во физ.-мат. л-ры, 1960. – 430 с.
123. Paradkar M. M. Discrimination and classification of beet and cane sugars and their inverts in maple syrup by FT-Raman spectroscopy / M. M. Paradkar, J. Irudayaraj, S. Sakhamuri // Applied Eng. in Agricult. – 2002. – Vol. 18, No 3. – P. 379-383.
124. Suchánek M. Qualitative analysis of green coffee by infrared spectrometry / M. Suchánek, H. Filipová, K. Volka, I. Delgadillo, A. N. Davies // Fresenius J. of Anal. Chem. – 1996. – Vol. 354, No 3. – P. 327-332.
125. Cozzolino D. Chemometrics and visible-near infrared spectroscopic monitoring of red wine fermentation in a pilot scale / D. Cozzolino, M. Parker, R. G. Dambergs, M. Herderich, M. Gishen // Biotechnol. and Bioeng. – 2006. – Vol. 95, No 6. – P. 1101-1107.
126. Granitto P. M. Modern data mining tools in descriptive sensory analysis: a case study a Random forest approach / P. M. Granitto, F. Gasperi, F. Biasioli, E. Trainotti, C. Furlanello // Food Quality and Preference. – 2007. – Vol. 78. – P. 681-689.
127. Timm H. An extension to possibilistic fuzzy cluster analusis / H. Timm, C. Borgelt, C. Döring, R. Kruse // Fuzzy Sets and Syst. – 2004. – Vol. 147. – P. 3-16.
128. Tsai C.-An. Multi-class clustering and prediction in the analysis of microarray data / C.-An. Tsai, Te-C. Lee, I-C. Ho, U.-C. Yang, C.-H. Chen, J. J. Chen // Mathematical Biosci. – 2005. – Vol. 193. – P. 79-100.
129. Leski J. Towards a robust clustering / J. Leski // Fuzzy Sets and Syst. – 2003. – Vol. 137. – P. 215-233.
130. Dumitrescu D. Fuzzy sets and their application to clustering and training / D. Dumitrescu, B. Lazzerini, L. C. Jain. – Boca Raton, London, New York, Washington : CRC Press, 2000. – 593 с.
131. Zahid N. Fuzzy clustering on k-nearest-neighbours rule / N. Zahid, O. Abouela, M. Limouri, A. Essaid // Fuzzy Sets and Syst. – 2001. – Vol. 120. – P. 239-247.

132. Wu K.-L. Alternative C-means clustering algorithms / K.-L. Wu, M.-S. Yang // Pattern Recognition. – 2002. – Vol. 35. – P. 2267-2278.
133. Shahin M. A. Fuzzy logic model for predicting peanut maturity / M. A. Shahin, B. P. Verma, E. W. Tollner // Transactions of the American Soc. of Agricult. Engineers. – 2000. – Vol. 43, No 2. – P. 483-490.
134. Андреев И. М. Описание алгоритма CART / И. М. Андреев // Методы. Алгоритмы. Программы. – 2004. – Т. 3–4, No 7-8. – С. 48-53.
135. Шитиков В. К. Количественная гидроэкология: методы системной идентификации [Электронный ресурс] / В. К. Шитиков, Г. С. Розенберг, Т. Д. Зинченко. – Тольятти : Институт экологии Волжского бассейна РАН, 2003. – 463 с. – Режим доступа : http://www.ievbran.ru/kiril/Library/Book1/content0/content0.htm
136. Zhang M. H. Application of boosting to classification problems in chemometrics / M. H. Zhang, Q. S. Xu, F. Daeyaert, P. J. Lewi, D. L. Massart // Anal. Chim. Acta. – 2005. – Vol. 544. – P. 167-176.
137. Dobra A. V. Classification and regression tree construction. PhD Thesis [Electronic Resource] / A. V. Dobra – Cornell University, 2003. – 186 p. – Way of access : http://www.cise.ufl.edu/~adobra/old/papers/phd-thesis.pdf
138. Шампандер А. Дж. Искусственный интеллект в компьютерных играх: как обучить виртуальные персонажи реагировать на внешние воздействия. Гл. 26. Деревья классификации и регрессии [Электронный ресурс] / А. Дж. Шампандер. – М. : Вильямс, 2007. – С. 385-401.– Режим доступа : http://www.williamspublishing.com/PDF/ 978-5-8459-1170-4/part.pdf
139. Khoshgoftaar T. M. Using classification trees for software quality models: Lessons learned / T. M. Khoshgoftaar, E. B. Allen, A. Naik, W. D. Jones, J. P. Hudepohl // Int. J. of Software Eng. and Knowledge Eng. – 1999. – Vol. 9, No 2. – P. 217-231.
140. Mutihac L. Mining in chemometrics / L. Mutihac, R. Mutihac // Anal. Chim. Acta. – 2008. – Vol. 612. – P. 1-18.
141. Nobel W. S. What is a support vector machine? / W. S. Nobel // Nature Biotechnol. – 2006. – Vol. 24, No 12. – P. 1565-1567.
142. Brereton R. G. Support Vector Machines for classification and regression / R. G. Brereton, G. R. Lloyd // Analyst. – 2010. – Vol. 135. – P. 230-267.
143. Esme E. Fuzzy c-means based support vector machines classifier for perfume recognition / E. Esme, B. Karlik // Applied Soft Computing. – 2016. – Vol. 46. – P. 452-458.
144. Madeo R. C. B. Gesture phase segmentation using support vector machines / R. C. B. Madeo, S. M. Peres, C. A. de Moraes Lima // Expert Systems with Applications. – 2016. – Vol. 56. – P. 100-115.
145. Лившиц Ю. Метод опорных векторов: лекция № 7 курса «Алгоритмы для Интернета» [Электронный ресурс] / Ю. Лившиц, 2006. – 9 с. – Режим доступа : http://yury.name/internet/07ianote.pdf

146. Lin Y. On the support vector machines. Technical Report № 1029 [Electronic Resource] / Y. Lin. – Madison : University of Wisconsin, Department of Statistics, 2000. – 25 p. – Way of access : http://www.svms.org/ tutorials/Lin2000.pdf
147. Воронцов К. В. Лекции по методу опорных векторов / К. В. Воронцов, 2007. – 18 с. [Электронный ресурс]. – Режим доступа : http://www.ccas.ru/voron/download/SVM.pdf
148. Браславский П. Система автоматического реферирования новостных сообщений на основе машинного обучения [Электронный ресурс] / П. Браславский, В. Густелев // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды 9-й Всеросс. конф. RCDL'2007. – Переславль-Залесский: Изд-во «Университет города Переславля», 2007. – С. 142–147.– Режим доступа : http://www.kansas.ru/pb/paper/rcdl2007.pdf
149. Erästö P. Support vector machines – backgrounds and practice. Dissertation for the Degree of Licentiate of Philosophy [Electronic Resource] / P. Erästö. – Helsinki: Rolf Nevanlinna Institute, 2001. – 78 p. – Way of access : http://www.svms.org/ tutorials/Erasto2001.pdf
150. Burges C. J. C. A tutorial on support vector machines for pattern recognition / C. J. C. Burges // Data Mining and Knowledge Discovery. – 1998. – Vol. 2, No 2. – P. 121-167.
151. Burbidge R. Support vector machines: some perspectives from probability and statistics [Electronic Resource] / R. Burbidge // Imperial College Statistics Seminars, 2001. – 53 p. – Way of access : http://www.svms.org/ tutorials/Burbidge2001.pdf
152. Setiawan M. A. Partial correlation metric based classifier for food product characterization / M. A. Setiawan, R. K. Rao, S. Lakshminarayanan // J. of Food Eng. – 2009. – Vol. 90. – P. 146-152.
153. Эсбенсен К. Анализ многомерных данных. Избранные главы: пер. с англ. С.В. Кучерявского под ред. О.Е. Родионовой / К. Эсбенсен. – Черноголовка : Институт проблем химической физики РАН, 2005. – 160 с.
154. Родионова О. Проекционные методы в системе Excel : учебное пособие [Электронный ресурс] / О. Родионова, А. Померанцев // Российское хемометрическое общество.– Режим доступа : http://www.chemometrics.ru/ materials/textbooks/projection.htm#Ch1.3
155. Lakshminarayanan S. Modelling and control of multivariable process: the dynamic projection to latent structures approach / S. Lakshminarayanan, L. Sirish, K. Nandakumar // AIChe J. – 1997. – Vol. 43. – P. 2307-2323.
156. Abdi H. Partial least square regression, projection on latent structure regression (PLS-Regression) [Electronic Resource] / H. Abdi // Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2. – Wiley, 2010. – P. 97-106. – Way of access : http://www.utdallas.edu/~herve/abdi-wireCS-PLS2010.pdf

157. Unay D. Stem and calyx recognition on Jonagold apples by pattern recognition / D. Unay, B. Gosselin // J. of Food Eng. – 2007. – Vol. 78, No 2. – P. 597-605.
158. Nunes A. Estimation of olive oil acidity using FT-IR and partial least squares regression / A. Nunes, J. Martins, A. S. Barros, A. C. Galvis-Sánchez, I. Delgadillo // Sensing and Instrumentation for Food Quality and Safety. – 2009. – Vol. 3, No 3. – P. 187-191.
159. Lachenmeier D. W. Rapid quality control of spirit drinks and beer using multivariate data analysis of Fourier transform infrared spectra / D. W. Lachenmeier // Food Chem. – 2007. – Vol. 101, No 2. – P. 825-832.
160. Lin P. Fast discrimination of varieties of sugar based on spectroscopy technology / P. Lin, Y. M. Chen, Y. He // Spectrosc. and Spectral Analysis. – 2009. – Vol. 29, No 2. – P. 382-385.
161. Armenta S. Determination of edible oil parameters by near infrared spectrometry / S. Armenta, S. Garrigues, M. de la Guardia // Anal. Chim. Acta. – 2007. – Vol. 596, No 2. – P. 330-337.
162. Mizrach A. Yeast detection in apple juice using Raman spectroscopy and chemometric methods / A. Mizrach, Z. Schmilovitch, R. Korotic, J. Irudayaraj, R. Shapira // Trans. of the ASABE. – 2007. – Vol. 50, No 6. – P. 2143-2149.
163. Andrade J. M. Classification of commercial apple beverages using a minimum set of mid-IR wavenumbers selected by Procrustes rotation / J. M. Andrade, M. P. Gomez-Carracedo, E. Fernandez, A. Elbergali, M. Kubista, D. Prada // Analyst. – 2003. – Vol. 128. – P. 1193-1199.
164. Ruth S. Geographical origin, cultivar and harvesting year verification of European and non-European olive oils using proton transfer reaction mass spectrometry with multivariate data analysis / S. Ruth, J. L. Kiers, W. Akkermans, R. Perez, A. H. Koot, E. Perri, M. Pellegrino, R. J. M. Moreno, C. Guillou, A. Rossignol-Castera // Proc. of the 5th Intern. Technical Symp. on Food Processing, Monitoring Technol. in Bioprocesses and Food Quality Management. – Potsdam, Germany. – 2009. – P. 791–797.
165. Ruth S. Butter and butter oil classification by PTR-MS / S. Ruth, A. Koot, W. Akkermans, N. Araghipour, M. Rozijn, M. Baltussen, A. Wisthaler, T. Märk, R. Frankhuizen // Europ. Food Research and Technol. – 2008. – Vol. 227, No 1. – P. 307-317.
166. Galtier O. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra / O. Galtier, N. Dupuy, Y. Le Dréau, D. Ollivier, C. Pinatel, J. Kister, J. Artaud // Anal. Chim. Acta. – 2007. – Vol. 595, No 1-2. – P. 136-144.
167. Родионова О. Е. Интервальный метод обработки результатов многоканальных экспериментов : дис. … докт. физ.-мат. наук : 01.04.01 / О. Е. Родионова. – М. : Ин-т хим. физики им. Н. Н. Семенова РАН, 2008. – 272 с.

Глава 2

РОБАСТНІ АЛГОРИТМИ КЛАСИФІКАЦІЇ БАГАТОВИМІРНИХ ХІМІЧНИХ ДАНИХ ЗА ДОПОМОГОЮ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

2.1. Основні характеристики та архітектура нейронних мереж

Штучні нейронні мережі (ШНМ) – це математичні моделі, що складаються з елементарних одиниць обробки інформації (нейронів) (рис. 2.1), певним чином сполучених один з одним та із зовнішнім середовищем [1–5].

Штучна нейронна мережа нагадує діяльність мозку в двох аспектах: а) знання потрапляють до нейронної мережі ззовні і використовуються в процесі навчання; б) для накопичення знань застосовуються зв'язки між нейронами, які називають синаптичними вагами. У літературі нейронні мережі часто називають нейрокомп'ютерами, мережами зв'язків, коннекційними мережами або моделями, інтелектуальними системами [1–5].

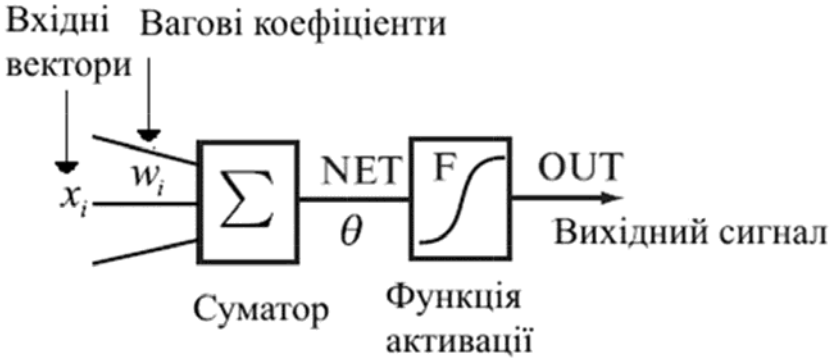

Рис. 2.1. Схема штучного нейрона

Складовою частиною мережі є штучний нейрон. На вхід нейрона поступає деяка множина вхідних векторів (сигналів) x_i ($i = 1, 2, ..., n$). Кожен вхід помножується на відповідний ваговий коефіцієнт нейрона w_i (коефіцієнт міжнейронного зв'язку), всі добутки підсумовуються, визначаючи рівень активації нейрона NET:

$$NET = \sum_{i=1}^{n} x_i w_i + \theta, \tag{2.1}$$

де θ – пороговий рівень нейрона (чи зміщення нейрона).

Цей сигнал (далі – NET) перетворюється активаційною функцією F, формуючи на виході нейронний сигнал OUT [6]:

$$OUT = F(NET). \tag{2.2}$$

Основні види функцій активації [7, 8] представлено в табл. 2.1.

Таблиця 2.1

Функції активації нейронів

Назва	Формула
Лінійна	$OUT = k \cdot NET$, $k = const$
Логістична	$OUT = \frac{1}{1+e^{-NET}}$
Гіперболічний тангенс	$OUT = \frac{e^{NET} - e^{-NET}}{e^{NET} + e^{-NET}}$
Експоненціальна	$OUT = e^{-NET}$
Радіальна базисна	$F_{RadBas} = e^{-\left[\frac{\sqrt{\sum_{i=1}^{n}(x_i - w_i)^2}}{2\delta}\right]^2}$, δ – параметр відхилення або згладжування
«Жорстка сходинка»	$OUT = \begin{cases} 0, NET < \theta \\ 1, NET \geq \theta \end{cases}$
Синусоїдальна	$OUT = \sin(NET)$ для $NET = \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ або $NET = [-\pi, \pi]$
SOFTMAX	$OUT = \frac{e^{NET}}{\sum_i e^{NET_i}}$

Нейрони об’єднуються в шари. Можна виділити три типи нейронів залежно від виконуваних в мережі функцій [9]:

1) вхідні нейрони – слугують для прийому та розподілу вхідних векторів і не виконують розрахунків;
2) приховані нейрони – складають основу нейронних мереж та здійснюють перетворення вхідних сигналів за вищенаведеними формулами (2.1) і (2.2); мережа може включати один або декілька шарів прихованих нейронів;
3) вихідні нейрони – формують виходи або результати роботи нейронної мережі; розрахунки в них також виконуються за формулами (2.1) і (2.2).

Алгоритми штучних нейронних мереж надзвичайно різноманітні за своїми конфігураціями. Виділяють такі типи нейронних мереж:

– однонаправлені і двонаправлені,
– рекурентні,
– мережі на основі радіальних базисних функцій і самоорганізовані карти [1, 10, 11].

Нейронні мережі класифікують за різними ознаками [8, 12]:

– тип вхідної інформації (аналогові і бінарні);
– тип навчання (нейронні мережі «з навчанням» і «без навчання»);
– характер налаштування вагових коефіцієнтів (нейронні мережі з фіксованими або динамічними зв’язками);
– топологія (одношарові і багатошарові мережі);
– характер зв’язків (мережі з прямими, перехресними, зворотними або латеральними зв’язками).

Побудова нейронної мережі складається з двох етапів [8, 13]:

1) вибір архітектури мережі – ініціалізація мережі, визначення кількості нейронів, типу з’єднання нейронів, виду активаційних (передавальних) функцій;
2) навчання мережі – вибір способу навчання (алгоритми локальної оптимізації з розрахунком часткових похідних першого порядку або з розрахунком часткових похідних першого і другого порядків, стохастичні алгоритми оптимізації, алгоритми глобальної оптимізації) і підбір значень вагових коефіцієнтів; зміст навчання нейромережі полягає в мінімізації функціонала похибки.

Питання про кількість проміжних (прихованих) нейронів є важливим, оскільки мережа з великою кількістю нейронів моделює складніші залежності і, отже, схильна до перенавчання, а мережа з невеликою кількістю нейронів може виявитися недостатньо ефективною для вирішення поставленого завдання.

І. Й. Баскін запропонував [14] трьохвибірковий підхід на основі формування навчальної, тестової та контрольної вибірок, який повинен запобігати «перенавчанню» нейромереж при розв'язанні задач прогнозування фізико-хімічних характеристик органічних сполук.

Застосовуючи штучні нейронні мережі для розв'язання задач ідентифікації в якісному хімічному аналізі, для контролю «перенавчання» нейромереж ми варіювали кількість прихованих нейронів і визначали оптимальну їх кількість за навчальною і тестовою вибірками. Цей підхід є досить простим. Зауважимо, що хіміки не завжди мають достатню кількість зразків, що робить неможливим формування навчальної і контрольної вибірок. Існуючі підходи до оцінки кількості нейронів у прихованих шарах прийнятні лише для однорідних нейронних мереж [15, 16].

Практичне застосування алгоритмів нейронних мереж дуже різноманітне. Виконувані ними функції можна розділити на декілька основних груп: апроксимація й інтерполяція, розпізнавання та класифікація образів; стиснення даних; прогнозування; ідентифікація; управління; асоціація. Нейронні мережі володіють такою цінною властивістю, як здатність до навчання (мережу можна навчити розв'язанню певної задачі, виконавши алгоритм навчання), здатність до узагальнення (після навчання мережа може працювати із зашумленими чи спотвореними даними, даючи на виході правильний результат) [6, 7, 17].

Інша важлива властивість нейронних мереж, що свідчить про їх потенціал і широкі прикладні можливості, полягає в паралельній обробці інформації одночасно всіма нейронами. Завдяки цій здатності при великій кількості міжнейронних зв'язків процес обробки інформації значно прискорюється.

Мережа володіє рисами штучного інтелекту. Натренована на обмеженій множині навчальних вибірок, вона узагальнює накопичену інформацію і виробляє очікувану реакцію стосовно даних, що в процесі навчання не оброблялися.

2.2. Значення штучних нейронних мереж для хімії

Штучні нейронні мережі знайшли широке застосування в хімії. Це пов'язано з тим, що вони є потужними засобами обробки даних нелінійного типу. Сферами використання нейронних мереж у хімії є, зокрема, планування експерименту (наприклад, пошук оптимальних умов проведення хроматографічного чи атомно-абсорбційного аналізу), моделювання й інтерпретація сигналів інструментальних методів, складання пептидних карт, вивчення кількісного взаємозв'язку будови та активності / властивостей молекул (Quantitative Structure Activity / Property Relationship, QSAR / QSPR).

Спільно з методологією QSAR / QSPR нейронні мережі визначають математичні залежності між активністю / властивостями молекул та їх топологічними, фізико-хімічними, геометричними й електронними дескрипторами, оцінюють найбільш інформативні з них, а також передбачають властивості нових сполук [18–20].

У хроматографії нейронні мережі використовують, зокрема, при плануванні експерименту − прогнозуванні індексів утримування Ковача [21], часу утримування [22−24], індексу міграції для міцелярної електрокінетичної хроматографії [25]. Градуювальні криві в методі атомно-абсорбційного аналізу також будують із використанням штучних нейронних мереж [26, 27]. Нейронні мережі активно використовують для інтерпретації та моделювання спектрів різної природи: спектрів ЯМР [28−30], ІЧ-спектрів [31], спектрів рухливості йонів [32], спектрів лазерної індукованої флуоресценції [33], гамма-спектрів [34], масспектрів [35, 36], електронних спектрів поглинання в УФ і видимому діапазонах [37], оптимізації умов проточно-інжекційного аналізу зі спектрофотометричним детектуванням [38]. Капілярний зонний електрофорез, важливий для дослідження пептидів, комбінують із нейронними мережами, основне завдання яких полягає в прогнозуванні електрофоретичної рухливості пептидів [39, 40]. Відзначимо також застосування рекурентної нейронної мережі для класифікації білків [41].

У табл. 2.2 наведено основні парадигми штучних нейронних мереж, що використовуються для розв'язання хімічних задач [9, 42−45].

Незважаючи на різноманітне і широке застосування ШНМ, питання про їх характеристики (архітектуру, типи функцій активації, кількість прихованих нейронів, методи навчання), необхідні і достатні для надійного розв'язання того чи іншого класу задач, залишається відкритим.

Таблиця 2.2

Основні парадигми нейронних мереж

Назва	Сфера застосування	Недоліки	Переваги
Мережа радіальної основи (Radial Basis Function Network)	Розпізнавання образів, класифікація	Об'єм мережі	Проста архітектура, швидке навчання
Мережа прямої передачі сигналу (Feed Forward Network)	Розпізнавання образів, класифікація, прогнозування	Низька швидкість навчання	Дозволяє вирішити численні практичні завдання
Мережа Кохонена (Kohonen's Neural Network)	Кластерний аналіз, класифікація, розпізнавання образів	Мережа може бути використана для кластерного аналізу лише тоді, коли заздалегідь визначено кількість кластерів	Мережа здатна функціонувати в умовах перешкод
Рекурентна мережа Елмана (Elman's Network)	Розпізнавання образів, класифікація, асоціативна пам'ять	Мережа має невелику місткість	Дозволяє відновити спотворені сигнали

Синтез мережі визначається задачею дослідження та особливостями первинного масиву експериментальних даних, тому користувач підбирає оптимальні параметри мережі для розв'язання свого конкретного завдання методом «спроб і помилок». У зв'язку з цим актуальним є розроблення рекомендацій щодо вибору характеристик нейронних мереж, зокрема для розв'язання задач якісного хімічного аналізу.

Слід згадати проблему, пов'язану з алгоритмами класифікації «з навчанням», – формування навчальної вибірки. Навчальна вибірка (training set) – це вибірка, за якою проводять налаштування параметрів алгоритму класифікації. Навчальна вибірка містить усю інформацію про задачу, і від її репрезентативності залежить ефективність будь-якого алгоритму навчання. Запропоновано декілька шляхів вибору навчальної вибірки: випадковий алгоритм; різні

варіації алгоритму найближчого сусіда; способи, засновані на об'єднанні концепції поетапного введення зразків до навчальної вибірки і стратифікації [46−48]. Запропоновані способи формування навчальної вибірки вимагають багато часу і є громіздкими, при цьому не дають відповідей на важливі питання: «Скільки зразків із відомою класовою приналежністю треба взяти для наступної класифікації зразків із невідомою класовою приналежністю? Наскільки можуть і наскільки повинні ці зразки розрізнятися за своїми властивостями? Як перевірити однорідність зразків навчальної вибірки»? Тому важливим нерозв'язаним питанням є процедура формування навчальної вибірки й оцінка її однорідності та репрезентативності.

2.3. Алгоритми деяких нейронних мереж

Надамо короткий опис деяких поширених алгоритмів штучних нейронних мереж [1−14].

Мережа Кохонена

Нехай наявна множина *n* об'єктів, кожен з яких характеризується *m*-вимірним вектором параметрів $\boldsymbol{x_i}$, $\{x_{i1}, x_{i2}, \ldots, x_{im}\}$, і цю множину слід розбити на *K* кластерів (груп), до яких входять об'єкти з подібними властивостями. Мережа Кохонена (Kohonen's Neural Network) розв'язує цю задачу. Вона відноситься до нейронних мереж із самоорганізацією на основі конкуренції нейронів.

Критерієм близькості об'єктів *A* та *B* слугує Евклідова відстань між ними $d_{AB} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$. Якщо відстань d_{AB} менше певного порогу σ, вважають, що об'єкти належать одному кластеру; в іншому разі − різним кластерам.

Мережа Кохонена має два шари: вхідний і вихідний. Кількість нейронів дорівнює кількості класів *K*. В якості нейронів виступають зважені суматори $s_j = b_j + \sum_{i=1}^{m} w_{ij} x_i$, де *j* – номер нейрона, w_{ij} – вага *i*-го входу *j*-го нейрона, b_j – пороговий рівень. Кожний нейрон мережі сполучений з усіма компонентами вхідного вектора $\boldsymbol{x_i}$. Сигнал s_j

обробляється конкуруючою функцією активації, яка працює за правилом «переможець забирає все»: визначається суматор із максимальним значенням виходу, і саме цьому нейрону Кохонена на виході приписується логічна одиниця, а інші нейрони видають нуль (рис. 2.2).

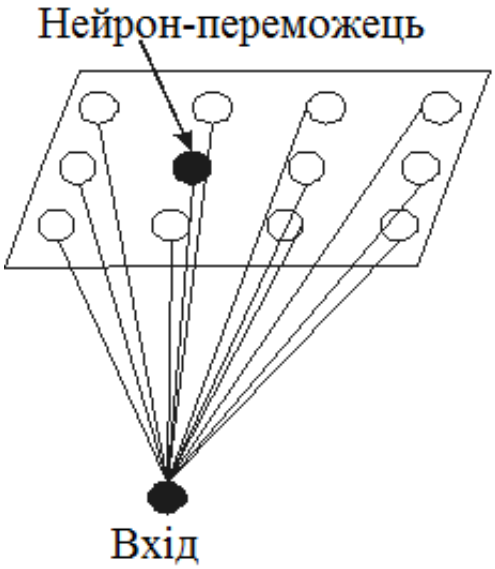

Рис. 2.2. Схема роботи функції конкуренції в моделі Кохонена

«Нейрон-переможець» $\boldsymbol{s_j}$ визначається з умови $d(\boldsymbol{x},\boldsymbol{s}_v)=\min\limits_{1\le j\le K} d(\boldsymbol{x},\boldsymbol{s}_j)$, де v – номер нейрона-переможця, $d(\boldsymbol{x},\boldsymbol{s_j})$ – відстань між векторами $\boldsymbol{x}$ та $\boldsymbol{s_j}$, $d_j=\sqrt{\sum\limits_{i=1}^{n}\left(x_i-s_{ij}\right)^2}$. Вагові коефіцієнти «нейрона-переможця» коректуються згідно з правилом Кохонена:

$$s_{vj}(t+1)=s_{vj}(t)+\eta\cdot(x_j(t)-s_{vj}(t)), \quad (2.3)$$

де x_j – j-й елемент вхідного вектора; s_{vj} – ваговий коефіцієнт, що характеризує зв'язок між j-м елементом вхідного вектора і нейроном v; η – крок навчання або коефіцієнт швидкості навчання (додатне число, менше одиниці); t – номер ітерації.

Розрахунок міри близькості і коригування вагових коефіцієнтів тривають, поки мережі не будуть пред'явлені всі вхідні вектори.

LVQ-мережі

Розвитком мережі Кохонена є LVQ-мережі (Learning Vector Quantization), які містять прихований конкуруючий шар (виконує кластеризацію векторів) і лінійний шар вихідних нейронів (співвідносить кластери з цільовими класами, заданими користувачем). LVQ-мережі інформацію про приналежність векторів до певних класів використовують для того, щоб віднести до одного класу вектори, які активізують один і той самий нейрон. Це досягається за рахунок використання різного знака перед кроком навчання η у формулі (2.3). Таке правило гарантує, що при правильній класифікації «нейрон-

переможець» наближається до вхідних векторів, а при неправильній класифікації віддаляється від них.

Мережа прямого поширення сигналу

«Класична» мережа прямого поширення сигналу (Feed Forward Neural Network, FFNN) характеризується передачею інформації з попередніх шарів на наступні (рис. 2.3). Часто FFNN містить прихований шар нейронів із сигмоїдальними функціями активації, тоді як вихідний шар містить нейрони з лінійними функціями активації. Вихідний шар містить стільки нейронів, скільки класів об’єктів міститься в навчальній вибірці. Кількість вхідних нейронів дорівнює кількості характеристик досліджуваних зразків.

Каскадна нейронна мережа

Каскадна нейронна мережа (Cascade Neural Netowrk, CNN) – мережа прямої передачі, в якій підбір структури мережі відбувається паралельно з її навчанням, шляхом додавання на кожному етапі навчання одного прихованого нейрона. Кожен черговий нейрон підключається до вхідних векторів і до всіх уже існуючих прихованих нейронів.

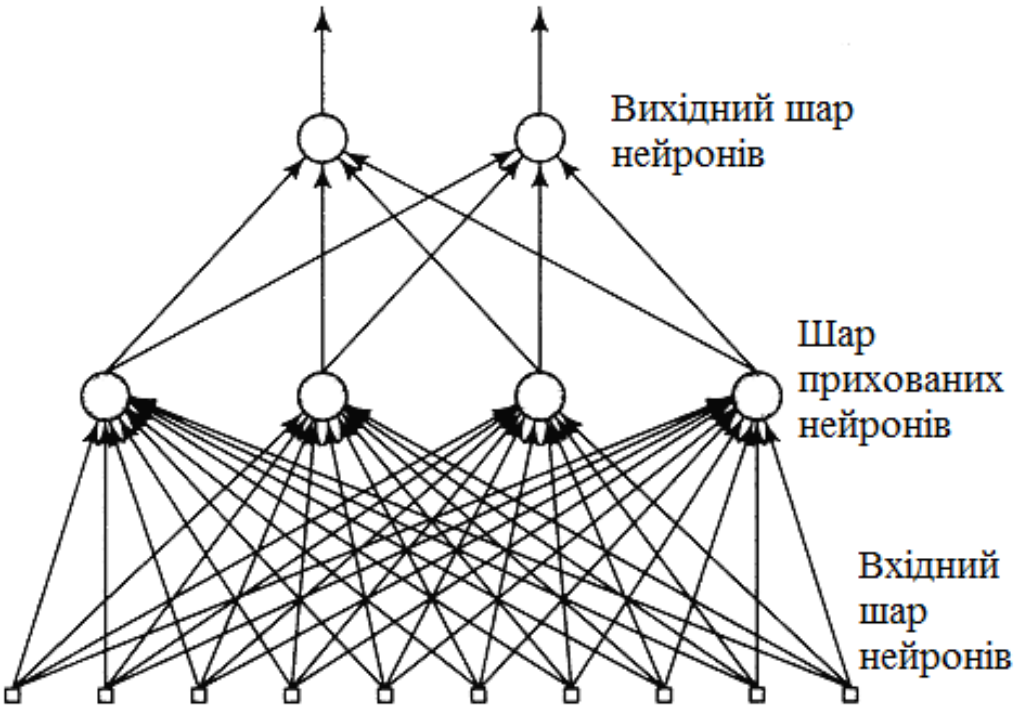

Рис. 2.3. Схема мережі прямого поширення з одним прихованим і одним вихідним шарами

Динамічна нейронна мережа

Динамічна нейронна мережа (Dynamic Neural Network, DNN) містить лінії затримки, тому вхід мережі потрібно розглядати як послідовність векторів, що подаються на мережу в певні моменти часу.

Мережа Елмана

Мережа Елмана (Elman's Neural Network, ENN) відноситься до рекурентних нейронних мереж і характеризується наявністю зворотного зв'язку між прихованим і вхідним шаром, що дозволяє врахувати передісторію спостережуваних процесів і накопичити інформацію для вироблення правильної стратегії управління.

Для реалізації алгоритмів двошарових нейромереж «з навчанням» використали різні комбінації функцій активації, алгоритми навчання і кількість прихованих нейронів h з метою формулювання рекомендацій щодо вибору параметрів мереж (табл. 2.3).

Таблиця 2.3

Параметри нейронних мереж

Тип параметра	Опис
Метод навчання	Левенберга–Марквардта, Пауела–Бієле, алгоритм зворотного поширення помилки
Функція активації	Гіперболічний тангенс, лінійна, логістична

Обрані функції активації є безперервними і диференційованими на всій числовій осі функції активації (рис. 2.4).

Логістична функція має властивість «посилювати» слабкі сигнали і запобігати насиченню від великих сигналів, оскільки вони відповідають областям аргументів, де сигмоїд має пологий схил. Функція гіперболічний тангенс симетрична відносно точки (0, 0), що є перевагою порівняно з логістичною кривою [7, 8].

Для ініціалізації вагових коефіцієнтів і зміщень нейронів мережі ми використали алгоритм Д. Нгуєна і Б. Відроу, який дозволяє значно прискорити процес навчання й усунути зупинку алгоритму навчання в локальних мінімумах. Згідно з цим алгоритмом, вагові коефіцієнти прихованих нейронів h ініціалізувалися в межах $\left[-\sqrt[l]{h},\sqrt[l]{h}\right]$, де l – кількість входів нейрона. Початкові ваги вихідних нейронів вибираються випадковим чином з інтервалу $\left[-0.5,0.5\right]$ [10, 49]. Зміщення

прихованих нейронів ініціалізувалися випадковим чином в інтервалі $[-l^r, l^r]$, де r – розмірність вхідних векторів; зміщення вихідних нейронів призначали однаковими.

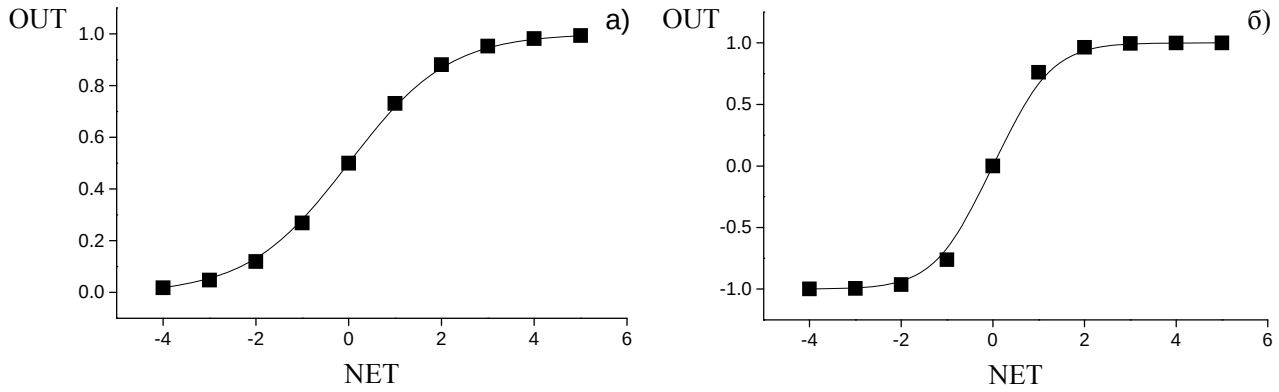

Рис. 2.4. Приклади активаційних функцій: а) логістична функція, б) гіперболічний тангенс

Алгоритм зворотного поширення похибки (Error Back Propagation) – це ітераційний градієнтний алгоритм навчання [9]. Вказаний алгоритм мінімізує напівсуму квадратів різниць між бажаною величиною виходу d_k і реально отриманими на виходах мережі значеннями y_k для кожного об'єкта *k*:

$$E = \frac{1}{2}\sum_{k=1}^{S}\sum_{i=1}^{Y}\left(d_k^i - y_k^i\right)^2, \qquad (2.4)$$

де S – кількість об'єктів у навчальній вибірці, Y є кількістю виходів багатошарової мережі.

Підсумовування ведеться по всіх нейронах вихідного шару і за всіма зразками, що обробляє мережа. Мінімізацію E забезпечує зміна вагових коефіцієнтів у такий спосіб:

$$\Delta w_{ij}^{(q)} = -\eta\frac{\partial E}{\partial w_{ij}}, \qquad (2.5)$$

де w_{ij} – ваговий коефіцієнт нейрона шару q, η – коефіцієнт швидкості навчання $(0 < \eta < 1)$.

Відповідно до правила диференціювання складної функції,

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y^i} \frac{dy^i}{dNET^i} \frac{\partial NET^i}{\partial w_{ij}}, \qquad (2.6)$$

де NET^i – зважена сума вхідних сигналів нейрона i, тобто аргумент функції активації (похідна функції активації має бути визначеною на всій осі абсцис).

Алгоритм зворотного поширення помилки коригує параметри налаштування в напрямі найшвидшого зменшення функціонала помилки. При фіксованому значенні параметра швидкості навчання алгоритм може зациклитися поблизу вузького мінімуму або «застрягти» в дрібних локальних мінімумах. У зв'язку з цим, алгоритм зворотного поширення помилки потребує модифікації. Необхідно а) ввести момент інерції, що дозволяє долати нерівності поверхні помилки і не зупинятися в локальних мінімумах, і б) ввести процедуру адаптивного підбору коефіцієнта швидкості навчання [10, 13, 14]. З урахуванням введення параметра моменту інерції μ зміна вагового коефіцієнта в процесі навчання на *t*-й ітерації має вид $\Delta w_{ij}^t = -\eta \frac{\partial E}{\partial w_{ij}} + \mu \, \Delta w_{ij}^{t+1}$, $\mu = 0.9$. Процедура адаптивного підбору коефіцієнта швидкості навчання заснована на таких міркуваннях: призначається початкове значення $\eta = 0.01$; при зменшенні критерію якості навчання мережі η збільшується в 1.05 рази; при збільшенні критерію якості навчання мережі більш ніж в 1.04 рази починається корекція η у бік зменшення з коефіцієнтом 0.7.

Алгоритм Левенберга–Марквардта належить до квазіньютонівських алгоритмів оптимізації. В ньому певним чином (наближено, але досить близько до точної оцінки) знаходять матрицю Гессе других часткових похідних функціонала помилки. Матрицю Гессе H обчислюють як $H \cong J^T J$, де J – матриця Якобі похідних функціоналів помилки окремо для кожного вихідного нейрона і для кожного об'єкта в навчальній вибірці за параметрами, що налаштовуються; градієнт розраховують за формулою $g = J^T e$, де e – вектор помилок мережі [10, 14].

Алгоритм зв'язаних градієнтів, зокрема метод Пауела–Бієле, на першій ітерації починає пошук у напрямі антиградієнта $p_0 = -q_0$. Після вибору напряму визначають крок пошуку a_t, на величину якого слід змінити параметри налаштування $w_{ij}(t+1) = w_{ij}(t) + a(t) \cdot p(t)$. Потім визначають наступний напрям пошуку як лінійну комбінацію нового напряму найшвидшого спуску і вектора руху в зв'язаному напрямі $p(t+1) = -g(t+1) + \beta(t+1) \cdot p(t)$, де β – константа [13].

Критерієм зупинки навчання алгоритмів нейронних мереж було досягнення заданого значення середньої квадратичної помилки (Mean Squared Error, mse):

$$mse = \frac{\sum_{i=1}^{S}\sum_{i=1}^{Y}\left(d_k^i - y_k^i\right)^2}{S}. \qquad (2.7)$$

Окремого розгляду вимагає ймовірнісна нейронна мережа (Probabilistic Neural Network, PNN) – модифікація радіальних базисних нейронних мереж. PNN містить прихований шар нейронів із радіальною базисною функцією активації (див. табл. 2.1), кожен з яких призначений для зберігання окремого еталонного зразка (немає необхідності визначати кількість прихованих нейронів і вид функції активації). Вихідний шар імовірнісної мережі – це конкуруючий шар, який підраховує ймовірність приналежності вхідного вектора до того чи іншого класу і, врешті-решт, зіставляє вектор із тим класом, ймовірність приналежності до якого найвища. Навчання ймовірнісної мережі передбачає попереднє проведення кластеризації для визначення центрів класів, для чого найчастіше використовують алгоритм k-середніх. Основна проблема при реалізації PNN полягає у визначенні параметра згладжування функції активації δ: значення параметра має бути досить великим, щоб перешкоджати перенавчанню, але не настільки великим, щоб радіальна базисна функція оголошувала однаково значущими всі значення входу. Вагові коефіцієнти першого шару формуються з використанням вхідних векторів із навчальної множини. Вагові коефіцієнти другого шару відповідають цільовим векторам навчальної вибірки.

2.4. Оцінка надійності і стійкості алгоритмів нейронних мереж

Важливою характеристикою алгоритмів класифікації є їх стійкість (робастність)[1)] до наявності в масивах чисельних характеристик похибок, розподіл яких відрізняється від нормального, і пропусків (дані з пропусками особливо часто зустрічаються при зверненні до масивів архівних даних).

Для перевірки стійкості реалізованих алгоритмів до наявності в масивах даних похибок, розподіл яких відрізняється від нормального, використовували модель «грубих промахів» [51−53].

У вихідні дані характеристик аналізованих об'єктів (тестової вибірки) вносили похибки ε, які розраховували за формулою

$$\varepsilon = \left[(100 - q)\cdot \varepsilon_{Gauss}(0,\sigma) + q \cdot \varepsilon_{Laplas}(0,\sigma)\right]/100, \qquad (2.8)$$

де q, % – інтенсивність «грубих промахів»; ε_{Gauss} – випадкова величина, розподілена за законом Гауса з нульовим середнім і стандартним відхиленням σ; ε_{Laplas} – випадкова величина, що підкоряється розподілу Лапласа з нульовим середнім і стандартним відхиленням $\sigma = \frac{\sqrt{2}}{a}$:

$$p(\varepsilon) = \frac{a}{2}\exp\{-a|\varepsilon - \beta|\},\ \varepsilon \in (-\infty, \infty), \qquad (2.9)$$

де $p(\varepsilon)$ – густина розподілу, $a > 0$ – параметр масштабу, $-\infty < \beta < +\infty$ – параметр зсуву.

Оскільки коефіцієнт ексцесу розподілу Лапласа $\gamma_2 = 3$ більший, ніж нормального розподілу ($\gamma_2 = 0$), із зростанням q підвищується ймовірність появи серед ε «грубих промахів» – похибок, що більше ніж у два-три рази перевищують стандартні відхилення σ [52]. До початкових даних характеристик об'єктів, що класифікуються, вносили похибки при значеннях q = 0, 10, ..., 100 %.

---

1) Стійкість (robustness, ruggedness) – здатність алгоритму видавати прийнятні результати при зміні окремих параметрів; вплив, що викликається зміною одного чи декількох факторів, на результат [50].

Для вивчення стійкості алгоритмів класифікації до наявності в даних пропусків використовували таку процедуру: видаляли 10 і 20 % значень характеристик зразків тестової вибірки і заповнювали отримані пропуски середніми значеннями відповідних властивостей.

Ненадійність результатів класифікації оцінювали як частку неправильно класифікованих зразків тестової вибірки:

$$P = \frac{n_{error}}{n} \cdot 100\ \%, \qquad (2.10)$$

де n_{error} – кількість неправильно класифікованих зразків тестової вибірки, n – загальна кількість зразків тестової вибірки. Надійність результатів алгоритмів класифікації $R = 100 - P$, %.

Особливу увагу приділяли перевірці змістовності класифікації. Контрольну вибірку для перевірки правильності класифікації річкових і джерельних вод складали зразки, відібрані в році, наступному після навчання алгоритмів штучних нейронних мереж. При цьому кількість визначених у цих зразках концентрацій металів відрізнялася від кількості відповідних властивостей еталонів, які використовували для навчання алгоритмів. Можливість отримання змістовної ідентифікації також перевіряли, використовуючи запропоновану процедуру кластеризації за відсутності інформації про кількість класів і навчальну вибірку (див. главу 4). Знайдене при цьому розбиття об'єктів на класи порівнювали з відомим з умов відбору зразків. Окрім цього, правильність класифікації оцінювали за допомогою процедури крос-валідації.

2.5. Апробація алгоритмів нейронних мереж «без навчання» та «з навчанням» на модельних і тестових наборах даних

Для вивчення роботи алгоритмів нейронних мереж при різних керуючих параметрах і розробки рекомендацій з формування оптимального складу навчальної вибірки алгоритми нейронних мереж випробували на наборах модельних і тестових даних. Використали модельні дані, що характеризуються складними структурами: двоієрархічною (рис. 2.5) і дугоподібною з рівною відстанню між об'єктами кожної з дуг (рис. 2.6) [53].

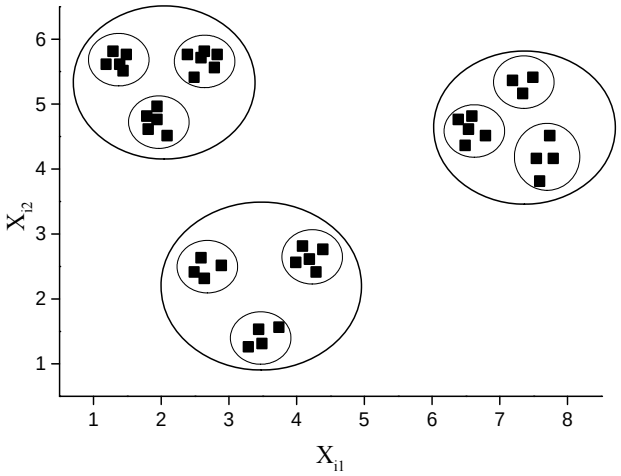

Рис. 2.5. Двоієрархічна структура даних

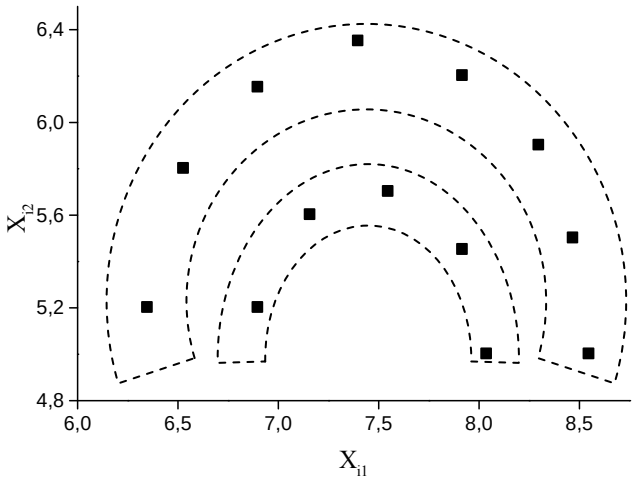

Рис. 2.6. Дугоподібна структура даних

Як еталонні (з надійно встановленою структурою) використали два набори даних, отриманих експериментально. Перший з них містить характеристики квітів ірису (відомий тестовий набір для випробування алгоритмів класифікації і кластерного аналізу) [54], другий – характеристики зразків італійських вин (набір даних, що також широко використовують для тестування алгоритмів класифікації) [55]. Масив даних про квіти ірису містить відомості про 4 характеристики 150 зразків ірису (довжина і ширина чашолистка, довжина і ширина пелюстки). Зразки ірисів поділені на три класи, в кожному з яких міститься по 50 зразків. Масив даних про вина містить результати визначення 13 ознак 178 зразків вин, що належать до трьох класів.

Значення характеристик досліджуваних масивів даних наведено в табл. 2.4, 2.5 та в Додатку (табл. Д1, Д2).

Першим кроком застосування алгоритмів «із навчанням» є формування оптимального складу навчальної вибірки. Навчальна вибірка містить всю інформацію про поставлену задачу, тому ефективність функціонування будь-якої системи, що навчається, залежить від репрезентативності навчальної вибірки [56]. Для формування оптимального складу навчальної вибірки використовували ймовірнісну нейронну мережу, що характеризується простою архітектурою. При цьому приймали, що оптимальний об'єм навчальної вибірки, визначений за допомогою ймовірнісної мережі, можна

використовувати для навчання нейронних мереж й інших типів. Таким чином, досліджувані нейронні мережі знаходяться в рівних умовах, і можна судити про те, яка із запропонованих архітектур нейронних мереж є більш надійною для класифікації конкретної тестової вибірки. Під оптимальним об'ємом навчальної вибірки розуміли таку кількість зразків, яка забезпечувала 100% надійність класифікації зразків тестової вибірки. Наявні набори даних розділяли випадковим чином на навчальну і тестову вибірки при різному їх співвідношенні. Параметр *T*, % показує, яка частка зразків від їхньої загальної кількості знаходиться в навчальній вибірці:

$$T = \frac{S}{M} \cdot 100\,\%, \qquad (2.11)$$

де *S* – кількість зразків у навчальній вибірці, *M* − загальна кількість зразків.

Таблиця 2.4

Дані з дугоподібною структурою

Номер об'єкта	Значення першої ознаки	Значення другої ознаки	Номер об'єкта	Значення першої ознаки	Значення другої ознаки
1	6.35	5.20	8	8.55	5.00
2	6.53	5.80	9	6.90	5.20
3	6.90	6.15	10	7.16	5.60
4	7.40	6.35	11	7.55	5.70
5	7.92	6.20	12	7.92	5.45
6	8.30	5.90	13	8.04	5.00
7	8.47	5.50			

Таблиця 2.5

Дані з двоієрархічною структурою

Номер об'єкта	Значення першої ознаки	Значення другої ознаки	Номер об'єкта	Значення першої ознаки	Значення другої ознаки
1	1.20	5.60	22	3.45	1.52
2	1.30	5.80	23	3.50	1.30
3	1.40	5.60	24	3.75	1.55
4	1.45	5.50	25	4.00	2.55
5	1.50	5.75	26	4.10	2.80

Продовження табл. 2.5

6	2.40	5.75	27	4.15	2.60
7	2.50	5.40	28	4.20	2.40
8	2.60	5.70	29	4.25	2.75
9	2.65	5.80	30	6.40	4.75
10	2.80	5.55	31	6.50	4.35
11	2.84	5.75	32	6.55	4.60
12	1.80	4.80	33	6.60	4.80
13	1.82	4.60	34	6.80	4.50
14	1.95	4.75	35	7.20	5.35
15	1.95	4.95	36	7.35	5.15
16	2.10	4.50	37	7.50	5.40
17	2.50	2.40	38	7.55	4.15
18	2.60	2.62	39	7.60	3.80
19	2.65	2.30	40	7.75	4.50
20	2.90	2.50	41	7.80	4.15
21	3.30	1.25			

Безпомилкова класифікація зразків із дугоподібною структурою при використанні ймовірнісної мережі ($\delta = 0.1$)[2)] спостерігається при $T = 70\,\%$, зразків із двоієрархічною структурою – при $T = 60\,\%$, зразків ірисів – при $T = 70\%$, зразків вин – при $T = 87\,\%$ (рис. 2.7).

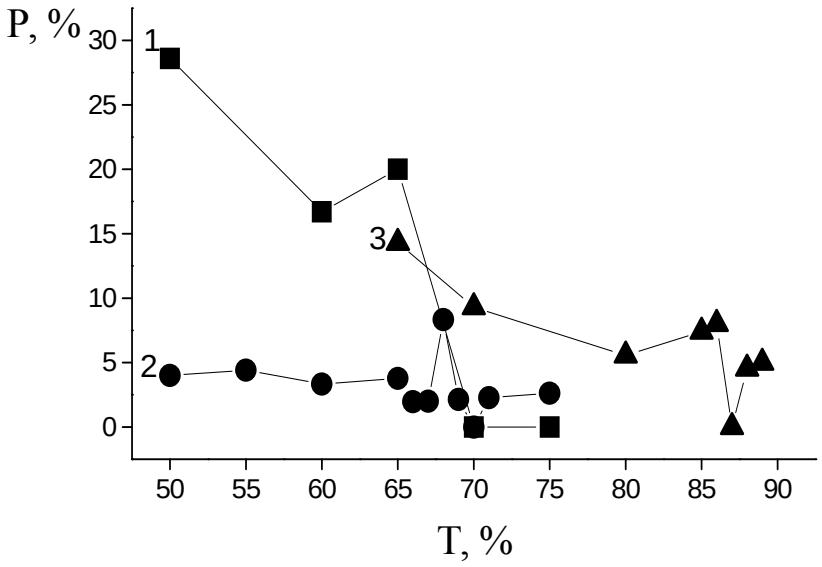

Рис. 2.7. Залежність значень частки неправильно класифікованих зразків тестової вибірки від об'єму навчальної вибірки для ймовірнісної нейронної мережі: 1 – дані з дугоподібною структурою, 2 – зразки ірисів, 3 – зразки вин

[2)] Значення $\delta = 0.1$ не є єдино можливим. Наприклад, для надійної класифікації даних із дугоподібною структурою оптимальне значення відхилення радіальної базисної функції знаходиться в межах $0.02 < \delta < 0.6$. У тексті вказуємо одне з можливих значень δ.

У випадку даних із двоієрархічною структурою навчальна вибірка формувалася так, щоб мінімум два-три представники всіх дев'яти класів входили до її складу для повнішого опису поставленого завдання.

Для ідентифікації зразків вин, характеристикам яких притаманний великий розмах значень – від частки одиниці до декількох сотень, попередньо провели автомасштабне перетворення даних (приведення до нульового середнього значення й одиничної дисперсії) [57]:

$$x_i^{norm} = \frac{x_i - \bar{x}}{std(x)}, i = 1, 2, ..., n, \qquad (2.12)$$

де x^{norm} – безрозмірне значення характеристики для i-го зразка, отримане внаслідок автомасштабного перетворення, x_i – вихідне значення характеристики для i-го зразка, $\bar{x}$ – середнє значення характеристики в зразках, $std(x)$ – стандартне відхилення значень характеристики в зразках, n – кількість зразків.

Параметри алгоритмів нейронних мереж з навчанням, що забезпечують стовідсоткову надійність класифікації тестових наборів даних, представлені в табл. 2.6. Для кожного алгоритму нейронної мережі визначили оптимальні алгоритм навчання, комбінацію функцій активації для прихованого / вихідного шарів і кількість прихованих нейронів [58, 59]. Під оптимальними розуміли такі параметри, що забезпечують коректне навчання мережі та правильне віднесення до відповідних класів зразків тестової вибірки. Вибір методу навчання і функцій активації детально описано далі в розд. 2.6.

Проілюструємо вибір оптимальної кількості прихованих нейронів і кроку навчання на прикладі мережі прямої передачі сигналу і LVQ-мережі (рис. 2.8, 2.9). Критерієм зупинки навчання було досягнення значення середнього квадратичного відхилення, що дорівнює 0.001 (див. формулу (2.4)).

Рис. 2.8, 2.9 ілюструють явища «недонавчання» та «перенавчання» нейронної мережі прямої передачі сигналу: мережа з $h < 6$ не здатна розв'язати поставлену задачу, мережа з $h > 7$ правильно навчається («заучує» приклади), але не здатна працювати з новими даними. Подібні твердження стосуються і LVQ-мережі для $h < 6$ і $h > 11$.

Таблиця 2.6

Оптимальні параметри нейронних мереж із навчанням

Алгоритм ШНМ	Параметри ШНМ			
	Дугоподібні дані	Двоієрархічні дані	Зразки ірисів	Зразки вина
FFNN	метод Левенберга–Марквардта, функції активації: гіперболічний тангенс, лінійна			
	h=6, 7	h=11, 12	h=7, 19	h=14
CNN	метод Левенберга–Марквардта, функції активації: гіперболічний тангенс, лінійна			
	h=6	h=6	h=7	h=12
DNN	метод Левенберга–Марквардта, функції активації: гіперболічний тангенс, лінійна			
	h=8, 17	h=9–14	h=8, 15	h=14
ENN	метод зворотного поширення помилки, функції активації: гіперболічний тангенс, гіперболічний тангенс		метод зворотного поширення помилки, функції активації: гіперболічний тангенс, лінійна	
	h=7	h=10	h=9	h=12

Для динамічної мережі та мережі прямої передачі сигналу є декілька варіантів оптимальних значень кількості прихованих нейронів для надійної класифікації модельних і тестових даних. Рекомендують [13] робити вибір на користь простішої моделі нейронної мережі.

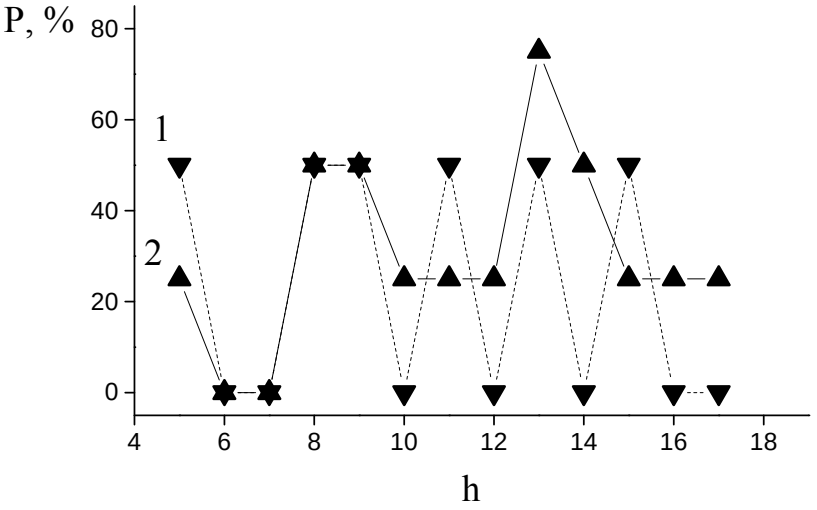

Рис. 2.8. Залежність значень частки неправильно класифікованих зразків навчальної і тестової вибірок даних із дугоподібною структурою від кількості прихованих нейронів (h) для мережі прямої передачі сигналу: 1 – навчальна вибірка, 2 – тестова вибірка

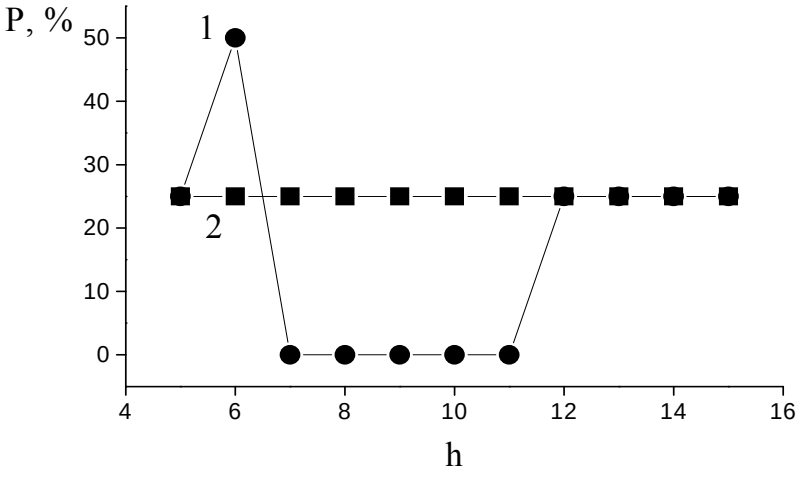

Рис. 2.9. Залежність значень частки неправильно класифікованих зразків даних із дугоподібною структурою від кількості нейронів для LVQ-мережі при різних значеннях кроку навчання: 1 – η = 0.03; 2 – η = 0.01, 0.1, 0.05, 0.04

Для підтвердження або спростування цієї рекомендації оцінили стійкість мережі прямої передачі сигналу та динамічної мережі при відповідних оптимальних значеннях кількості прихованих нейронів до наявності похибок у початкових даних. Початкові значення параметрів, що характеризують дані із дугоподібною і двоієрархічною структурою і зразки ірисів, модифікували шляхом внесення похибок за формулою (2.5).

Проілюструємо внесення в дані «грубих промахів» на прикладі однієї з характеристик квітів ірису – довжин чашолистків (рис. 2.10). Внесені похибки показані на рис. 2.11 для $q = 10\ \%$ і $q = 80\ \%$.

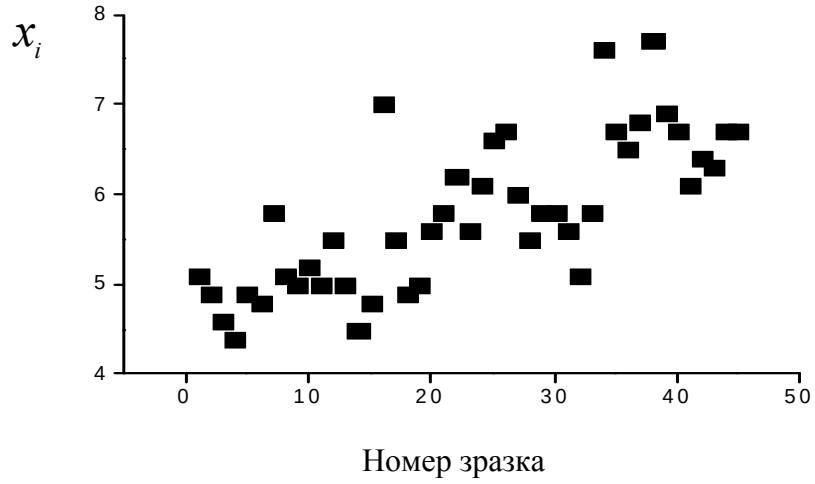

Рис. 2.10. Довжини чашолистка зразків квітів ірису, що знаходяться в тестовій вибірці (x_i)

Залежності, представлені на рис. 2.12 для мережі прямої передачі сигналу з 7 і 19 нейронами у прихованому шарі при інших параметрах, вказаних у табл. 2.8, дозволяють зробити висновок про те, що мережа з меншою кількістю прихованих нейронів не поступається мережі з більшою кількістю нейронів за стійкістю до наявності в даних похибок.

Подібні залежності спостерігалися і для динамічної мережі при $h = 8$ і $h = 15$ при класифікації зразків ірису (рис. 2.13), і при $h = 8$ і $h = 17$ при класифікації даних із дугоподібною структурою (рис. 2.14). Прості моделі мереж стійкіші при невеликих значеннях внесених промахів: $0 < q \le 30$ %, що також говорить на їх користь, оскільки ймовірність появи більшої кількості промахів у хімічному експерименті мала. У зв'язку з цим ми також рекомендуємо робити вибір на користь простішої нейронної мережі. Далі будуть наведені найпростіші оптимальні архітектури нейронних мереж.

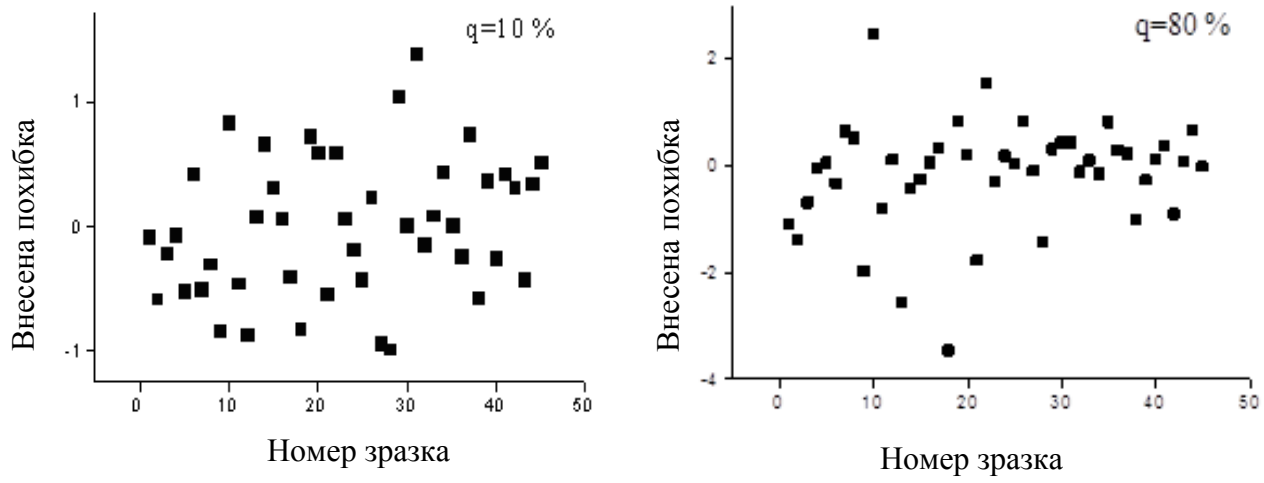

Рис. 2.11. Похибки, внесені в довжини чашолистків квітів ірису при $\sigma = 0.1 \cdot \bar{x}_i$

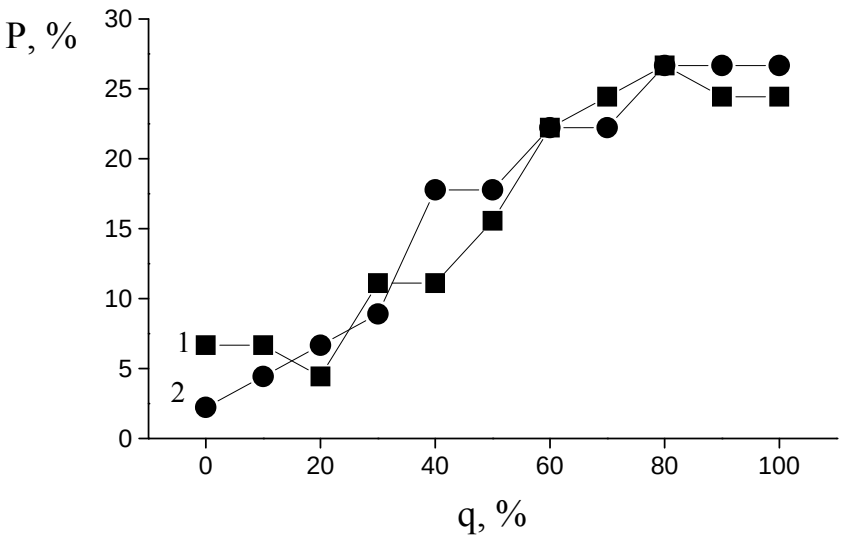

Рис. 2.12. Залежність часток неправильно ідентифікованих зразків ірису від інтенсивності «грубих промахів», внесених у первинні характеристики, для мережі прямої передачі сигналу: 1 – мережа з $h = 19$, 2 – мережа з $h = 7$

Відзначимо, що нейронні мережі без навчання і традиційні алгоритми класифікації не дозволяють виявити реальну структуру даних (табл. 2.7). При розрахунках за методом опорних векторів для всіх масивів даних використовували в якості ядра радіальну базисну функцію Гауса

$$K(x_i, x) = e^{-\gamma \|x_i - x\|^2}, \qquad (2.13)$$

де параметр $\gamma = 1/(2\sigma^2)$, σ^2 – дисперсія.

Таблиця 2.7

Результати класифікації модельних і тестових даних алгоритмами LVQ-мережі, мережі Кохонена і традиційними алгоритмами

Алгоритм	P, %			
	Дугоподібні дані	Двоієрархічні дані	Зразки ірисів	Зразки вин
Мережа Кохонена (η=0.01)	39	0	9	11
LVQ -мережа	0 ($7 \le h \le 11$, η = 0.03)	не класифікує	*	*
LDA	50	0	0	0
SVM	0	0	0	*
SIMCA	75	44	4	*
PLS-DA	50	44	7	8

* Алгоритм коректно не навчається.

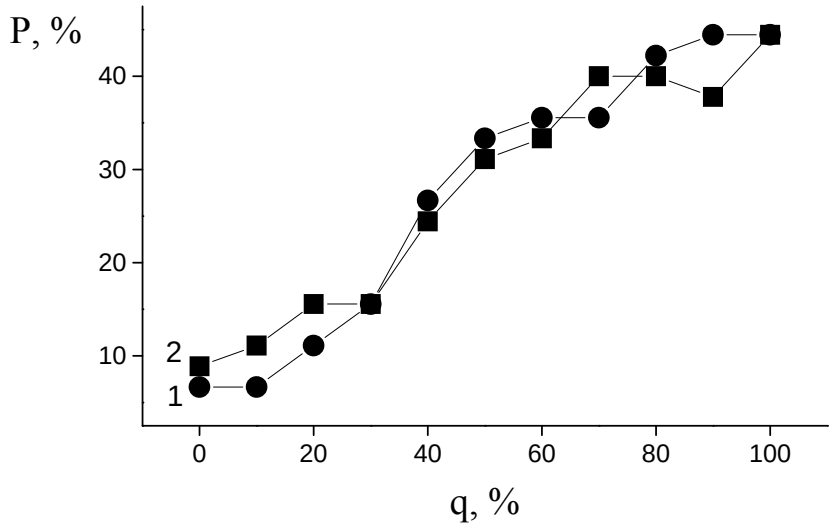

Рис. 2.13. Залежність часток неправильно ідентифікованих зразків ірису від інтенсивності «грубих промахів», внесених у первинні характеристики, для динамічної мережі: 1 – мережа з $h = 8$, 2 – мережа з $h = 15$

При використанні формального моделювання аналогій класів та дискримінантного аналізу за допомогою регресії на латентні структури досліджувані класи модельних і тестових даних моделювали за допомогою однієї головної компоненти / прихованої змінної. Методи SIMCA і PLS-DA, а також LVQ-мережа виявилися найменш ефективними. Збільшення кількості головних компонент / прихованих змінних для SIMCA і PLS-DA, відповідно, не дозволило зменшити помилку класифікації і правильно навчити метод для ідентифікації зразків вин.

Табл. 2.8 містить основні параметри традиційних методів класифікації: для SVM – кількість опорних векторів N_{sv} і значення параметра ядра γ в формулі (2.13), для SIMCA і PLS-DA – значення сумарної дисперсії ознак, що описують головні компоненти / приховані змінні Var, для SIMCA вони наведені для кожного класу.

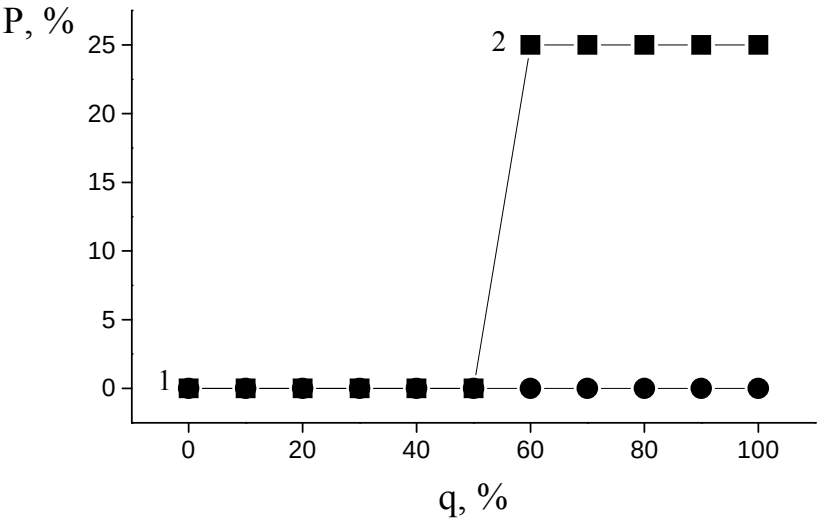

Рис. 2.14. Залежність часток неправильно класифікованих даних із дугоподібною структурою від часток інтенсивності «грубих промахів», внесених у первинні характеристики, для динамічної мережі: 1 – мережа з $h = 8$, 2 – мережа з $h = 17$

Таблиця 2.8

Параметри традиційних методів класифікації

Метод	Параметри алгоритмів			
	Дугоподібні дані	Двоієрархічні дані	Зразки ірисів	Зразки вина
SVM	$N_{sv} = 8$, $\gamma = 0.316$	$N_{sv} = 20$, $\gamma = 0.0316$	$N_{sv} = 28$, $\gamma = 0.00316$	$N_{sv} = 51$, $\gamma = 0.000316$
SIMCA	$Var_1 = 62\,\%$, $Var_2 = 60\,\%$	$Var_{1-9} \approx 60\,\%$	$Var_1 = 61\,\%$, $Var_2 = 67\,\%$, $Var_3 = 65\,\%$	$Var_{1-3} = 25\,\%$
PLS-DA	$Var = 73\,\%$	$Var = 72\,\%$	$Var = 74\,\%$	$Var = 37\,\%$

Результати класифікації модельних і тестових даних дозволяють характеризувати алгоритми нейронних мереж з навчанням як перспективні класифікаційні інструменти.

2.6. Оптимізація параметрів штучних нейронних мереж і стійкість класифікації з навчанням до варіювання вихідних даних (класифікація розчинників за їх сольватохромними характеристиками)

Оптимізацію параметрів штучних нейронних мереж та стійкість класифікації до варіювання вихідних даних розглянемо на прикладі класифікації розчинників за їх сольватохромними характеристиками.

Сольватохромні параметри Тафта–Камлета (параметр основності розчинників як акцепторів водневого зв'язку β [60], параметр кислотності розчинників як донорів водневого зв'язку α [61], параметр полярності і поляризовності розчинників π^* [62]) широко застосовуються в хімії. Регулярно створюються нові сольватохромні зонди для дослідження різних реакційних середовищ і поверхні матеріалів [63−66]. Реакційні середовища класифікують і характеризують на основі значень їх сольватохромних параметрів, використовуючи багатовимірні статистичні методи [67, 68]. Значення сольватохромних параметрів із часом уточнюються (див., наприклад, [69]). Окрім цього, в літературі зустрічаються випадки відсутності значень одного з соль-

ватохромних параметрів для деяких речовин. Наприклад, у роботі [70] для 11 зі 185 наведених органічних розчинників значення параметра π^* є невідомими, в роботі [71] представлено значення тільки параметра π^* для 229 розчинників, а в роботі [72] – значення параметрів α і π^* для 35 кислот. Тому класифікаційні алгоритми повинні не лише забезпечувати надійну ідентифікацію сполук за набором сольватохромних параметрів, але й бути стійкими як до невеликого варіювання значень сольватохромних параметрів, так і до наявності у вихідних даних пропусків.

Досліджуваний нами масив даних включав 56 розчинників, що характеризуються трьома сольватохромними параметрами α, β і π^* [73, 74] (див. Додаток, табл. Д3).

У роботі [73] запропоновано розбиття цих розчинників на 6 груп: 1) слабкоосновні розчинники (аліфатичні етери й заміщені аліфатичні аміни); 2) апротонні полярні розчинники (циклічні етери, кетони, естери і нітрили); сильнополярні й основні розчинники (піридини, аміди, сульфоксиди, сечовини, фосфороаміди); 4) відносно полярні розчинники (ароматичні сполуки, галогенопохідні і полігалогенопохідні аліфатичних вуглеводнів); 5) розчинники, здатні утворювати водневі зв'язки (спирти і вода); 6) аліфатичні вуглеводні.

Особливу складність обробці таких даних надає великий розмах значень сольватохромних параметрів усередині кожного класу. Наприклад, параметр π^* усередині класу 4 коливається від 0.28 до 0.82, параметр β усередині класу 5 – від 0.18 до 1.01, параметр α усередині класу 2 – від 0 до 0.22.

Таким чином, обробка масиву даних про 56 розчинників, що характеризуються трьома сольватохромними параметрами, є цікавою як з хімічної точки зору, так і з точки зору перевірки працездатності алгоритмів нейронних мереж на класах, що перекриваються, й оцінки стійкості алгоритмів до наявності в даних пропусків і невеликого варіювання чисельних значень характеристик (наявності в даних похибок) [75].

Оптимальний об'єм навчальної вибірки знаходили за допомогою ймовірнісної мережі, як і для розглянутих раніше модельних і тестових наборів даних. Процедуру визначення оптимального об'єму навчальної вибірки повторювали чотири рази для кожного значення

T, % (див. формулу (2.7)) з метою розроблення рекомендацій щодо формування репрезентативної навчальної вибірки.

Ненадійність класифікації розчинників оцінювали за формулою (2.6). На рис. 2.15 наведено середні значення ненадійності ідентифікації розчинників по чотирьох парах навчальних і тестових вибірок. Імовірнісна мережа правильно класифікує розчинники тестової вибірки, починаючи з $T = 60\,\%$. Це значення об'єму навчальної вибірки використовували при навчанні нейронних мереж й інших типів.

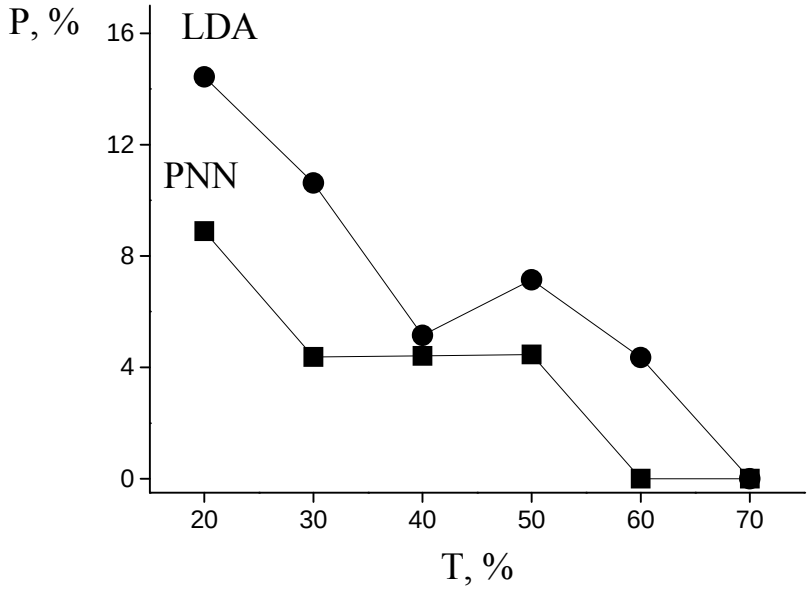

Рис. 2.15. Залежність значень частки розчинників тестової вибірки, що неправильно класифікуються, від об'єму навчальної вибірки для ймовірнісної мережі ($\delta = 0.1$) і лінійного дискримінантного аналізу

Традиційні алгоритми класифікації вимагають більшої кількості навчальних зразків. Наприклад, лінійний дискримінантний аналіз правильно класифікує розчинники тестової вибірки, починаючи з $T = 70\,\%$; при $T = 60\,\%$ дискримінантний аналіз правильно класифікує розчинники лише для однієї пари навчальної та тестової вибірок. Цю пару навчальної і тестової вибірок застосували і для тестування традиційних методів класифікації та для оцінки робастності алгоритмів.

Для кожного алгоритму нейронної мережі при випадково вибраній невеликій кількості прихованих нейронів (h=9) визначали метод навчання і комбінацію функцій активації, що забезпечують прийнятну (мінімальну з отриманих значень) ненадійність ідентифікації розчинників тестової вибірки (табл. 2.9–2.12). Потім, змінюючи кількість прихованих нейронів, досягали стовідсоткової надійності класифікації

розчинників (рис. 2.16). Оптимальна кількість прихованих нейронів для різних типів нейронних мереж знаходиться в межах $10 < h < 14$.

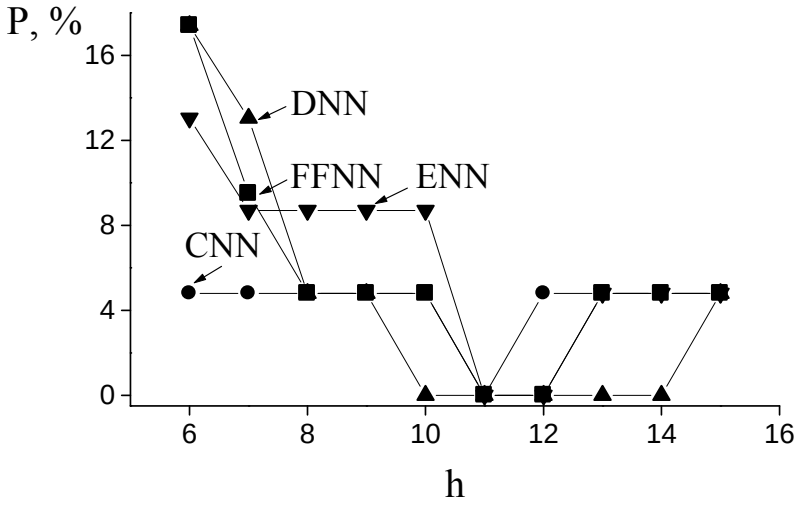

Рис. 2.16. Ненадійність класифікації розчинників тестової вибірки при різних кількостях прихованих нейронів для чотирьох типів нейронних мереж

Таблиця 2.9

Ненадійність класифікації розчинників (P, %) при різних функціях активації та методах навчання для мережі Елмана

Метод навчання	Функції активації для прихованого / вихідного шарів			
	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс	логістична / лінійна	логістична / логістична
Левенберга–Марквардта	35	30	48	52
Пауела–Бієле	57	44	52	39
Алгоритм зворотного поширення помилки	17	**9**	39	44

Таблиця 2.10

Ненадійність класифікації розчинників (P, %) при різних функціях активації та методах навчання для динамічної мережі

Метод навчання	Функції активації для прихованого / вихідного шарів			
	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс	логістична / лінійна	логістична / логістична
Левенберга–Марквардта	**5**	9	30	87
Пауела–Бієле	9	22	39	30

Продовження табл. 2.10

Алгоритм зворотного поширення помилки	35	83	83	61

Таблиця 2.11

Ненадійність класифікації розчинників (P, %) при різних функціях активації та методах навчання для каскадної мережі

Метод навчання	Функції активації для прихованого / вихідного шарів			
	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс	логістична / лінійна	логістична / логістична
Левенберга–Марквардта	9	**5**	9	9
Пауела–Бієле	9	35	9	17
Алгоритм зворотного поширення помилки	22	35	61	74

Таблиця 2.12

Ненадійність класифікації розчинників (P, %) при різних функціях активації та методах навчання для мережі прямої передачі сигналу

Метод навчання	Функції активації для прихованого / вихідного шарів			
	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс	логістична / лінійна	логістична / логістична
Левенберга–Марквардта	**5**	22	13	9
Пауела–Бієле	9	17	9	22
Алгоритм зворотного поширення помилки	30	57	48	30

Оптимальні параметри нейронних мереж, що забезпечують надійну класифікацію розчинників за значеннями їх сольватохромних параметрів, представлено в табл. 2.13.

Таблиця 2.13

Оптимальні параметри алгоритмів нейронних мереж для надійної класифікації розчинників

Параметр	Нейронна мережа			
	FFNN	CNN	DNN	ENN
Метод навчання	Левенберга–Марквардта	Левенберга–Марквардта	Левенберга–Марквардта	Алгоритм зворотного поширення помилки
Отримане значення *mse*	0.001	0.001	0.001	0.1
Кількість прихованих нейронів (мінімальна)	11	11	10	11
Функції активації	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс	гіперболічний тангенс / лінійна	гіперболічний тангенс / гіперболічний тангенс

Надійність класифікації розчинників за допомогою традиційних алгоритмів є низькою. У випадку PLS-DA і SIMCA досліджувані класи моделювали за допомогою двох головних компонент / прихованих змінних (табл. 2.14).

Таблиця 2.14

Значення параметрів та ненадійність класифікації розчинників деякими традиційними методами класифікації

Метод	SVM	SIMCA	PLS-DA
Значення параметрів	$N_{sv} = 24$, $\gamma = 0.00316$	$Var_1 = Var_3 = Var_5 = 100\ \%$, $Var_2 = 91\ \%$, $Var_4 = 83\ \%$,	$Var = 78\ \%$
P, %	13	26	13

Ефективність алгоритмів оцінювали за їх стійкістю до наявності в даних похибок, розподіл яких відрізняється від нормального, і пропусків. До початкових чисельних характеристик розчинників тестової вибірки вносили похибки відповідно до моделі «грубих промахів». Ступінь спотворення початкових значень параметрів внесеними «грубими промахами» оцінювали за значенням критерію

$$Q = \sum_{i}^{N} \left| x_i^{initial} - x_i^{changed} \right|, \qquad (2.14)$$

де N – кількість розчинників, $x_i^{initial}$ – початкові значення сольватохромних параметрів, $x_i^{changed}$ – значення сольватохромних параметрів після внесення похибок.

На рис. 2.17 наведено залежність критерію Q від частки внесених промахів q.

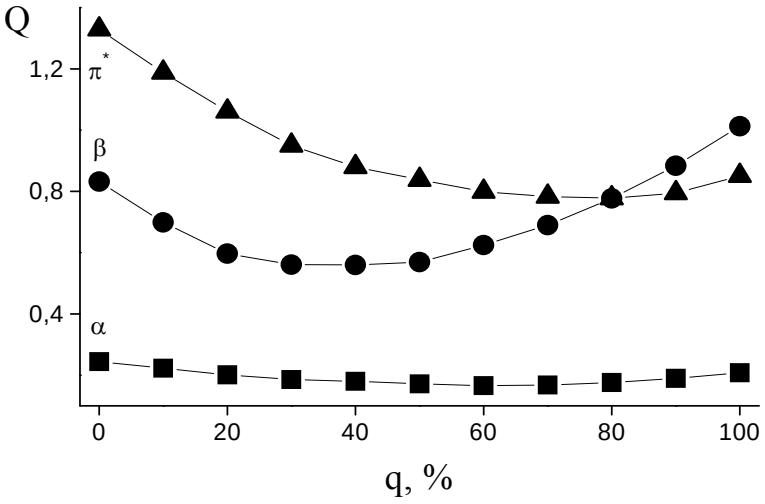

Рис. 2.17. Залежність критерію спотворення початкових значень параметрів від частки інтенсивності внесених «грубих промахів»

Для того щоб оцінити стійкість алгоритмів класифікації до наявності в даних пропусків, з масиву характеристик розчинників тестової вибірки випадковим чином видаляли 6 і 13 значень, що відповідає приблизно 10 і 20 % характеристик. Пропуски заповнювали середніми значеннями відповідного параметра по значеннях, що залишилися (табл. 2.15, нові значення характеристик виділено жирним шрифтом).

Таблиця 2.15

Початкові і змінені внесенням пропусків значення сольватохромних характеристик

Значення сольватохромних характеристик								
Початкові			Після внесення 10 % пропусків			Після внесення 20 % пропусків		
α	β	π*	α	β	π*	α	β	π*
0.00	0.46	0.24	0.00	0.46	0.24	**0.14**	0.46	0.24
0.00	0.47	0.27	0.00	0.47	0.27	0.00	0.47	0.27
0.00	0.22	0.73	0.00	0.22	0.73	0.00	**0.37**	0.73
0.00	0.20	0.69	0.00	0.20	0.69	0.00	0.20	0.69

Продовження табл. 2.15

0.08	0.48	0.71	0.08	0.48	0.71	**0.14**	0.48	0.71
0.00	0.45	0.55	0.00	0.45	0.55	0.00	**0.37**	**0.49**
0.00	0.76	0.88	0.00	0.76	0.88	0.00	0.76	0.88
0.00	0.69	0.88	0.00	0.69	**0.48**	**0.14**	0.69	0.88
0.00	0.77	0.92	0.00	0.77	0.92	0.00	0.77	0.92
0.00	0.80	0.83	**0.11**	0.80	0.83	0.00	0.80	**0.49**
0.00	0.71	0.14	0.00	0.71	0.14	0.00	0.71	0.14
0.00	0.39	1.01	0.00	0.39	1.01	**0.14**	0.39	1.01
0.00	0.64	0.92	0.00	0.64	0.92	**0.14**	0.64	0.92
0.68	1.01	0.41	0.68	1.01	0.41	0.68	**0.37**	0.41
0.79	0.88	0.47	**0.11**	0.88	0.47	0.79	0.88	0.47
0.83	0.77	0.54	0.83	**0.46**	0.54	0.83	0.77	0.54
0.90	0.52	0.92	0.90	0.52	0.92	0.90	0.52	0.92
0.00	0.00	-0.02	0.00	0.00	-0.02	0.00	0.00	-0.02
0.00	0.00	0.03	0.00	0.00	0.03	**0.14**	0.00	0.03
0.00	0.00	-0.04	0.00	0.00	-0.04	0.00	0.00	-0.04
0.00	0.00	0.00	0.00	0.00	**0.48**	0.00	0.00	0.00
0.00	0.53	0.76	0.00	0.53	0.76	0.00	**0.37**	0.76
0.00	0.49	0.87	0.00	0.49	**0.48**	0.00	**0.37**	0.87

Таблиця 2.16

Ненадійність алгоритмів класифікації (*P*, %) при обробці даних, в які було внесено «грубі промахи»

q, %	0	10	20	30	40	50	60	70	80	90	100
P_{PNN}	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
P_{LDA}	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	8.7	8.7
P_{DNN}	**4.3**	**0**	**0**	**8.7**	**4.3**	**4.3**	**8.7**	**0**	**4.3**	**4.3**	**4.3**
P_{CNN}	8.7	4.3	4.3	8.7	4.3	4.3	4.3	0	8.7	4.3	8.7
P_{FFN}	13	4.3	13	8.7	4.3	8.7	8.7	8.7	8.7	8.7	8.7
P_{ELMN}	0	0	4.3	8.7	0	8.7	0	8.7	4.3	8.7	8.7
P_{PLS-DA}	13	17	17	17	22	17	17	22	22	22	22
P_{SVM}	26	26	22	17	17	17	22	22	22	22	22
P_{SIMCA}	57	57	61	61	61	61	61	61	61	61	61

Значення сольватохромних параметрів сильно змінювалися як після внесення похибок за моделлю «грубих промахів», так і після внесення пропусків. При оцінці стійкості алгоритмів нейронних мереж для їх навчання використовували параметри з табл. 2.13. Ненадійність класифікації розчинників при обробці даних за наявності пропусків і похибок представлено в табл. 2.16 і 2.17.

Таблиця 2.17

Ненадійність алгоритмів класифікації (*P*, %) при обробці даних, в які було внесено пропуски

Кількість пропусків	6	13
P_{PNN}, P_{DNN}, P_{CNN}	**4.3**	**13.0**
P_{LDA}, P_{ELMN}	8.7	17.4
P_{FFN}	8.7	26.1
P_{PLS-DA}	30	30
P_{SVM}	26	26
P_{SIMCA}	70	65

Стійкість алгоритмів класифікації до наявності в даних похибок знижується в такому порядку:

$$PNN > DNN > ELMN \approx LDA \approx CNN > FFN > PLS-DA > SVM > SIMCA;$$

стійкість до наявності пропусків – у такому порядку:

$$CNN \approx PNN \approx DNN > LDA \approx ELMN > FFN > SVM > PLS-DA > SIMCA.$$

***

Таким чином, з'ясовано, що для надійної класифікації органічних розчинників, що характеризуються трьома сольватохромними параметрами, оптимальна кількість прихованих нейронів для різних типів нейронних мереж знаходиться в межах $10 < h < 14$.

В якості функцій активації доцільно використовувати гіперболічний тангенс і лінійну функцію спільно з алгоритмом навчання Левенберга–Марквардта.

Алгоритми нейронних мереж стійкіші до наявності в даних промахів і пропусків, ніж традиційні методи класифікації.

Найвищою стійкістю до наявності у вихідних даних промахів володіє ймовірнісна нейронна мережа

Література до глави 2

1. Agatonovic-Kustrin S. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research / S. Agatonovic-Kustrin, R. Beresford // J. of Pharm. and Biomed. Analysis. – 2000. – V. 22. – P. 717-727.
2. Arbib M. A. The handbook of brain theory and neural networks, 2nd edn. / M. A. Arbib. – Cambridge, e.a. : Bradford Book, 2002. – 1290 p.
3. Gupta M. M. Static and dynamic neural network: from fundamentals to advanced theory / M. M. Gupta, L. Jin, N. Homma. – Wiley, 2003. – 722 p.
4. Braspenning P. J. Artificial neural networks. An introduction to ANN theory and practice / P. J. Braspenning, F. Thuijsman, A. J. M. M. Weijters. – Berlin : Springer-Verlag, 1995. – 136 p.
5. Хайкин С. Нейронные сети: полный курс / С. Хайкин; пер. с англ. Н. Н. Куссуль, А. Ю. Шелестова. – 2-е изд., испр. – М. : Вильямс, 2006. – 1104 с.
6. Уоссермен Ф. Нейрокомпьютерная техника: теория и практика / Ф. Уоссермен; пер. с англ. Ю. А. Зуева, В. А. Точенова. – М. : Мир, 1992. – 184 с.
7. Заенцев И. В. Нейронные сети: основные модели / И. В. Заенцев. – Воронеж : Воронежский гос. ун-т, 1999. – 76 с. [Электронный ресурс]. – Режим доступа : http://nncourse.chat.ru/course.pdf
8. Дьяконов В. П. Matlab 6.5 SP1/7 SP2 + Simulink 5/6. Инструменты искусственного интеллекта и биоинформатики / В. П. Дьяконов, В. В. Круглов. – М. : СОЛОН-ПРЕСС, 2006. – 456 с.
9. Круглов В. В. Искусственные нейронные сети. Теория и практика / В. В. Круглов, В. В. Борисов. – 2-е изд. – М. : Горячая линия-Телеком, 2002. – 382 с.
10. Осовский С. Нейронные сети для обработки информации / С. Осовский; пер. с польск. И. Д. Рудинского. – М. : Финансы и статистика, 2002. – 344 с.
11. Каллан Р. Основные концепции нейронных сетей / Р. Каллан; пер. с англ. А. Г. Сивака. – М. : Вильямс, 2001. – 287 с.
12. Комарцова Л. Г. Нейрокомпьютеры / Л. Г. Комарцова, А. В. Максимов. – 2-е изд., перераб. и доп. – М. : Изд-во МГТУ имени Н. Э. Баумана, 2004. – 400 с.
13. Медведев В. С. Нейронные сети. MATLAB 6 / В. С. Медведев, В. Г. Потемкин. – М. : ДИАЛОГ-МИФИ, 2002. – 496 с.
14. Баскин И. И. Моделирование свойств химических соединений с использованием искусственных нейронных сетей и фрагментарных дескрипторов : дис…доктора физ.-мат. наук: 02.00.17 / Баскин Игорь Иосифович. – М. : Московский гос. ун-т им. М. В. Ломоносова, 2009. – 365 с.

15. Basheer I. A. Artificial neural networks: fundamentals, computing, design, and application / I. A. Basheer, M. Hajmeer // J. Microbiol. Meth. – 2000. – Vol. 43, No 1. – P. 3-31.
16. Marini F. Artificial neural networks in chemometrics: history, examples and perspectives / F. Marini, R. Bucci, A. L. Magri, A. D. Magri // Microchem. J. – 2008. – Vol. 88. – P. 178-185.
17. Akyol D. E. A review on evolution of producing scheduling with neural networks / D. E. Akyol, G. M. Bayhan // Comput. Ind. Eng. – 2007. – Vol. 53, No 1. – P. 95-122.
18. Niculescu S. P. Artificial neural networks and genetic algorithms in QSAR / S. P. Niculescu // J. Mol. Struct.: THEOCHEM. – 2003. – Vol. 622, No 1–2. – P. 71-83.
19. Jalali-Heravi M. QSAR study of heparanase inhibitors activity using artificial neural networks and Levenberg-Marquardt algorithm / M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah // Eur. J. Med. Chem. – 2008. – Vol. 43, No 3. – P. 548-556.
20. Stojic N. Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks / N. Stojic, S. Eric, I. Kuzmanovski // J. Mol. Graph. Model. – 2010. – Vol. 29, No 3. – P. 450-460.
21. Jalali-Heravi M. Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes / M. Jalali-Heravi, M. H. Fatemi // J. Chromatogr. A. – 2001. – Vol. 915, No 1-2. – P. 177-183.
22. Srecnik G. Optimization of artificial neural networks used for retention modeling in ion chromatography / G. Srecnik, Z. Debeliak, S. Cerjan-Stefanovic, M. Novic, T. Bolanca // J. Chromatogr. A. – 2002. – Vol. 973, No 1-2. – P. 47-59.
23. Zhao R. H. Application of an artificial neural network in chromatography – retention behavior prediction and pattern recognition / R. H. Zhao, B. F. Yue, J. Y. Ni, H. F. Zhou, Y. K. Zhang // Chemometr. Intell. Lab. – 1999. – Vol. 45, No 1-2. – P. 163-170.
24. Jalali-Heravi M. Use of self-training artificial neural networks in modeling of gas chromatography relative retention times of a variety of organic compounds / M. Jalali-Heravi, Z. Garakani-Nejad // J. Chromatogr. A. – 2002. – Vol. 945, No 1-2. – P. 173-184.
25. Fatemi M. H. Quantitative structure-property relationship studies of migration index in microemulsion electrokinetic chromatography using artificial neural networks / M. H. Fatemi // J. Chromatogr. A. – 2003. – Vol. 1102, No 1-2. – P. 221-229.
26. Hernandez-Caraballo E. A. Increasing the working calibration range by means of artificial neural networks for the determination of cadmium by graphite furnace atomic absorption spectrometry / E. A. Hernandez-

Caraballo, R. M. Avila-Gomez, F. Rivas, M. Burguera, J. L. Burguera // Talanta. – 2004. – Vol. 63, No 2. – P. 425-431.
27. Vander Heyden Y. The application of Kohonen neural networks to diagnose calibration problems in atomic absorption spectrometry / Y. Vander Heyden, P. Vankeerberghen, M. Novic, J. Zupan, D. L. Massart // Talanta. – 2000. – Vol. 51, No 3. – P. 455-466.
28. Masoum S. Discrimination of wines based on 2D NMR spectra using learning vector quantization neural networks and partial least squares discriminant analysis / S. Masoum, D. J.-R. Bouveresse, J. Vercauteren, M. Jalali-Heravi, D. N. Rutledge // Anal. Chim. Acta. – 2006. – Vol. 558, No 1–2. – P. 144-149.
29. Aires-de-Sousa J. Prediction of ^{1}H NMR chemical shifts using neural networks / J. Aires-de-Sousa, M. C. Hemmer, J. Gasteiger // Anal. Chem. – 2002. – Vol. 74, No 1. – P. 80-90.
30. Meiler J. Using neural network for ^{13}C NMR chemical shift prediction – comparison with traditional methods / J. Meiler, W. Maier, M. Will, R. Meusinger // J. Magn. Resonance. – 2002. – Vol. 157, No 2. – P. 242-252.
31. Cleva C. Advantages of a hierarchical system of neural-networks for the interpretation of infrared spectra in structure determination / C. Cleva, C. Cachet, D. Cabrol-Bass, T. P. Forrest // Anal. Chim. Acta. – 1997. – Vol. 348. – P. 255-265.
32. Bell S. Classification of ion mobility spectra by functional groups using neural networks / S. Bell, E. Nazarov, Y. F. Wang, G. A. Eiceman // Anal. Chim. Acta. – 1999. – Vol. 394, No 2-3. – P. 121-133.
33. Andrews J. M. Neural network approach to qualitative identification of fuels and oils from induced fluorescence spectra / J. M. Andrews, S. H. Lieberman // Anal. Chim. Acta. – 1994. – Vol. 285, No 1–2. – P. 237-246.
34. Yoshida E. Application of neural networks for the analysis of gamma-ray spectra measured with a Ge spectrometry / E. Yoshida, K. Shizuma, S. Endo, T. Oka // Nucl. Instrum. Meth. A. – 2002. – Vol. 484, No 1-3. – P. 557-563.
35. Belic I. Neural network methodologies for mass spectra recognition / I. Belic, L. Gyergyek // Vacuum. – 1997. – Vol. 48, No 7-9. – P. 633-637.
36. Jalali-Heravi M. Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network / M. Jalali-Heravi, M. H. Fatemi // Anal. Chim. Acta. – 2000. – Vol. 415, No 1–2. – P. 95-103.
37. Pulido A. Radial basis functions applied to the classification of UV-visible spectra / A. Pulido, I. Ruisanchez, F. X. Rius // Anal. Chim. Acta. – 1999. – Vol. 388, No 3. – P. 273-281.
38. Reshetnikova V. N. Application of artificial neural networks to predicttions in flow-injection spectrophotometry / V. N. Reshetnikova, V. V. Kuznetsov, S. S. Borodulin // J. of Analyt. Chem. − 2016. − Vol. 71, No. 3. − P. 243-247.

39. Mittermayr S. Mobility modeling of peptides in capillary electrophoresis / S. Mittermayr, M. Olajos, T. Chovan // Trends Anal. Chem. – 2008. – Vol. 27, No 5. – P. 407-417.
40. Jalali-Heravi M. Artificial neural network modeling of peptide mobility and peptide mapping in capillary zone electrophoresis / M. Jalali-Heravi, Y. Shen, M. Hassanisadi, M. G. Khaledi // J. Chromatogr. A. – 2005. – Vol. 1096, No 1-2. – P. 58-68.
41. Mishra P. Elman RNN based classification of proteins sequences on account of their mutual information / P. Mishra, P. N. Pandey // J. Theor. Biol. – 2012. – Vol. 311. – P. 40-45.
42. Smits J. R. M. Using artificial neural networks for solving chemical problems. Part I. Multi-layer feed-forward networks / J. R. M. Smits, W. J. Melssen, L. M. C. Buydens, G. Kateman // Chemometr. Intell. Lab. – 1994. – Vol. 22. – P. 165-189.
43. Melssen W. J. Using artificial neural networks for solving chemical problems. Part II. Kohonen self-organizing feature maps and Hopfield networks / W. J. Melssen, J. R. M. Smits, L. M. C. Buydens, G. Kateman // Chemometr. Intell. Lab. – 1994. – Vol. 23. – P. 267-291.
44. Zupan J. Neural Networks: a new method for solving chemical problems or just a passing phase? / J. Zupan, J. Gasteiger // Anal. Chim. Acta. – 1991. – Vol. 248. – P. 1-30.
45. Kateman G. Neural networks in analytical chemistry / G. Kateman // Chemometr. Intell. Lab. – 1993. – Vol. 19. – P. 135-142.
46. Cano J. R. On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining / J. R. Cano, F. Herrera, M. Lozano // Appl. Soft Comput. – 2006. – Vol. 6. – P. 323-332.
47. Analysis of new techniques to obtain quality training sets / J. S. Sanchez, R. Barandela, A. I. Marques, Alejo R., J. Badenas // Pattern Recogn. Lett. – 2003. – Vol. 24. – P. 1015-1022.
48. Carpenter S. E. Selection of optimum training sets for use in pattern recognition analysis of chemical data / S. E. Scott, G. W. Small // Anal. Chim. Acta. – 1991. – Vol. 249. – P. 305-321.
49. Nguyen D. Improving the learning speed of 2-layer neural networks by chhosing initial values of the adaptive weights / D. Nguyen, B. Widrow // Int. Joint Conf. on Neural Networks, June 1990. : Abstr. – San Diego, USA, 1990. – P. 21.
50. Руководство ЕВРАХИМ / СИТАК «Количественное описание неопределенности в аналитических измерениях» : пер. с англ. под ред. Л. А. Конопелько. – СПб. : ВНИИМ им. Д. И. Менделеева, 2002. – 149 с.
51. Айвазян С. А. Прикладная статистика: основы моделирования и первичная обработка данных. Справочное изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1983. – 471 с.

52. Вучков И. Прикладной линейный регрессионный анализ / И. Вучков, Л. Бояджиева, Е. Солаков; пер. с болгарского Ю. П. Адлера. – М. : Финансы и статистика, 1987. – 239 с.
53. Холин Ю. В. Количественный физико-химический анализ комплексообразования в растворах и на поверхности химически модифицированных кремнеземов: содержательные модели, математические методы и их приложения : монография / Ю. В. Холин. – Харьков : Фолио, 2000. – 288 с.
54. Iris Data Set (1988). UCI Machine Learning Repository. [Electronic Resource] – Way of access : http://archive. ics.uci.edu/ml/datasets/Iris
55. Wine Data Set (1991). UCI Machine Learning Repository. [Electronic Resource] – Way of access : http://archive. ics.uci.edu/ml/datasets/wine
56. Debska B. Application of artificial neural network in food classification / B. Debska, B. Guzowska-Swider // Anal. Chim. Acta. – 2011. – Vol. 705. – P. 283-291.
57. Шараф М. А. Хемометрика: пер. с англ. / М. А. Шараф, Д. Л. Иллмэн, Б. Р. Ковальски. – Ленинград : Химия, 1989. – 272 с.
58. Краснянчин Я. Н. Надежность идентификации аналитов с помощью искусственных нейронных сетей / Я. Н. Краснянчин, А. В. Пантелеймонов, Ю. В. Холин // Вісн. Харк. нац. ун-ту. – 2010. – No 895. Хімія. Вип. 18 (41). – С. 39-46.
59. Краснянчин Я. Н. Некоторые аспекты параметризации искусственных нейронных сетей в задачах качественного химического анализа / Я. Н. Краснянчин, А. В. Пантелеймонов, Ю. В. Холин // Вісн. Харк. нац. ун-ту. – 2010. – No 932. Хімія. Вип. 19 (42). – С. 170-181.
60. Kamlet M. J. The solvatochromic comparison method. I. The β-scale of solvent hydrogen-bond acceptor (HBA) basicities / M. J. Kamlet, R. W. Taft // J. Am. Chem. Soc. – 1976. – Vol. 98, No 2. – P. 377-383.
61. Kamlet M. J. The solvatochromic comparison method. 2. The α-scale of solvent hydrogen-bond donor (HBD) acidities / M. J. Kamlet, R. W. Taft // J. Am. Chem. Soc. – 1976. – Vol. 98, No 10. – P. 2886-2894.
62. Kamlet M. J. The solvatochromic comparison method. 6. The π^* scale of solvent polarities / M. J. Kamlet, J. L. Abboud, R. W. Taft // J. Am. Chem. Soc. – 1977. – Vol. 99, No 18. – P. 6027-6038.
63. Spange S. Probing surface basicity of solid acids with an aminobenzodifurandione dye as the solvatochromic probe / S. Spange, S. Prause, E. Vilsmeier, W. R. Thiel // J. Phys. Chem. B. – 2005. – Vol. 109. – P. 7280-7289.
64. Fidale L. C. Probing the dependence of the properties of cellulose acetates and their films on the degree of biopolymer substitution: use of solvatochromic indicators and thermal analysis / L. C. Fidale, C. Ibbrucker, P. L. Silva, C. M. Lucheti, T. Heinze, O. A. El Seoud // Cellulose. – 2010. – Vol. 17. – P. 937-951.

65. Luchetti L. Kinetic and spectroscopic investigations of aqueous micelles of cationic surfactants / L. Luchetti // Centr. Eur. J. Chem. – 2010. – Vol. 8, No 6. – P. 1318-1322.
66. Silva P. L. Solvation in pure liquids: what can be learned from the use of pairs of indicators / P. L. Silva, P. A. R. Pires, M. A. S. Trassi, O. A. El Seoud // J. Phys. Chem. B. – 2008. – Vol. 112. – P. 14976-14984.
67. Katritzky A. R. The classification of solvents by combining classical QSPR methodology with principal component analysis / A. R. Katritzky, D. C. Fara, M. Kuanar, E. Hur, M. Karelson // J. Phys. Chem. A. – 2005. – Vol. 109, No 45 – P. 10323-10341.
68. Gramatica P. Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of volecular descriptors / P. Gramatica, N. Navas, R. Todeschini // Trends Anal. Chem. – 1999. – Vol. 18, No 7. – P. 461-471.
69. Reichardt C. Solvatochromic dyes as solvent polarity indicators / C. Reichardt // Chem. Rev. – 1994. – Vol. 94. – P. 2319-2358.
70. Marcus Y. The properties of organic liquids that are relevant to their use as solvating solvents / Y. Marcus // Chem. Soc. Rev. – 1993. – Vol. 22. – P. 409-416.
71. Laurence C. The empirical treatment of solvent-solute interactions: 15 years of π^* / C. Laurence, P. Nicolet, M. T. Dalati, J.-L. M. Abboud, R. Notari // J. Phys. Chem. – 1999. – Vol. 98. – P. 5807-5816.
72. Spange S. Probing the surface polarity of various silicas and other moderately strong solid acids by means of different genuine solvatochromic dyes / S. Spange, E. Vilsmeier, Y. Zimmermann // J. Phys. Chem. B. – 2000. – Vol. 104. – P. 6417-6428.
73. de Juan A. Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach / A. de Juan, G. Fonrodona, E. Casassas // Trends Anal. Chem. – 1997. – Vol. 16, No 1. – P. 52-62.
74. Snyder L. R. Classification of the solvent properties of common liquids / L. R. Snyder // J. Chromatogr. – 1974. – Vol. 92. – P. 223-230.
75. Pushkarova Y. The classification of solvents based on solvatochromic characteristics: the choice of optimal parameters for artificial neural networks / Y. Pushkarova, Y. Kholin // Centr. Eur. J. Chem. – 2012. – Vol. 10, No 4. – P. 1318-1327.

Глава 3

ІДЕНТИФІКАЦІЯ ГЕОГРАФІЧНОГО ПОХОДЖЕННЯ З ВИКОРИСТАННЯМ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

3.1. Ідентифікація зразків річкових і джерельних вод

Одним з інструментів, необхідних для розв'язання комплексної проблеми якості води, є якісний хімічний аналіз, особливо ідентифікація зразків вод за їх складом, походженням та властивостями. Для розв'язання цієї задачі все частіше залучають хемометричні методи (метод головних компонент, факторний, кластерний і дискримінантний аналізи, штучні нейронні мережі та інші) [1−7].

Ми використали апарат нейронних мереж для встановлення походження (віднесення до того або іншого заздалегідь вибраного класу) зразків річкових (забруднених промисловими стоками) і джерельних вод м. Харкова на основі даних про вміст важких і перехідних металів [8].

Аналізований масив даних включав 22 зразки річкових вод і 24 зразки джерельних вод із різних річок і джерел м. Харкова, відібраних у різні сезони впродовж 2008–2010 років. У зразках вод визначали концентрації купруму, цинку, плюмбуму, кадмію, мангану, феруму, кобальту і нікелю. Табл. 3.1–3.3 містять середні значення концентрацій металів у зразках вод, визначені за п'ятьма паралельними вимірюваннями [8].

Таблиця 3.1

Концентрації металів у зразках джерельних вод (мг/л)

Джерело, рік відбору та аналізу проби	Метал							
	Zn	Cu	Mn	Fe	Cd	Pb	Co	Ni
Саржин Яр «Харківська-1», 2010 р.	0.009	0.010	0.030	0.048	0.0024	0.025	0.0084	0.008
	0.013	0.008	0.024	0.072	0.0024	0.028	0.0063	0.009
	0.012	0.008	0.026	0.050	0.0018	0.025	0.0078	0.009

Таблиця 3.2

Концентрації металів у зразках річкових вод (мг/л)

Річка, рік відбору і аналізу проби	Метал							
	Zn	Cu	Mn	Fe	Cd	Pb	Co	Ni
Немишля 2008 р.	0.015	0.0055	0.023	0.100	0.0021	0.035	0.0084	0.020
	0.013	0.0045	0.041	0.120	0.0015	0.044	0.0088	0.018
	0.008	0.0073	0.017	0.028	0.0024	0.035	0.0094	0.014
	0.012	0.0073	0.013	0.038	0.0022	0.047	0.0063	0.013
	0.012	0.0078	0.433	0.019	0.0022	0.046	0.0140	0.020
Харків 2009 р.	0.005	0.0034	0.002	0.022	0.0017	0.041	0.0105	0.016
	0.104	0.0103	0.007	0.019	0.0019	0.041	0.0112	0.012
	0.015	0.0043	0.003	0.017	0.0017	0.034	0.0090	0.008
	0.007	0.0052	0.003	0.016	0.0039	0.031	0.0085	0.010
	0.010	0.0052	0.003	0.017	0.0030	0.023	0.0097	0.007
Лопань 2009 р.	0.047	0.0043	0.005	0.011	0.0032	0.031	0.0070	0.012
	0.109	0.0069	0.003	0.017	0.0052	0.043	0.0068	0.014
	0.065	0.0069	0.133	0.022	0.0060	0.038	0.0078	0.013
	0.010	0.0078	0.006	0.017	0.0030	0.028	0.0075	0.019
	0.011	0.0087	0.040	0.023	0.0023	0.038	0.0107	0.016
Уди 2010 р.	0.012	0.0056	0.279	0.069	н/в	н/в	0.0125	0.011
	0.017	0.0056	0.003	0.146	н/в	н/в	н/в	0.004
	0.005	0.0032	н/в	0.011	н/в	н/в	0.0071	0.004
	0.030	0.0286	0.015	0.097	0.0090	0.200	0.0554	0.030
	0.009	0.0063	0.023	0.046	н/в	0.038	0.0089	0.039
	0.008	0.0110	0.013	0.043	0.0021	0.010	0.0047	0.007
	0.011	0.0080	0.028	0.046	0.0015	0.025	0.0078	0.009

Таблиця 3.3

Концентрації металів у зразках річкових вод, для яких не визначені концентрації нікелю (2011 р.)

Річка	Метал						
	Zn	Cu	Mn	Fe	Cd	Pb	Co
Немишля	0.052	0.033	0.096	0.213	0.0035	0.025	0.005
Харків	0.034	0.030	0.066	0.180	0.0035	0.016	0.007
Лопань	0.036	0.034	0.037	0.160	0.0035	0.024	0.008
Уди	0.042	0.033	0.047	0.069	0.0031	0.009	0.003

Для кожної річки і джерела кількість відібраних проб (об'ємом по 500 мл) коливалася від 2 до 7. Час транспортування і зберігання проб від моменту їх відбору до обробки не перевищував доби. Вміст металів у зразках вод визначали методом атомно-абсорбційної спектроскопії [8–12] в полум'ї пропан-бутан-повітря і ацетилен-повітря. Відносні стандартні відхилення визначення концентрацій (сумарні відносні невизначеності) не перевищували 3 % [12]. Правильність методик перевірили за методом «введено-знайдено».

Табл. 3.4 містить значення гранично допустимих концентрацій металів у питній воді [13].

Таблиця 3.4

Гранично допустимі концентрації металів у питній воді (мг/л)

Метал	Нормативи для питної води		
	водопровідної	з колодязів і каптажів джерел	фасованої з пунктів розливу і бюветів
Цинк	1.0	не визначається	1.0
Купрум	1.0	не визначається	1.0
Манган	0.05	0.5	0.05
Ферум	0.2	1.0	0.2
Кадмій	0.001	не визначається	0.001
Плюмбум	0.010	не визначається	0.010
Кобальт	0.1	не визначається	0.1
Нікель	0.02	не визначається	0.02

Оскільки концентрації металів у зразках вод варіювалися в діапазоні від тисячних до десятих мг/л, перед застосуванням алгоритмів класифікації провели автомасштабне перетворення початкових даних за формулою (2.12).

Обробка масивів експериментальних даних ускладнюється наявністю пропусків: для трьох зразків джерельних вод не визначено вміст плюмбуму та кобальту; для чотирьох зразків річкових вод – вміст мангану, кадмію, плюмбуму і кобальту (концентрації металів нижчі за межі виявлення). Перед застосуванням класифікаційних алгоритмів пропуски заповнили нулями.

На рис. 3.1 представлено залежності частки неправильно класифікованих зразків вод тестової вибірки від об'єму навчальної вибірки для річкових і джерельних вод для ймовірнісної мережі. Ймовірнісна мережа правильно ідентифікує зразки вод тестової вибірки, починаючи

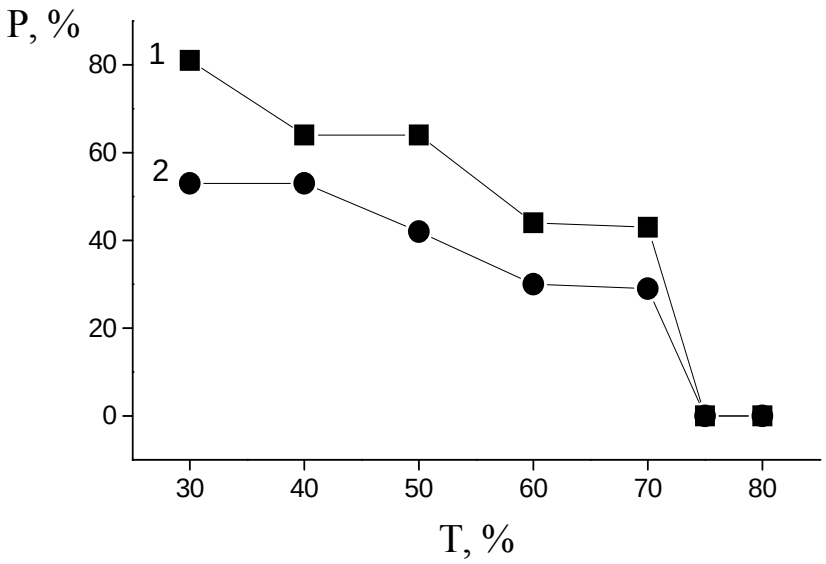

Рис. 3.1. Залежність часток неправильно класифікованих зразків вод тестової вибірки від об'єму навчальної вибірки для ймовірнісної нейронної мережі: 1 – зразки річкових вод, 2 – зразки джерельних вод

з $T = 75\,\%$, що еквівалентно 16 і 18 зразкам річкових і джерельних вод, відповідно. Це значення об'єму навчальної вибірки використовували при навчанні нейронних мереж й інших типів.

Як метод навчання для каскадної мережі, мережі з прямим поширенням сигналу і динамічної мережі використовували алгоритм Левенберга–Марквардта; для мережі Елмана – алгоритм зворотного поширення помилки (при виборі методу навчання керувалися встановленою оптимальною архітектурою мереж для класифікації модельних / тестових даних і класифікації розчинників).

Параметри навчання і тестування реалізованих нейронних мереж, а також ненадійність ідентифікації зразків вод представлено в табл. 3.5. Для реалізованих нейронних мереж визначено оптимальне число нейронів прихованого шару й оптимальну комбінацію функцій активації для прихованого і вихідного шарів.

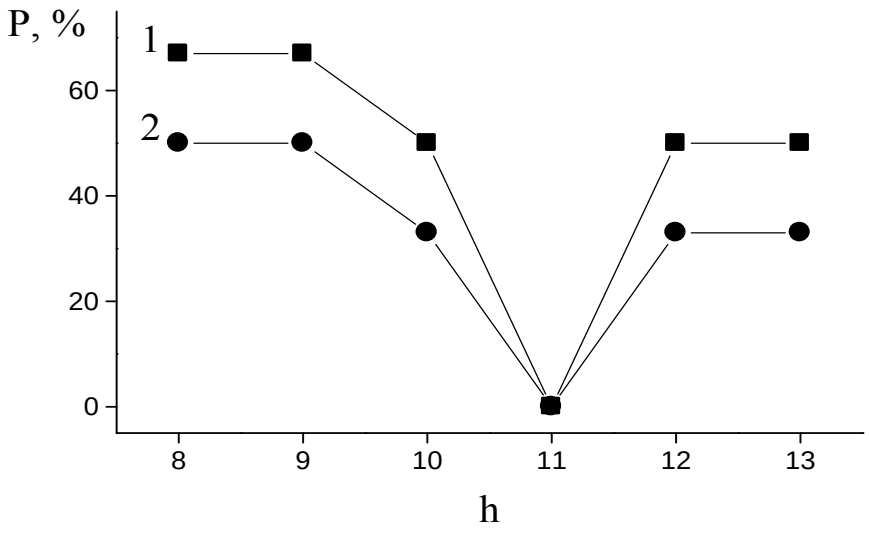

Рис. 3.2. Залежність частки неправильно класифікованих зразків вод від кількості нейронів прихованого шару для динамічної мережі: 1 – зразки річкових вод, 2 – зразки джерельних вод

Рис. 3.2–3.5 ілюструють залежності часток неправильно класифікованих зразків вод від кількості нейронів прихованого шару для алгоритмів нейромереж. Використання саме цих параметрів забезпечувало коректне навчання мережі і задовільну ідентифікацію зразків вод тестової вибірки (окрім мережі Елмана).

При використанні для ідентифікації зразків вод лінійного дискримінантного аналізу, методу опорних векторів, формального незалежного моделювання аналогій класів і дискримінантного аналізу за допомогою регресії на латентні структури ненадійність P є неприйнятно високою. В табл. 3.6 наведено мінімальні значення P, отримані при моделюванні класів за допомогою однієї головної компоненти в методі SIMCA для зразків річкових і джерельних вод та за допомогою трьох і однієї головних компонент в методі PLS-DA для зразків річкових і джерельних вод, відповідно.

Таблиця 3.5

Параметри нейронних мереж і ненадійність ідентифікації зразків вод

Нейронна мережа	Кількість прихованих нейронів для ідентифікації зразків річкових / джерельних вод	Функції активації для прихованого / вихідного шарів	P, % для зразків річкових / джерельних вод
Каскадна	9 / 10	гіперболічний тангенс / гіперболічний тангенс	17 / 0
З прямим поширенням сигналу	9 / 11	гіперболічний тангенс / лінійна	17 / 0
Елмана	11 / 12	гіперболічний тангенс / лінійна	17 / 17
Динамічна	11 / 11	гіперболічний тангенс / лінійна	0 / 0
Ймовірнісна	16 / 18	радіальна базисна / конкуруючий шар	0 / 0

Таким чином, одержані результати свідчать, що ідентифікацію зразків вод за даними про вміст мікроелементів доцільно здійснювати із застосуванням імовірнісної та динамічної нейронних мереж.

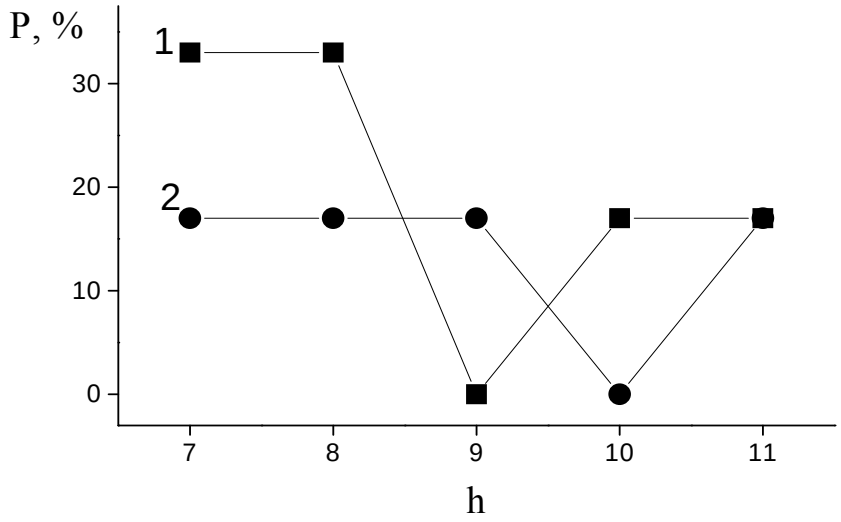

Рис. 3.3. Залежність частки неправильно класифікованих зразків вод від кількості нейронів прихованого шару для каскадної мережі: 1 – зразки річкових вод, 2 – зразки джерельних вод

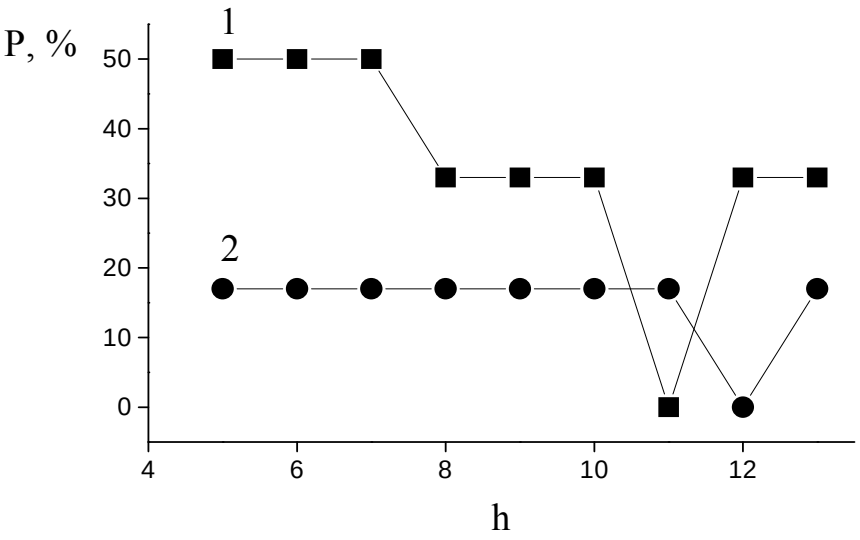

Рис. 3.4. Залежність частки неправильно класифікованих зразків вод від кількості нейронів для мережі Елмана: 1 – зразки річкових вод, 2 – зразки джерельних вод

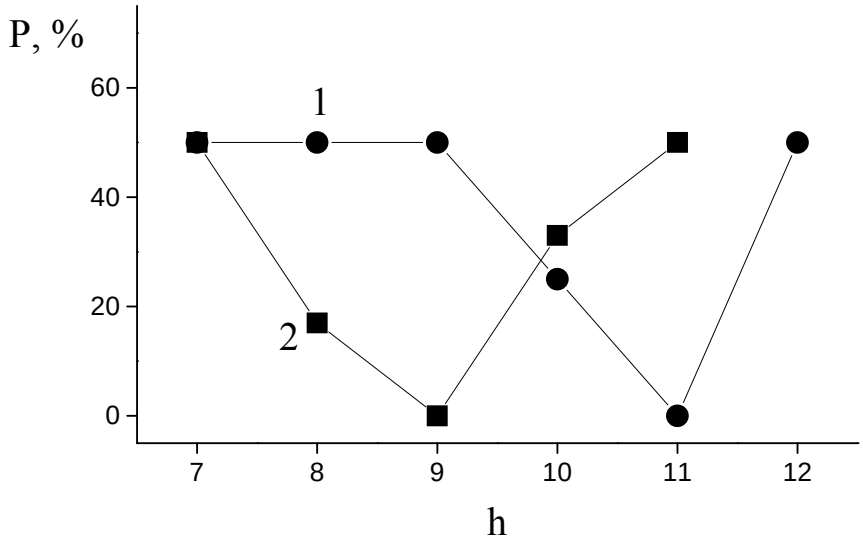

Рис. 3.5. Залежність частки неправильно класифікованих зразків вод від кількості нейронів для мережі прямої передачі сигналу: 1 – зразки річкових вод, 2 – зразки джерельних вод

Для перевірки правильності класифікації використали незалежну контрольну вибірку, яка містила чотири зразки річкових вод, відібраних у 2011 році (табл. 3.3). В цих зразках були визначені концентрації тих самих металів (окрім нікелю), що і в зразках, отриманих у 2008–2010 роках.

Таблиця 3.6

Результати ідентифікації зразків вод за допомогою традиційних алгоритмів класифікації

Метод	LDA	SVM	SIMCA	PLS-DA
P, % для зразків річкових / джерельних вод	67 / 33	50 / 67	50 / 50	50 / 17
Параметри для зразків річкових вод	–	$N_{sv} = 16$, $\gamma = 1$	$Var_1 = 70\,\%$, $Var_2 = 75\,\%$, $Var_3 = 55\,\%$, $Var_4 = 64\,\%$	$Var = 70\,\%$
Параметри для зразків джерельних вод	–	$N_{sv} = 18$, $\gamma = 0.00316$	$Var_1 = 100\,\%$, $Var_2 = 61\,\%$, $Var_3 = 62\,\%$, $Var_4 = 59\,\%$, $Var_5 = 62\,\%$, $Var_6 = 100\,\%$	$Var = 35\,\%$

При використанні ймовірнісної та динамічної нейронних мереж надійність ідентифікації зразків вод за даними про вміст 7 металів склала 100 %.

Додатково дослідили стійкість алгоритмів ШНМ до наявності в даних пропусків. Використали чотири зразки річкових вод, аналіз яких проводився в 2011 р. і в яких концентрації іонів нікелю не були визначені (табл. 3.3).

Для навчання нейронних мереж використали 22 зразки вод із табл. 3.2 і перелічені в табл. 3.4 функції активації. Застосовували ті ж алгоритми навчання, що використовували раніше для ідентифікації зразків річкових і джерельних вод, дані про склад яких наведено в табл. 3.1, 3.2: для каскадної мережі, мережі з прямим поширенням сигналу і динамічної мережі – алгоритм Левенберга–Марквардта; для мережі Елмана – алгоритм зворотного поширення помилки.

Відсутні значення концентрацій іонів нікелю в даних контрольної вибірки замінили середньою концентрацією (0.0144 мг/л) цього елемента в зразках навчальної вибірки (в даних про склад зразків річкових вод, відібраних і проаналізованих у 2008–2010 роках).

Дані, наведені в табл. 3.7, дозволяють порівняти надійність різних алгоритмів класифікації. Використання традиційних алгоритмів дало найгірші результати; найкращі результати (100% надійність) забезпечила динамічна нейронна мережа. Можна дійти висновку, що застосування нейронних мереж забезпечує надійну ідентифікацію зразків річкових вод навіть за відсутності однієї з характеристик, що описують зразки навчальної вибірки (еталонів).

Таблиця 3.7

Результати ідентифікації зразків річкових вод за наявності в даних пропусків

Алгоритм	Оптимальна кількість нейронів	Мінімальна ненадійність, P, %
Каскадна мережа	14	25
Мережа з прямим поширенням сигналу	15	25
Мережа Елмана	15	25
Динамічна мережа	14	0
Ймовірнісна мережа	22	25
LDA, SVM*, SIMCA**, PLS-DA**	–	75

* У методі опорних векторів використовували $N_{sv} = 22$ і $\gamma = 1$.

** При варіюванні головних компонент / прихованих змінних для SIMCA і PLS-DA від 1 до 3 ненадійність не змінюється.

3.2. Ідентифікація географічного походження овочів і фруктів

Однією з актуальних задач якісного хімічного аналізу є перевірка справжності продуктів харчування і харчової сировини та виявлення фальсифікованої продукції. Серед способів фальсифікації харчових продуктів і сільськогосподарської сировини досить поширеною є повна або часткова їх підміна замінниками іншого найменування або сорту з позначенням регіону, що має на ринку кращу репутацію [14]. Ми визнали за необхідне вивчити застосовність хемометричних і статистичних методів для ідентифікації географічного походження продуктів харчування рослинного походження.

Існують непоодинокі докази того, що рівень забруднення об'єктів довкілля (атмосферного повітря, ґрунтів, поверхневих вод) іонами металів-токсикантів позначається на вмісті цих металів у продуктах рослинного походження (див., наприклад, [15, 16]). Це дозволило висунути гіпотезу, що за вмістом металів-токсикантів у рослинних продуктах можна ідентифікувати географічне походження цих продуктів. Встановленням географічного походження вважали ідентифікацію типу ландшафту[3)], який чинить вплив на мікроелементний склад продукту [18].

Для перевірки цієї гіпотези використали експериментальні дані про вміст металів в овочах і фруктах із різних районів м. Харкова і Харківської області, отримані у 2008–2010 роках [18]. Типи ландшафтів, характерні для Харківської області, представлені на рис. 3.6 [19].

Аналізований масив даних містив 58 зразків картоплі і 22 зразки яблук (табл. 3.8, 8 груп зразків картоплі та 4 групи зразків яблук, номери груп відповідають номерам типів ландшафтів).

Таблиця 3.8

Райони відбору та кількість зразків картоплі та яблук

Тип ландшафту відповідно до рис. 3.6	Кількість зразків картоплі	Кількість зразків яблук
1	11	3
2	6	–
3	–	–
4	5	4
5	2	–
6	5	–
7	8	–
8	–	–
9	10	2
10	11	13

У зразках рослинного матеріалу визначали концентрації купруму, цинку, плюмбуму, кадмію, мангану, феруму, кобальту, нікелю, алюмінію і хрому.

---

3) Ландшафт – природно-територіальний комплекс, що має чітке географічне положення та характеризується певною взаємодією людської діяльності і природних компонент рельєфу, ґрунтів, клімату, поверхневих і підземних вод [17].

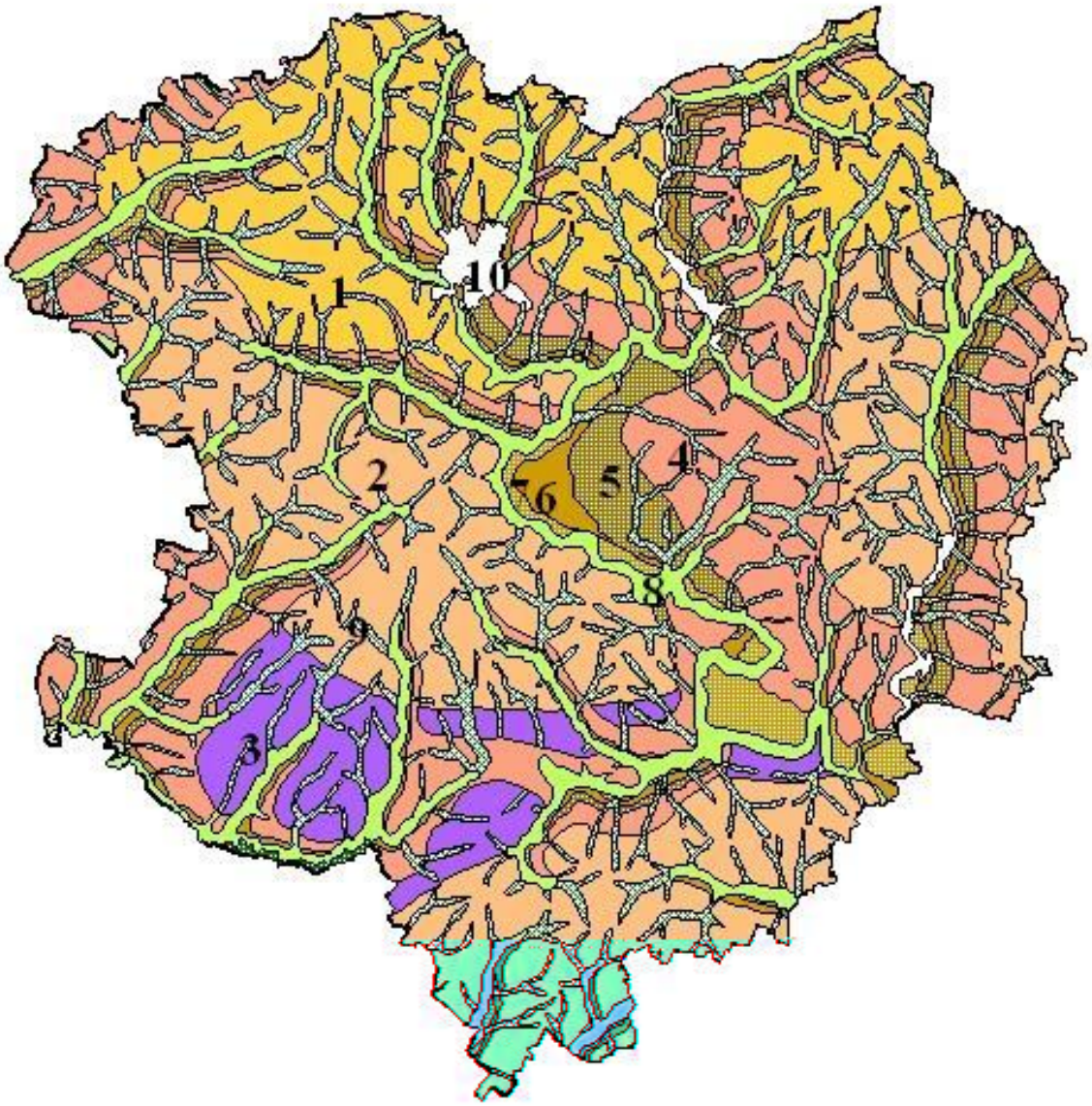

Рис. 3.6. Типи ландшафтів Харківської області (масштаб 1:400000)

1 – міжрічкові рівнини лесові, підведені і відносно вирівняні;
2 – міжрічкові рівнини лесові, відносно вирівняні, розчленовані ярами і балками;
3 – міжрічкові рівнини лесові, розчленовані ярами і балками;
4 – долинні рівнини лесові, сильно розчленовані ярами і балками;
5 – долинні рівнини лесові, розчленовані ярами і балками;
6 – долинні рівнини лесові плоскі;
7 – долинні рівнини дрібногорбисті;
8 – долинні рівнини плоскі і слабкохвилясті;
9 – балочно-долинні рівнини;
10 – м. Харків (урбосистема)

Табл. 3.9 і 3.10 містять середні значення концентрацій металів за трьома паралельними вимірюваннями.

Таблиця 3.9

Концентрації металів у зразках картоплі (мг/кг сухої маси)

Тип ландшафту (рис. 3.6)	Метал									
	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
10	14.40	8.11	9.44	2.90	0.78	0.82	2.40	0.87	0.36	0.08
10	12.30	1.90	3.70	1.90	0.10	н/н	4.40	0.10	0.10	0.01
10	22.10	14.60	5.41	1.17	0.26	0.51	7.40	0.10	0.31	0.06
10	23.00	13.00	4.92	1.20	0.24	0.49	7.30	0.11	0.30	0.08
10	22.00	14.90	5.20	1.24	0.30	0.50	7.20	0.10	0.36	0.07
10	13.60	6.41	5.17	1.75	0.59	0.60	6.42	0.21	0.19	0.09
10	12.00	5.76	6.00	1.80	0.44	0.53	6.40	0.20	0.20	0.03
10	30.09	5.20	10.10	3.20	1.40	0.80	2.40	0.30	0.82	0.12
7	30.50	4.70	8.14	2.90	0.60	0.54	6.40	0.24	0.29	0.06
7	9.20	14.10	7.60	4.70	0.46	0.31	3.00	0.72	0.16	0.06
7	6.70	7.20	4.10	2.10	0.21	0.24	2.10	0.40	0.07	0.03
7	10.40	9.46	8.00	2.52	0.96	0.58	1.76	0.44	0.31	0.12
7	10.40	9.46	8.00	2.52	0.96	0.58	1.76	0.44	0.31	0.12
7	6.60	7.40	4.10	3.30	0.50	0.41	3.80	0.44	0.21	0.06
1	10.60	18.40	9.43	3.84	0.95	0.55	2.80	0.72	0.30	0.11
1	9.21	10.90	8.11	3.12	0.58	0.34	2.30	0.64	0.24	0.08
1	30.09	5.20	10.10	3.20	1.40	0.80	2.40	0.30	0.82	0.12
10	11.30	7.60	7.10	3.10	0.40	0.18	2.60	0.40	0.20	0.10
10	10.00	7.20	6.80	3.00	0.39	0.20	2.40	0.38	0.20	0.10
6	8.10	8.20	4.90	4.60	0.46	0.56	4.60	0.70	0.30	0.12
4	5.10	2.15	5.60	1.40	0.22	0.40	5.12	0.30	0.24	0.09
4	11.40	1.70	5.80	2.10	0.20	0.25	2.10	0.30	0.16	0.02
2	14.60	7.45	4.70	2.92	0.49	0.63	4.32	0.41	0.21	0.08
6	20.34	6.41	4.05	6.08	0.62	1.76	3.02	0.50	0.48	0.06
9	31.00	5.00	9.64	2.86	1.12	0.80	2.54	0.30	0.56	0.03
9	17.80	10.60	8.65	6.90	1.00	0.57	6.20	1.40	0.43	0.10
1	9.90	8.10	7.43	5.10	0.60	0.31	4.10	1.00	0.21	0.07
5	9.86	10.20	7.94	4.10	0.56	0.68	4.80	0.68	0.26	0.10
6	10.60	7.80	5.10	2.60	0.51	0.48	3.20	0.50	0.34	0.09

Продовження табл. 3.9

2	12.60	6.30	7.00	1.35	1.06	0.36	2.40	0.46	0.35	0.11
2	10.00	8.00	5.60	4.20	0.58	0.61	4.00	0.80	0.26	0.08
6	10.40	7.00	5.20	2.00	0.36	0.48	3.40	0.35	0.30	0.11
4	28.50	3.81	10.20	1.63	0.38	0.84	3.60	0.62	0.14	0.05
4	32.50	15.00	9.60	5.40	0.25	1.35	3.60	1.12	0.32	0.15
4	19.80	7.60	5.80	2.80	0.40	0.54	4.80	0.28	0.21	0.12
6	27.40	4.40	8.20	2.98	0.49	0.60	7.10	0.20	0.29	0.05
2	32.80	4.90	9.40	3.10	0.52	0.62	7.90	0.23	0.26	0.07
1	20.60	7.70	10.10	4.40	0.51	0.32	5.00	0.12	0.13	0.07
2	23.00	6.69	11.80	5.31	1.40	0.90	4.10	0.22	0.14	0.09
1	10.40	8.76	5.90	4.80	0.60	0.66	4.40	0.86	0.24	0.06
5	10.40	8.90	5.80	4.80	0.54	0.68	4.90	0.69	0.28	0.11
1	21.60	12.40	9.30	3.20	0.72	0.65	2.80	0.70	0.51	0.11
1	10.80	15.60	8.40	3.70	0.91	0.33	2.70	0.30	0.24	0.08
2	12.60	6.30	7.00	1.35	1.06	0.36	2.40	0.46	0.35	0.11
1	6.90	7.40	5.00	4.10	0.40	0.32	3.60	0.44	0.20	0.10
1	8.40	7.90	5.10	4.65	0.42	0.46	4.10	0.62	0.20	0.18
1	8.40	2.06	2.34	0.14	0.11	0.15	4.60	0.02	0.11	0.01
9	54.80	8.60	12.60	4.10	0.50	1.00	3.86	1.94	0.60	0.26
9	51.20	10.80	18.30	4.90	0.62	0.96	3.90	1.80	0.62	0.30
9	18.30	8.90	9.70	3.90	3.80	1.14	5.10	0.60	0.32	0.14
9	22.00	6.94	10.20	4.46	3.62	6.00	4.71	1.20	1.50	0.89
7	21.40	7.00	10.40	4.40	3.50	5.90	4.84	1.10	1.38	0.87
7	24.20	7.21	12.80	4.68	3.50	6.00	4.90	1.14	1.40	0.90
9	18.40	6.04	9.27	1.70	0.50	1.02	1.70	3.48	0.36	0.10
9	17.90	5.43	10.00	1.54	0.46	0.96	1.52	3.61	0.40	0.09
9	14.60	5.92	8.30	1.62	0.48	1.00	1.80	3.30	0.32	0.10
9	14.10	5.83	9.10	1.56	0.42	0.96	1.62	3.27	0.36	0.10
10	35.24	5.59	10.54	2.50	1.12	1.39	7.60	0.17	0.43	3.40

Сумарна відносна невизначеність результатів аналізів не перевищувала 10 %. Прецизійність аналізу в умовах повторюваності (збіжність) і внутрішньолабораторну прецизійність аналізу (внутрішньолабораторну відтворюваність) характеризували значеннями відносних середньоквадратичних відхилень. Для концентрацій Fe, Mn, Zn, Cu і Al вони не перевищували 4.0 %, для концентрацій Ni, Pb,

Co, Cr і Cd – 5.0 %. Правильність методик перевіряли за методом «введено-знайдено» на стандартних розчинах.

Таблиця 3.10

Концентрації металів у зразках яблук (мг/кг сухої маси)

Тип ландшафту (рис. 3.6)	Метал									
	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
10	12.40	2.10	4.24	2.60	0.20	0.13	3.60	0.90	0.26	0.04
10	9.60	2.00	2.10	0.62	0.12	0.82	4.60	0.39	0.30	0.10
10	6.30	2.40	1.00	0.60	0.38	0.46	1.60	0.00	0.23	0.03
10	16.30	3.00	2.40	2.44	0.25	0.54	3.40	0.46	0.40	0.20
10	16.90	2.20	0.72	0.39	0.14	1.20	н/н	0.42	0.36	0.12
10	12.28	2.30	4.61	1.04	0.19	0.74	4.20	0.51	0.30	0.10
10	7.10	1.52	2.00	1.70	0.24	0.28	1.00	0.01	0.14	0.03
10	5.80	1.50	2.60	1.40	0.21	0.30	0.86	0.01	0.11	0.02
10	4.80	0.96	1.50	0.86	0.20	0.60	2.20	0.42	0.16	0.12
10	6.70	1.10	1.80	0.90	0.28	0.84	2.00	0.48	0.23	0.20
10	10.40	2.60	2.90	1.20	0.46	1.12	3.20	0.70	0.34	0.23
10	9.00	1.65	1.85	0.52	0.11	0.76	4.00	0.32	0.24	0.11
10	23.50	1.45	4.19	2.00	0.20	0.61	4.50	0.71	0.30	0.03
1	6.40	1.20	1.45	1.74	0.33	0.26	1.60	0.93	0.14	0.06
1	16.00	1.30	3.62	2.40	0.31	0.30	2.40	0.20	0.43	0.05
1	6.80	2.40	2.00	0.46	0.24	1.10	3.20	0.26	0.21	0.20
9	17.80	2.20	1.20	0.39	0.11	0.84	5.00	0.41	0.20	0.20
9	10.20	2.10	2.60	0.84	0.14	1.10	5.10	0.50	0.36	0.21
4	9.12	2.10	1.85	1.10	0.20	0.26	3.00	0.33	0.21	0.12
4	10.40	2.44	5.10	1.86	0.29	0.16	3.10	0.92	0.24	0.06
4	9.70	1.94	4.06	2.00	0.16	0.10	3.70	0.81	0.20	0.03
4	21.00	1.60	3.80	1.65	0.21	0.54	3.80	0.61	0.27	0.09

Табл. 3.11 містить значення гранично допустимих концентрацій металів в овочах і фруктах [20, 21].

Таблиця 3.11

Гранично допустимі концентрації металів в овочах і фруктах

ГДК, мг/кг	Метал								
	Fe	Mn	Zn	Cu	Ni	Pb	Co	Cr	Cd
Овочі	50.00	20.00	10.00	5.00	0.50	0.50	1.00	0.20	0.03
Фрукти	–	–	10.00	5.00	–	0.40	–	–	0.03

За критерієм χ^2 перевірили гіпотезу про нормальність розподілу концентрацій металів у зразках картоплі і яблук, а за 3s-критерієм − наявність у масиві даних спостережень, що різко виділяються (викидів). Гіпотезу про те, що концентрації Ni, Pb, Co, Cr і Cd в зразках картоплі розподілені нормально, відкинули, а деякі концентрації металів у картоплі були визнані викидами (табл. 3.12, 3.13). Але ці дані з масиву даних не виключали, оскільки для розподілів із хвостами, довшими, ніж хвости нормального розподілу, отримання результатів, що суттєво відрізняються від середнього значення, є досить імовірним. Високі концентрації металів у деяких зразках (наприклад, кадмію і плюмбуму в зразках картоплі) вказують на значний антропогенний вплив на екосистеми в місцях відбору проб. Так, у Зміївському районі Харківської області працюють теплоелектростанція і 10 промислових підприємств (для порівняння: в Краснокутському районі всього 4 підприємства). Не дивно, що при переході від одного регіону до іншого спостерігаються істотні коливання концентрацій металів-забруднювачів.

Таблиця 3.12

Результати перевірки гіпотези про нормальність розподілу концентрацій металів у зразках картоплі і виявлення спостережень, що різко виділяються

Параметр	Метал									
	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
$\chi^2_{експ}$	6.8	9.1	1.0	1.7	66.2	76.1	1.8	33.0	39.6	420.0
$\chi^2_{крит}(f)$ *	7.8 (3)	11.1 (5)	3.8 (2)	7.8 (3)	12.6 (4)	6.0 (3)	6.0 (3)	12.6 (4)	3.8 (2)	11.1 (5)
Кількість викидів	2	–	–	–	4	1	–	4	1	1

*Тут і в табл. 3.13 $\chi^2_{крит}(f)$ – критичні значення критерію χ^2 для рівня значущості 5 % і f ступенів свободи.

Таблиця 3.13

Результати перевірки гіпотези про нормальність розподілу концентрацій металів у зразках яблук

Параметр	Метал									
	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
$\chi^2_{експ}$	1.7	2.2	3.5	1.7	1.3	0.8	0.9	1.0	1.6	3.2
$\chi^2_{крит}(f)$	3.8 (1)									

Для знаходження залежностей між характеристиками зразків картоплі і яблук та їх географічним походженням до масивів даних застосували комплекс статистичних і хемометричних процедур.

Із статистичних методів використовували непараметричні алгоритми – розрахунок коефіцієнтів рангової кореляції Спірмена, критерію Уілкоксона–Манна–Уітні та його узагальнення – критерію Краскела–Уолліса і параметричний алгоритм – розрахунок коефіцієнтів кореляції Пірсона. З хемометричних процедур застосовували метод головних компонент та ймовірнісну нейронну мережу.

Коефіцієнт рангової кореляції Спірмена слугує мірою лінійного зв'язку між двома величинами незалежно від виду їх розподілу. Якщо знайдене значення коефіцієнта більше за критичне при відповідних рівні значущості та числі ступенів свободи, то кореляцію вважають статистично значущою [22]. Коефіцієнт кореляції Пірсона використовують для обробки нормально розподілених даних. Інтерпретуючи значення коефіцієнта кореляції (*r)*, приймають: $r > 0.9$ – дуже сильна кореляція, $r = 0.7 - 0.9$ – сильна кореляція, $r = 0.5 - 0.7$ – середня кореляція, $r = 0.2 - 0.5$ – слабка кореляція, $r < 0.2$ – дуже слабка кореляція [23].

Критерій Уілкоксона–Манна–Уітні призначений для оцінки відмінностей між двома незалежними вибірками. Відмінність / подібність вибірок визначали за вмістом металу, що чинить найбільший вплив на ідентифікацію зразків. Якщо розраховане значення критерію вище за критичне, то приймають гіпотезу про відсутність відмінностей між вибірками (H_0), інакше – гіпотезу про наявність суттєвої відмінності між вибірками (H_1).

За критерієм Краскела–Уолліса перевіряють рівність медіан декількох вибірок. Цей критерій заснований на рангах, а не на початкових спостереженнях, тому він є інваріантним по відношенню до будь-якого монотонного перетворення шкали вимірювань. Ми за допомогою критерію Краскела–Уолліса визначали метал, вміст якого чинить найбільший вплив на віднесення зразка рослинного матеріалу до тієї чи іншої групи. Якщо розраховане значення критерію більше за критичне, то тестований параметр істотно змінюється залежно від групи зразків, в іншому випадку статистично значущі відмінності для тестованого параметра у групах відсутні [24].

Ймовірнісну нейронну мережу застосували як класифікаційний інструмент для ідентифікації зразків картоплі та яблук за їх географічним походженням з тієї причини, що цей алгоритм є стійким до порушення гіпотези про нормальний розподіл експериментальних похибок.

Перед застосуванням PNN концентрації металів у зразках рослинного матеріалу піддали автомасштабному перетворенню (2.12). Для одного зразка картоплі вміст плюмбуму та для одного зразка яблук вміст алюмінію не були визначені (концентрації виявилися нижчими за межу виявлення); відсутні дані заповнили нулями. Для реалізації ймовірнісної мережі значення параметра відхилення радіальної базисної функції активації задавали 0.1. Для навчання мережі використовували навчальну вибірку об'ємом 79 % (46 зразків картоплі і 17 зразків яблук). У результаті обробки аналізованих масивів даних PNN зіткнулися з проблемою ідентифікації деяких груп зразків: для картоплі – зразків, відібраних із ландшафтів типів 4, 5 і 6, для яблук – зразків із ландшафтів типів 1 і 9. Це можна пояснити недостатньою кількістю зразків, відібраних із цих типів ландшафтів. У зв'язку з неможливістю ідентифікувати всі групи зразків перевірили можливість об'єднання груп за допомогою критеріїв Краскела–Уолліса та Уілкоксона–Манна–Уітні.

В табл. 3.14 представлено результати розрахунку критерію Краскела–Уолліса. Встановлено, що на ідентифікацію зразків картоплі найбільший вплив чинить вміст кобальту, а на ідентифікацію зразків яблук – плюмбуму.

Таблиця 3.14

Результати розрахунку критерію Краскела–Уолліса

Продукт	Метал									
	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
Картопля*	12.9	7.0	13.7	11.9	15.0	17.4	7.0	**25.0**	16.4	4.4
Яблука**	2.0	1.2	2.8	3.9	6.3	**7.2**	6.0	2.1	0.6	3.8

* Критичні значення критерію для рівнів значущості 10 і 5 % при числі ступенів свободи 7 дорівнюють 12.0 і 14.1, відповідно.

** Критичні значення критерію для рівнів значущості 10 і 5 % при числі ступенів свободи 3 дорівнюють 6.2 і 7.8, відповідно.

Результати розрахунку критерію Уілкоксона–Манна–Уітні представлені в табл. 3.15 і 3.16. У випадку зразків картоплі в одну групу об'єднали зразки, а) відібрані з ландшафтів типів 5 і 9, оскільки було прийнято гіпотезу H_0 про відсутність відмінностей між вибірками; і б) відібрані з ландшафтів типів 4 і 6, оскільки також було прийнято гіпотезу про відсутність відмінності між вибірками, а зразки, відібрані з цих типів ландшафтів, з використанням алгоритму PNN не ідентифікувалися.

Таблиця 3.15

Результати розрахунку критерію Уілкоксона–Манна–Уітні для зразків картоплі

Метал	Кобальт						
Тип ландшафту	1	2	4	5	6	7	9
10	H_0	H_1	H_0	H_1	H_0	H_1	H_1
1		H_0	H_0	H_0	H_0	H_0	H_1
2			H_0	H_0	H_0	H_0	H_1
4				H_0	**H_0**	H_0	H_1
5					H_0	H_0	**H_0**
6						H_0	H_1
7							H_1

Таблиця 3.16

Результати розрахунку критерію Уілкоксона–Манна–Уітні для зразків яблук

Метал	Плюмбум			
Тип ландшафту	10	1	9	4
10		H_0	H_0	H_1
1			**H_0**	H_0
9				H_0

У випадку зразків яблук в одну групу об'єднали зразки, відібрані з ландшафтів типів 1 і 9.

Об'єднання груп узгоджується з близькістю характеристик відповідних типів ландшафтів: типи ландшафтів 4 і 6 – долинні; тип ландшафту 9 пронизує всю карту ландшафтів у вигляді тонких «жилок», і видається можливим об'єднувати його з будь-яким іншим типом ландшафту.

Вищезгадані дії дозволили правильно ідентифікувати зразки тестових вибірок із застосуванням PNN, а також при поєднанні PNN і PCA (табл. 3.17). У табл. 3.18 наведено власні значення головних компонент. Перші чотири компоненти описують 77.8 % загальної дисперсії ознак зразків картоплі; перші три компоненти описують 69.7 % загальної дисперсії ознак зразків яблук.

Таблиця 3.17

Результати обробки масивів даних PNN і PCA після об'єднання груп

Метод	Параметр	Продукт	
		Картопля	Яблука
PNN+PCA	Кількість головних компонент	4	3
	Навчальна вибірка	78 % (45 зразків)	77 % (17 зразків)
	Тестова вибірка	22 % (13 зразків)	23 % (5 зразків)
PNN	Навчальна вибірка	91 % (53 зразки)	82 % (18 зразків)
	Тестова вибірка	9 % (5 зразків)	18 % (4 зразки)

Таблиця 3.18

Власні значення головних компонент

Компонента	Зразки картоплі	Зразки яблук
1	3.66	2.90
2	1.57	2.73
3	1.33	1.34
4	1.22	–

Як приклад на рис. 3.7 наведено графік рахунків для зразків яблук. Для зразків картоплі попередня обробка даних за методом головних компонент дозволила суттєво збільшити об'єм тестової вибірки. На рис. 3.8 наведено фрагмент карти типів ландшафтів із точками відбору зразків яблук, на рис. 3.9 – карту ландшафтів із точками відбору зразків картоплі.

Для того щоб з'ясувати, чи можна скоротити кількість металів, концентрації яких необхідно визначати у зразках картоплі та яблук для надійної ідентифікації їх географічного походження, розрахували коефіцієнти рангової кореляції Спірмена (табл. 3.19) та коефіцієнти кореляції Пірсона (табл. 3.20).

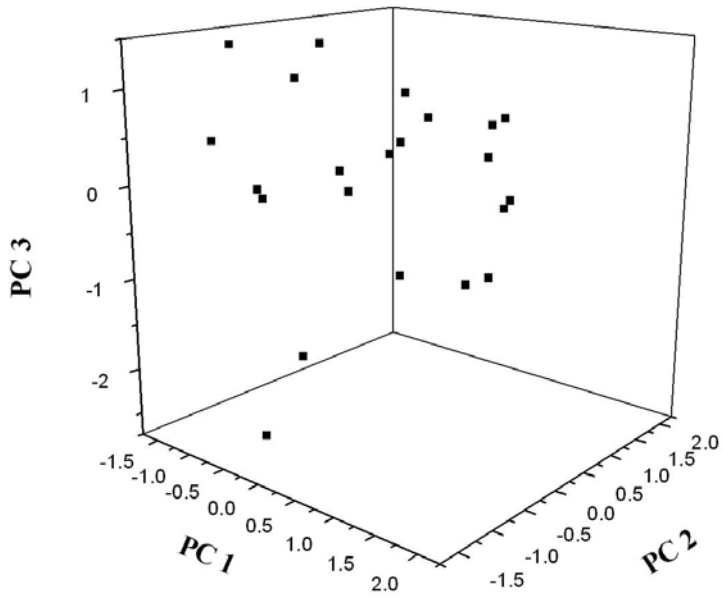

Рис. 3.7. Графік рахунків для зразків яблук

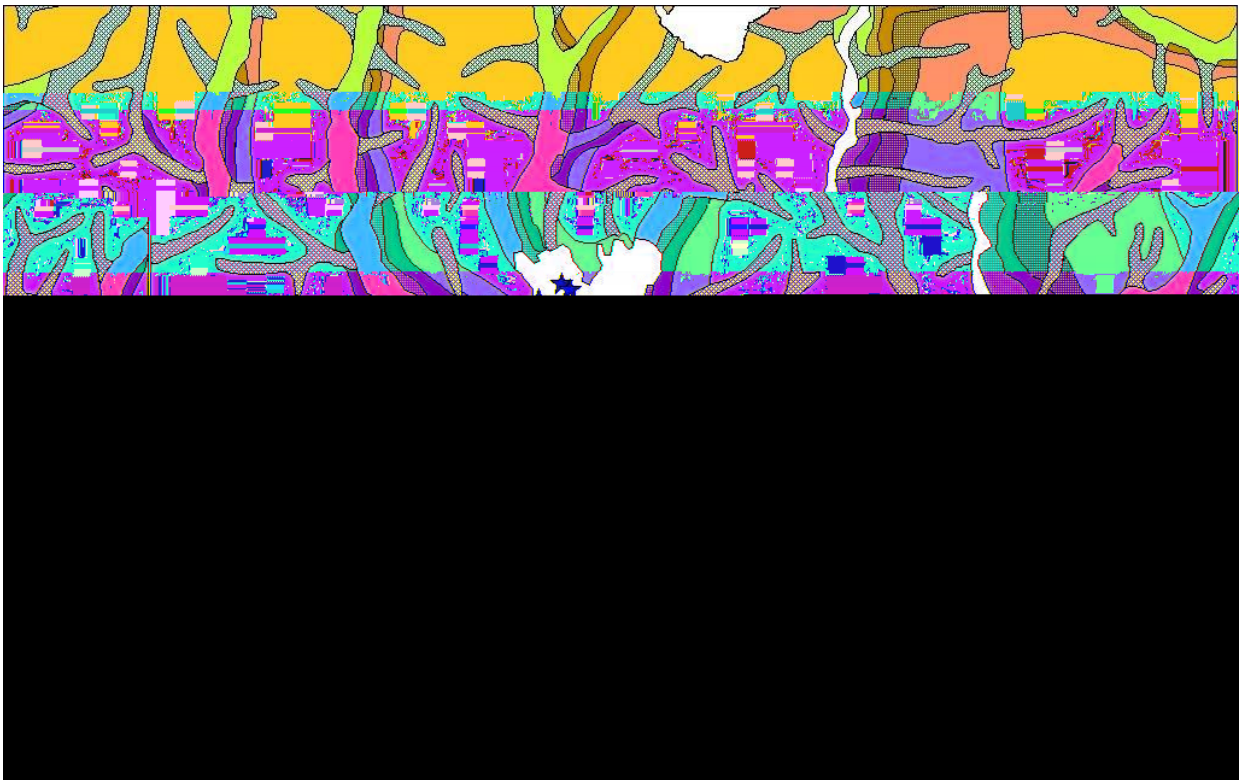

Рис. 3.8. Фрагмент карти типів ландшафтів Харківської області з точками відбору зразків яблук (колір зірки відповідає певній групі зразків, відібраній на відповідних типах ландшафту)*

* *тип ландшафту № 10*

типи ландшафтів № 1, 9

тип ландшафту № 4

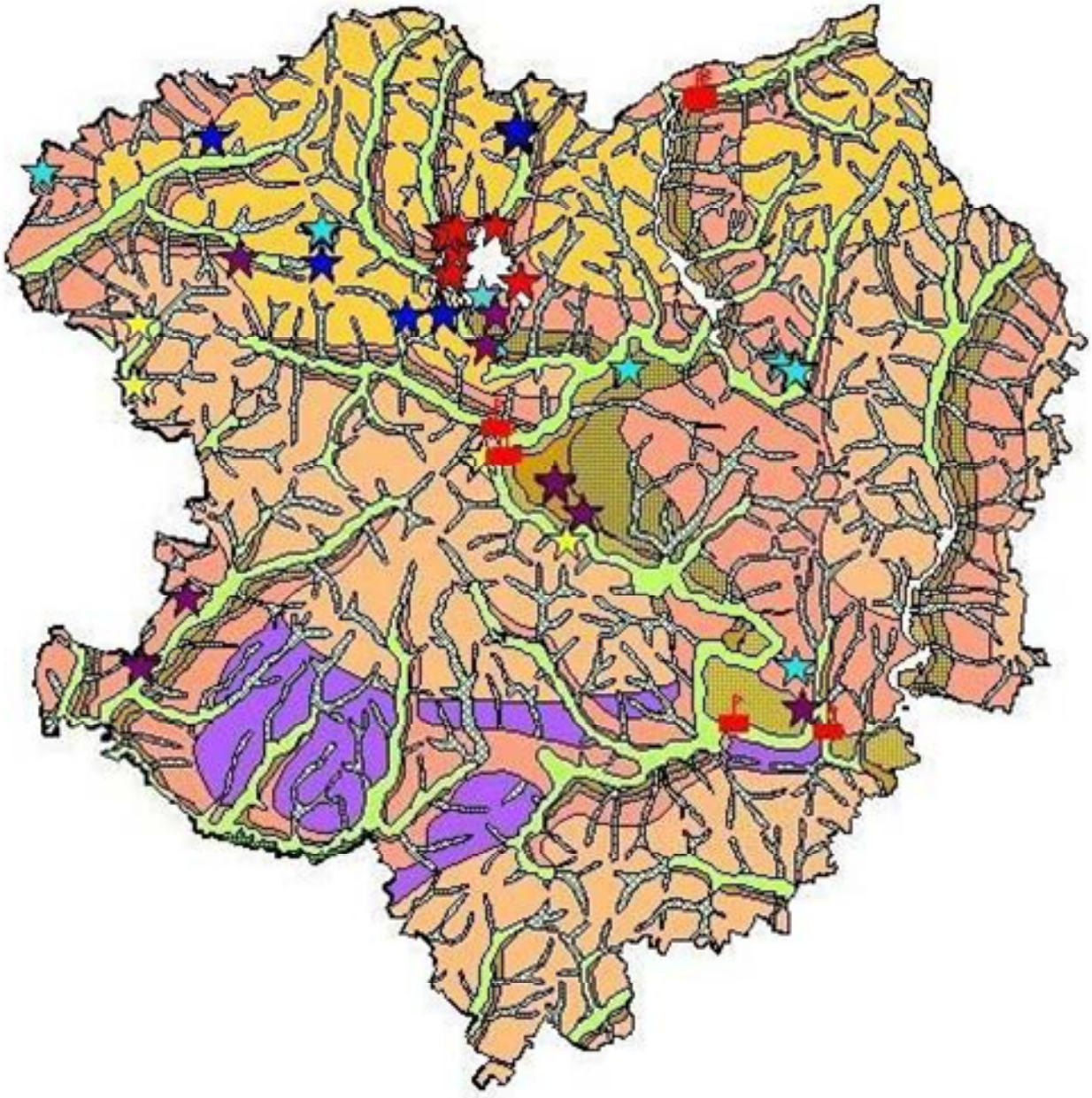

Рис. 3.9. Карта типів ландшафтів Харківської області з точками відбору зразків картоплі (значок відповідає певній групі зразків, відібраній на відповідних типах ландшафту)*

* *тип ландшафту № 10*

тип ландшафту № 1

тип ландшафту № 2

типи ландшафтів № 4, 6

тип ландшафту № 7

типи ландшафтів № 5, 9

Таблиця 3.19

Значення коефіцієнтів рангової кореляції Спірмена для концентрацій металів у зразках картоплі

Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd	Метал
1.19	**6.35**	0.11	2.48	5.84	**1.84**	0.48	4.81	1.55	Fe
	0.07	**3.24**	0.76	0.26	0.33	2.18	0.73	2.03	Mn
		2.17	**5.47**	**6.02**	0.69	2.91	4.50	3.80	Zn
			3.40	2.00	0.91	3.40	0.70	2.28	Cu
				4.43	0.35	2.08	5.19	3.87	Ni
					0.35	**4.07**	**7.05**	4.04	Pb
						3.37	0.82	0.59	Al
							3.62	3.52	Co
								5.24	Cr
$t_{крит.}$ (f =56, рівень значущості (α=0.05) = 1.96; $t_{крит.}$ (f =56, (α=0.01) = 2.79									

Таблиця 3.20

Значення коефіцієнтів кореляції Пірсона для характеристик зразків яблук*

Zn	Cu	Ni	Pb	Al	Co	Cr	Cd	Метал
1	2	3	4	5	6	7	8	9
0.35	0.30			0.38	0.27	0.58		Fe
			0.24	0.27		0.40	0.36	Mn
	0.65			0.41	0.52	0.24		Zn
		0.26			0.41			Cu
						0.38	**0.75**	Pb
					0.36	0.30	0.31	Al
							0.33	Cr

* Значення коефіцієнтів, менші 0.2, не наведено.

З кожної пари корельованих концентрацій металів у зразках картоплі виключили по одному металу, концентрація якого характеризується найбільшою дисперсією (див. табл. 3.21). Так, для зразків картоплі виключили концентрації Fe, Mn, Zn, Pb і Cd. Для зразків яблук у результаті розрахунку коефіцієнтів кореляції Пірсона виявили лише одну пару корельованих вмістів металів, тому суттєве скорочення кількості металів, концентрації яких треба визначати, неможливе.

Застосувавши до скороченого масиву даних, що характеризують зразки картоплі, ймовірнісну мережу, отримали високий об’єм тестової вибірки – 22 % (13 зразків), усі зразки якої ідентифікуються правильно.

Таблиця 3.21

Значення дисперсій концентрацій металів у зразках картоплі

Метал	Fe	Mn	Zn	Cu	Ni	Pb	Al	Co	Cr	Cd
Дисперсія	107.98	12.16	7.74	2.04	0.69	1.53	3.05	0.70	0.09	0.22

***

Таким чином, можна вважати доведеним той факт, що апарат штучних нейронних мереж є ефективним засобом встановлення походження зразків вод навіть за відсутності відомостей про концентрацію одного з них, а найбільшою ефективністю володіють динамічна та ймовірнісна мережі.

Використання комплексу статистичних методів спільно з методом головних компонент та ймовірнісною мережею забезпечило надійну ідентифікацію географічного походження рослинної сировини, підтвердивши гіпотезу про зв'язок географічного походження овочів і фруктів із вмістом у них важких і перехідних металів.

Література до глави 3

1. Ouyang Y. Evaluation of river water quality monitoring stations by principal component analysis / Y. Ouyang // Water Res. – 2005. – Vol. 39, No 12. – P. 2621-2635.
2. Parinet B. Principal component analysis: an appropriate tool for water quality evaluation and management–application to a tropical lake system / B. Parinet, A. Lhote, B. Legube // Ecol. Model. – 2004. – Vol. 178, No 3–4. – P. 2953-311.
3. Stanimirova I. Chemometric analysis of the water purification process data / I. Stanimirova, M. Polowniak, R. Skorek, A. Kita, E. John, F. Buhl, B. Walczak // Talanta. – 2007. – Vol. 74, No 1. – P. 153-162.
4. Rodrigues P. M. S. M. Multivariate analysis of the water quality variation in the Serra da Estrela (Portugal) Natural Park as a consequence of road deicing with salt / P. M. S. M. Rodrigues, R. M. M. Rodrigues, B. H. F. Costa, A. A. L. Tavares Martins, J. C. G. Esteves da Silva // Chemometr. Intell. Lab. – 2010. – Vol. 102, No 2. – P. 130-135.
5. Terrado M. Distribution and assessment of surface water contamination by application of chemometric and deterministic models / M. Terrado, M.-P. Lavigne, S. Tremblay, S. Duchesne, J.-P. Villeneuve, A. N. Rousseau, D. Barceló, R. Tauler // J. Hydrol. – 2009. – Vol. 369, No 3-4. – P. 416-426.

6. Groselj N. Verification of the geological origin of bottled mineral water using artificial neural networks / N. Groselj, G. van der Veer, M. Tusar, M. Vracko, M. Novic // Food Chem. – 2010. – Vol. 118, No 4. – P. 941-947.
7. Brodnjak-Voncina D. Chemometrics characterisation of the quality of river water / D. Brodnjak-Voncina, D. Dobcnik, M. Novic, J. Zupan // Anal. Chim. Acta. – 2002. – Vol. 462, No 1. – P. 87-100.
8. Идентификация образцов воды источников и рек г. Харьков: сравнение методов многомерного анализа данных // Я. Н. Пушкарева, А. Б. Следзевская, А. В. Пантелеймонов, Н. П. Титова, О. И. Юрченко, В. В. Иванов, Ю. В. Холин // Вестн. Моск. ун-та. Сер. 2. Химия. – 2012. – Т. 53, No 6. – P. 405-412.
9. Yuferova E. V. Variations in the blank signal in the atomic-absorption determination of Zn in waters with low mineral content / E. V. Yuferova, Yu. I. Sizykh, A. N. Smagunova // J. Anal. Chem. – 1997. – Vol. 52, No 9. – P. 819-821.
10. Басаргин Н. Н. Групповое концентрирование меди, цинка и свинца в анализе природных и сточных вод / Н. Н. Басаргин, З. С. Сванидзе, Ю. Г. Розовский // Зав. лаб. – 1993. – Т. 59, No 2. – С. 8-9.
11. Yebra-Biurrun M. C. Determination of trace metals in natural waters by flame atomic absorption spectrometry following on-line ion-exchange preconcentration / M. C. Yebra-Biurrun, A. Bermejo-Barrera, M. P. Bermejo-Barrera, M. C. Barciela-Alonso // Anal. Chim. Acta. – 1995. – Vol. 303. – P. 341-345.
12. Юрченко О. И. Анализ питьевой воды Харькова на содержание микроэлементов и анионных поверхностно-активных веществ / О. И. Юрченко, Н. П. Титова, О. В. Козлова // Вісн. Харк. нац. ун-ту. – 2003. – No 596. Хімія. Вип. 10 (33). – С. 110-113.
13. Гігієнічні вимоги до води питної, призначеної для споживання людиною : ДСанПіН 2.2.4-171-10. – Україна, 2010. – 25 с. [Электронний ресурс]. – Режим доступу : http://zakon2.rada. gov.ua/laws/show/z0452-10
14. Luykx D. M. A. M. An overview of analytical methods for determining the geographical origin of food products / D. M. A. M. Luykx, S. M. van Ruth // Food Chem. – 2008. – Vol. 107, No 2. – P. 897-911.
15. Некос А. Н. Использование дисперсионного анализа для определения влияния природных и антропогенных факторов на формирование качества растительной продукции / А. Н. Некос, П. В. Семибратова, Е. В. Высоцкая, А. П. Порван, А. Л. Петухова // Вісник Харківського національного університету імені В. Н. Каразіна. Серія «Екологія». − Харків : Харківський національний університет імені В. Н. Каразіна. − 2012. − No 1004. − С. 79-90.

16. Некос А. Н. Трофогеографія: теорія і практика / А. Н. Некос, Ю. В. Холін. – Харків : Харківський національний університет імені В. Н. Каразіна, 2015. – 296 с.
17. Douglas I. Companion encyclopedia of geography. The environment and humankind / Douglas I., Huggett R., Robinson M. – London, New York : Routledge, 1996. – 1021 p.
18. Пушкарева Я. Н. Особенности идентификации географического происхождения овощей и фруктов с помощью хемометрических и статистических методов / Я. Н. Пушкарева, А. Б. Следзевская, П. В. Семибратова, А. Г. Гарбуз, А. Н. Некос, Ю. В. Холин // Методы и объекты химического анализа. – 2012. – Т. 7, No 4. – С. 184-191.
19. Атлас Харківської області / [наук. редкол.: І. І. Залюбовський та ін.]. – К. : Головне управління геодезії, картографії та кадастру при Кабінеті Міністрів України, 1993. – 44 с.
20. Предельно допустимые концентрации тяжелых металлов и мышьяка в продовольственном сырье и пищевых продуктах : СанПиН 42-123-4089-86. – М. : Изд-во стандартов, 1986. – 56 с.
21. Медико-биологические требования и санитарные нормы качества продовольственного сырья и пищевых продуктов : № 5061-89. – М. : Изд-во стандартов, 1990. – 181 с.
22. Орлов А. И. Прикладная статистика : учебник / А. И. Орлов. – М. : Экзамен, 2004. – 656 с.
23. Дубнов П. Ю. Обработка статистической информации с помощью SPSS / П. Ю. Дубнов. – М. : АСТ, НТ Пресс, 2004. – 221 с.
24. Sprent P. Applied nonparametric statistical methods / P. Sprent, N. C. Smeeton. – 3rd ed. – USA : Chapman & Hall / CRC, 2001. – 462 p.

Глава 4

КЛАСТЕРИЗАЦІЯ ОБ'ЄКТІВ БЕЗ АПРІОРНОЇ ІНФОРМАЦІЇ ПРО КІЛЬКІСТЬ КЛАСІВ

Здійснюючи класифікацію «з навчанням», дослідник розподіляє об'єкти навчальної вибірки між наперед визначеними класами; цей розподіл використовують для того, щоб побудувати правила класифікації. Вважають, що класи досить добре відокремлені один від одного в просторі даних, а всі об'єкти навчальної вибірки правильно розподілені за класами [1]. При класифікації «без навчання» навчальних вибірок немає, але кількість класів слід задати перед початком класифікації [2]. Отже, і класифікація «з навчанням», і класифікація «без навчання» вимагають інформації про кількість класів.

Розв'язуючи реальні задачі, хіміки досить часто стикаються з необхідністю аналізувати набори даних, для яких кількість однорідних груп невідома, і її слід визначити в ході обробки даних. Крім того, раціональні критерії для віднесення об'єкта до того чи іншого класу можуть бути неясними, погано формалізованими, нечіткими. Ці обставини обумовлюють необхідність комбінування різних хемометричних методів при проведенні класифікації [3−7].

Ця глава присвячена визначенню кількості класів і знаходженню стійкої класифікації за допомогою процедури, що поєднує мережу Кохонена та ймовірнісну мережу. Викладено результати апробації запропонованої процедури на масиві даних про 9 характеристик 76 органічних розчинників.

Класифікація розчинників є одним із прикладів задач класифікації, для яких однозначне розв'язання є принципово неможливим. Тому було важливо з'ясувати, чи надає запропонована процедура можливість одержувати хімічно значущі змістовні результати. Також класифікація розчинників за дев'ятьма характеристиками привертає особливу увагу, оскільки більшість відомих класифікацій ґрунтується лише на декількох параметрах [8−10].

Розроблену процедуру також було застосовано для кластеризації зразків вод із різних джерел і річок м. Харкова та зразків яблук різного походження.

4.1. Процедура кластеризації об'єктів без апріорної інформації про кількість класів

Проводити обробку масивів експериментальних даних, для яких кількість однорідних груп невідома, а критерії віднесення зразків до того чи іншого класу нечіткі або суперечливі, ми пропонуємо за процедурою, що поєднує мережу Кохонена «без навчання» з імовірнісною мережею «з навчанням».

Роль мережі Кохонена полягає в попередній кластеризації даних і знаходженні кількості однорідних груп досліджуваних об'єктів; роль PNN – у знаходженні стійкої класифікації досліджуваних об'єктів. Мережу Кохонена обрано з тієї причини, що вона за надійністю перевершує стандартні алгоритми кластерного аналізу. В табл. 4.1 порівнюються результати використання для обробки різних масивів даних алгоритму згаданої мережі та низки алгоритмів ієрархічної кластеризації («ближнього сусіда», «далекого сусіда», «середнього зв'язку», «центроїдного») з використанням Евклідової метрики близькості, а також методу k-середніх і нечіткого методу k-середніх. Вибір імовірнісної мережі був обумовлений тим, що їй властиві висока стійкість до наявності в даних похибок, розподіл яких відрізняється від розподілу Гауса, і пропусків, а також проста архітектура.

Алгоритм класифікації без апріорної інформації про кількість класів і без наявності навчальної вибірки включає такі кроки:

1) класифікація даних за допомогою мережі Кохонена при різних значеннях кількості нейронів (відповідно і кількості класів);

2) визначення груп зразків, які незалежно від кількості заданих нейронів віднесені мережею Кохонена до одного і того ж класу; використання цих зразків в якості першої навчальної вибірки для навчання ймовірнісної мережі;

3) випадкове формування невеликих вибірок приблизно однакового розміру зі зразків, що не увійшли до першої навчальної вибірки, і послідовного їх пред'явлення на вхід імовірнісної мережі ($\delta = 0.1$) як тестових вибірок;

4) включення зразків кожної тестової вибірки до навчальної вибірки після їх класифікації ймовірнісною мережею (навчальна вибірка збільшується, що забезпечує адекватну класифікацію наступних тестових вибірок);

5) проведення крос-валідації (leave-one-out cross validation) [11, 12] для перевірки й уточнення отриманої класифікації.

Схему запропонованої процедури класифікації наведено на рис. 4.1.

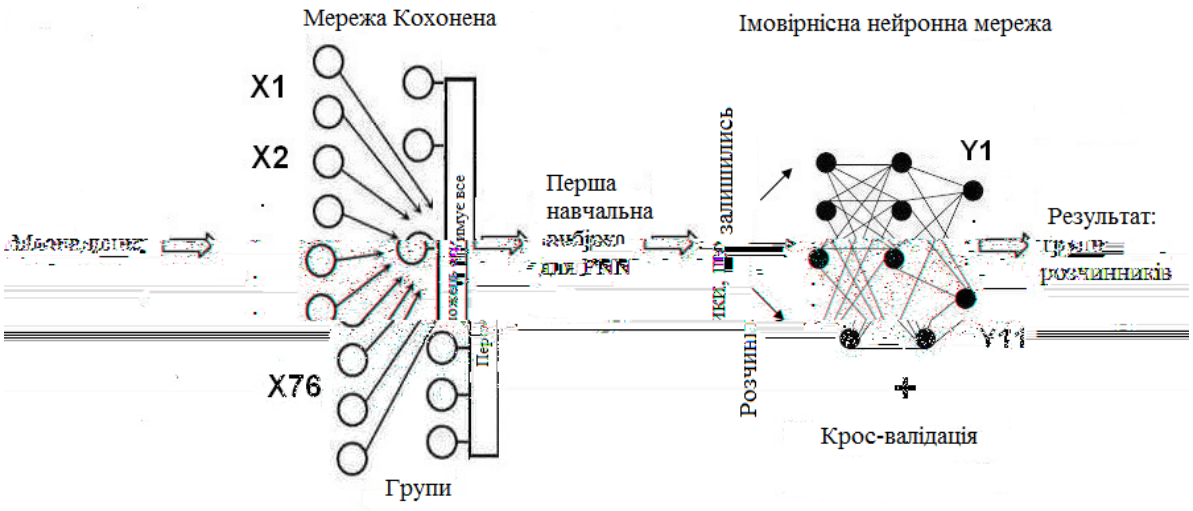

Рис. 4.1. Схема процедури класифікації, що ґрунтується на поєднанні мережі Кохонена та ймовірнісної мережі

Процедура крос-валідації полягала в послідовному виключенні об'єктів із навчальної вибірки, навчанні на об'єктах, що залишилися, і розпізнаванні виключених об'єктів. Досліджуваний масив даних розбивається на k вибірок, що не перетинаються та мають однакові розміри. На кожній ітерації навчання проводиться за k-1 вибірками, а тестування – на вибірці, що залишилася. Процедура крос-валідації повторюється k разів до тих пір, поки кожна з вибірок не буде використана в якості тестової.

Таблиця 4.1

Результати кластерного аналізу масивів даних

Метод	*P*, %					
	Дугопо-дібні дані	Двоієра-рхічні дані	Зразки ірисів	Зразки вин	Зразки річкових вод	Зразки джерель-них вод
мережа Кохонена (η=0.01)	**39**	**0**	**9**	**11**	**54**	**46**
k-середніх	46	0	11	26	86	91
нечіткий метод k-середніх	46	10	10	40	77	83
«ближнього сусіда»	54	0	33	69	77	87
«далекого сусіда»	46	0	15	29	77	83
«середнього зв'язку»	46	0	9	29	77	87
«центроїдний алгоритм»	46	0	32	38	82	95

На перших етапах алгоритму необхідно отримати не багато груп, що включають один-два зразки, а декілька груп, що містять якомога більшу кількість зразків.

4.2. Класифікація розчинників за фізико-хімічними характеристиками

Класифікація розчинників є прикладом однієї з найскладніших задач класифікації. Однозначної класифікації розчинників існувати не може, оскільки кожна з можливих класифікацій спрямована на розв'язання окремого кола задач і надає перевагу тим чи іншим характеристикам речовин [13]. Більшість існуючих класифікацій ґрунтується на використанні обмеженої кількості характеристик розчинників. Зважаючи на ці обставини, ми використали запропоновану процедуру для класифікації набору розчинників і вважали, що розглядати цю процедуру як застосовну можна в тому випадку, якщо вона дозволяє отримати хімічно змістовні результати.

Досліджували масив даних, що включав 76 розчинників, за набором із 9 фізико-хімічних характеристик:

- параметр розчинності Ґільдебранда, δ, $Дж^{1/2}см^{-3/2}$,
- поверхневий натяг, γ, 10^{-2} Н / м,
- дипольний момент молекули, μ, Д,
- відносна діелектрична проникність, ε,
- показник заломлення, n,
- емпіричний параметр кислотності розчинника як донора воднених зв'язків, α,
- емпіричний параметр полярності і поляризовності, π^*,
- емпіричний параметр полярності Райхардта, E_T, Ккал / моль,
- Str. – структурованість розчинника.

Числові значення параметрів, взяті з роботи [8], наведено в Додатку (табл. Д4).

У взятому з роботи [8] досліджуваному масиві даних наявні чотири пропуски: невідомі значення структурованості 1-деканолу, анізолу, гексаметилфосфораміду та пропіленкарбонату. Ці пропуски заповнили середнім значень структурованості інших 72 розчинників (воно становило 0.63) і надалі обробляли модифікований масив даних. Перед застосуванням процедури класифікації здійснили автомасштабне перетворення даних (див. формулу (2.12)). При виконанні кластеризації з використанням мережі Кохонена обирали таку кількість нейронів: $5 \le h \le 8$ (табл. 4.2).

Таблиця 4.2

Результати класифікації органічних речовин мережею Кохонена при різних кількостях нейронів (*h*) (тут і в наступних таблицях зразки, віднесені мережею Кохонена до одного і того ж класу незалежно від кількості заданих нейронів, відмічені жирним курсивом)

№	Розчинник	Віднесення до класу			
		$h = 5$	$h = 6$	$h = 7$	$h = 8$
1	н-пентан	***1***	***4***	***2***	***1***
2	н-гексан	***1***	***4***	***2***	***1***
3	н-гептан	***1***	***4***	***2***	***1***
4	н-октан	***1***	***4***	***2***	***1***
5	і-октан	***1***	***4***	***2***	***1***

Продовження табл. 4.2

6	н-декан	***1***	***4***	***2***	***1***
7	н-додекан	***1***	***4***	***2***	***1***
8	н-тетрадекан	1	4	2	2
9	н-гексадекан	1	4	2	2
10	циклогексан	***1***	***4***	***2***	***1***
11	метилциклогексан	***1***	***4***	***2***	***1***
12	цис-декалін	1	5	4	2
13	бензол	***3***	***5***	***4***	***2***
14	п-ксилол	***3***	***5***	***4***	***2***
15	мезитилен	***3***	***5***	***4***	***2***
16	гексафторбензол	1	4	1	2
17	тетрахлорметан	1	5	2	2
18	трихлоретилен	***3***	***5***	***2***	***2***
19	тетрахлоретилен	***3***	***5***	***2***	***2***
20	толуол	***3***	***5***	***2***	***2***
21	о-ксилол	***3***	***5***	***2***	***2***
22	м-ксилол	***3***	***5***	***2***	***2***
23	етилбензол	3	5	1	2
24	п-цимол	***3***	***5***	***4***	***2***
25	1,4-діоксан	***3***	***1***	***1***	***6***
26	фторбензол	3	1	1	8
27	дихлорметан	2	1	1	3
28	хлороформ	3	1	1	3
29	1,2-дихлоретан	2	1	5	8
30	1,1,1-трихлоретан	***3***	***1***	***1***	***6***
31	1,1,2,2-тетрахлоретан	4	2	5	8
32	1-хлорпропан	1	1	1	3
33	1,2,3-трихлорпропан	***4***	***1***	***5***	***8***
34	хлорбензол	***4***	***1***	***5***	***8***
35	о-дихлорбензол	4	2	5	5
36	м-дихлорбензол	4	5	7	8
37	1,2,4-трихлорбензол	***4***	***2***	***7***	***5***
38	дибромметан	***4***	***2***	***7***	***5***
39	бромоформ	***4***	***2***	***7***	***5***
40	1,2-диброметан	***4***	***2***	***7***	***5***
41	1-бромпропан	***3***	***1***	***1***	***6***
42	бромбензол	***4***	***2***	***7***	***5***
43	дийодметан	***4***	***2***	***7***	***5***
44	1-йодпропан	***4***	***1***	***5***	***8***

Продовження табл. 4.2

45	йодбензол	***4***	***2***	***7***	***5***
46	тетрагідрофуран	***3***	***1***	***1***	***6***
47	анізол	***4***	***1***	***5***	***8***
48	1-амінобутан	1	1	1	3
49	піридин	2	2	5	8
50	хінолін	***4***	***2***	***7***	***5***
51	дисульфід карбону	4	5	7	5
52	тетрагідротіофен	***4***	***1***	***5***	***8***
53	метанол	***5***	***6***	***3***	***7***
54	етанол	***5***	***6***	***3***	***7***
55	1-пропанол	***5***	***6***	***3***	***7***
56	1-бутанол	***5***	***6***	***3***	***3***
57	1-пентанол	***5***	***6***	***3***	***3***
58	1-гексанол	***5***	***6***	***3***	***3***
59	1-октанол	***5***	***6***	***3***	***3***
60	1-деканол	***5***	***6***	***3***	***3***
61	1,2-етандіол	5	6	6	7
62	вода	5	3	6	7
63	N-метилформамід	5	3	6	7
64	N,N-диметилформамід	***2***	***3***	***6***	***4***
65	N,N-диметилацетамід	***2***	***3***	***6***	***4***
66	N,N-метилпіролідон	***2***	***3***	***5***	***4***
67	гексаметилфосфорамід	***2***	***3***	***5***	***4***
68	диметилсульфоксид	***2***	***3***	***6***	***4***
69	о-крезол	***5***	***6***	***3***	***7***
70	пропіленкарбонат	***2***	***3***	***6***	***4***
71	ацетон	2	6	3	3
72	нітрометан	5	3	6	7
73	нітроетан	***2***	***3***	***6***	***4***
74	нітробензол	***2***	***3***	***5***	***4***
75	ацетонітрил	5	6	6	7
76	бензонітрил	***2***	***3***	***5***	***4***

Було виявлено десять груп розчинників, які, незалежно від кількості заданих у структурі мережі Кохонена нейронів, відносяться мережею до одного і того ж класу. Кожна така група містила не менше чотирьох розчинників. 53 речовини, представлені в табл. 4.3, використали в якості першої навчальної вибірки для навчання ймовірнісної нейронної мережі на другому етапі алгоритму.

23 речовини, що залишилися, випадковим чином розбили на три вибірки (дві вибірки містили по 8 розчинників, третя – 7 розчинників). Ці вибірки використали як тестові для класифікації ймовірнісною нейронною мережею (третій та четвертий етапи алгоритму).

Результати класифікації зразків тестових вибірок наведено в табл. 4.4–4.6.

Таблиця 4.3

Склад першої навчальної вибірки для навчання ймовірнісної мережі

Клас	Розчинник (номер розчинника в таблиці 4.2)
I	н-пентан (1), н-гексан (2), н-гептан (3), н-октан (4), і-октан (5), н-декан (6), н-додекан (7), циклогексан (10), метилциклогексан (11)
II	бензол (13), п-ксилол (14), мезитилен (15), п-цимол (24)
III	трихлоретилен (18), тетрахлоретилен (19), толуол (20), о-ксилол (21), м-ксилол (22)
IV	1,4-діоксан (25), 1,1,1-трихлоретан (30), 1-бромпропан (41), тетрагідрофуран (46)
V	1,2,3-трихлорпропан (33), хлорбензол (34), 1-йодпропан (44), анізол (47), тетрагідротіофен (52)
VI	1,2,4-трихлорбензол (37), дибромметан (38), бромоформ (39), 1,2-диброметан (40), бромбензол (42), дийодметан (43), йодбензол (45), хінолін (50)
VII	метанол (53), етанол (54), 1-пропанол (55), о-крезол (69)
VIII	1-бутанол (56), 1-пентанол (57), 1-гексанол (58), 1-октанол (59), 1-деканол (60)
IX	N,N-диметилформамід (64), N,N-диметилацетамід (65), диметилсульфооксид (68), пропіленкарбонат (70), нітроетан (73)
X	N,N-метилпіролідон (66), гексаметилфосфорамід (67), нітробензол (74), бензонітрил (76)

Таблиця 4.4

Результати класифікації першої тестової вибірки

Клас	Розчинник (номер розчинника в табл. 4.2)
II	н-тетрадекан (8)
III	цис-декалін (12), етилбензол (23), дисульфід карбону (51)
IV	1,2-дихлоретан (29)
VII	ацетон (71)
IX	нітрометан (72)

Таблиця 4.5

Результати класифікації другої тестової вибірки

Клас	Розчинник (номер розчинника в табл. 4.2)
I	гексафторбензол (16)
II	н-гексадекан (9), тетрахлорметан (17)
V	1,1,2,2-тетрахлоретан (31)
VIII	1-хлорпропан (32)
IX	ацетонітрил (75), 1,2-етандіол (61), вода (62)

Таблиця 4.6

Результати класифікації третьої тестової вибірки

Клас	Розчинник (номер розчинника в табл. 4.2)
I	хлороформ (28), 1-амінобутан (48)
IV	фторбензол (26), дихлорметан (27)
V	о-дихлорбензол (35)
VI	м-дихлорбензол (36), піридин (49)
IX	N-метилформамід (63)

Для проведення крос-валідації масив даних із характеристиками 76 розчинників розбили випадковим чином на 5 вибірок: 4 вибірки містили по 15 речовин, п'ята вибірка – 16 речовин. Процедуру крос-валідації повторювали 5 разів: кожну вибірку використовували один раз як тестову вибірку для тестування ймовірнісної мережі, а чотири інші вибірки – в якості навчальних для навчання ймовірнісної мережі. В результаті проведення процедури крос-валідації 12 розчинників змінили свою приналежність до класів (п'ятий етап алгоритму, табл. 4.7).

Таблиця 4.7

Результати крос-валідації

Розчинник (номер розчинника в табл. 4.2)	Клас	
	до крос-валідації	після крос-валідації
м-ксилол (22), толуол (20), о-ксилол (21), трихлоретилен (18), етилбензол (23)	3	2
о-крезол (69)	7	4
циклогексан (10), н-додекан (7), метилциклогексан (11)	1	2
піридин (49)	6	5
анізол (47)	5	6
1-бутанол (56)	8	7

Відзначимо, що хлороформ та 1-амінобутан були віднесені до I класу, що містить неполярні аліфатичні вуглеводні, незважаючи на явну відмінність їх властивостей. Найбільш вірогідним поясненням цьому факту слугує відсутність відповідної групи для хлороформу та 1-амінобутану у вихідній навчальній вибірці. У зв'язку з цим до навчальної вибірки додали додаткову XI групу. Хлороформ використовували як елемент нової групи в навчальній вибірці, а 1-амінобутан – як елемент тестової вибірки. В результаті застосування алгоритму PNN 1-амінобутан був віднесений до однієї групи з хлороформом. Віднесення хлороформу до групи, відмінної від інших галогенозаміщених вуглеводнів, узгоджується з класифікацією розчинників за Л. Снайдером [14] та М. Шастретом [15].

Остаточна класифікація 76 розчинників [16] представлена в табл. 4.8.

Обговоримо фізико-хімічні міркування, що пояснюють віднесення розчинників до тих чи інших класів.

Клас I містить неполярні аліфатичні вуглеводні з числом атомів карбону до 10 та гексафторбензол. Ці речовини характеризуються дуже низькою діелектричною проникністю і показниками заломлення нижче двох.

Клас II містить неполярні та слабкополярні аліфатичні вуглеводні з числом атомів карбону більше 10, циклічні й ароматичні вуглеводні, а також трихлоретилен та тетрахлорметан. У цих речовин діелектричні проникності та показники заломлення більше двох, чим можна пояснити розбиття аліфатичних вуглеводнів на два класи. Можна вважати, що гексафторбензол віднесений до I класу також на основі значень діелектричної проникності і показника заломлення.

Клас III містить цис-декалін, дисульфід карбону, тетрахлоретилен. Цис-декалін відрізняється від циклічних вуглеводнів II класу вищим значенням поверхневого натягу (значення поверхневого натягу цис-декаліну складає 31.6, тоді як значення цього параметра в II класі не перевищують 29.5). Полігалоненопохідні аліфатичних вуглеводнів (тетрахлорметан і тетрахлоретилен) віднесені до різних класів у зв'язку з відмінностями в значеннях поверхневого натягу (значення поверхневого натягу тетрахлорметану складає 26.1, а тетрахлоретилену – 31.3).

Класи IV–VI включають слабкополярні галогенопохідні аліфатичних і ароматичних вуглеводнів, гетероциклічні сполуки (хінолін

і піридин), ароматичні і циклічні ефіри (тетрагідрофуран, анізол, 1,4-діоксан), а також тетрагідротіофен і о-крезол. Ці класи речовин утворені на основі специфічної комбінації значень їх 9 параметрів, але відзначимо, що при переході від IV класу до VI спостерігається збільшення значень показника заломлення. Трихлоретилен (клас II) відрізняється від галогенопохідних аліфатичних вуглеводнів IV–VI класів меншими значеннями дипольного моменту, діелектричної проникності і структурованості. Єдиний представник фенолів, о-крезол, відрізняється від аліфатичних спиртів вищими значеннями більшості своїх характеристик (поверхневого натягу, показника заломлення, емпіричного параметра кислотності як донора водневих зв'язків, емпіричного параметра полярності і поляризовності, структурованості) і меншими значеннями дипольного моменту і діелектричної проникності.

Клас VII містить аліфатичні спирти з числом атомів карбону до 5 і ацетон; *клас VIII* – аліфатичні спирти з числом атомів карбону від 5 до 10 і 1-хлорпропан. 1-Хлорпропан характеризується найменшими значеннями поверхневого натягу і показника заломлення серед усіх галогенопохідних аліфатичних вуглеводнів, розглянутих при класифікації, що пояснює його віднесення до цього класу.

Класи IX і *X* містять високополярні сполуки. Розчинники цих груп відрізняються високими значеннями діелектричної проникності, емпіричного параметра полярності Райхардта і показника заломлення.

Хлороформ та 1-амінобутан формують *клас XI*. Хлороформ характеризується найвищим значенням емпіричного параметра кислотності як донора водневих зв'язків. Характеристики 1-амінобутану та хлороформу подібні. Наприклад, близькими є параметри заломлення і діелектричні проникності. Виділення хлороформу та 1-амінобутану в окремий клас слід вважати прийомом, що враховує значну відмінність характеристик вказаних речовин від інших розглянутих розчинників.

Про те, що отримана нами класифікація розчинників не є формальною, свідчить той факт, що вона близька до розбиття розчинників на групи, запропонованого для інших груп сполук і наборів характеристик.

У роботі М. Шастрета та співавторів [15] здійснено класифікацію 83 розчинників за набором із 8 параметрів (функції Кірквуда, молярної рефракції, параметра розчинності Ґільдебранда, показника заломлення,

температури кипіння, дипольного моменту, енергій вищої заповненої і нижчої незаповненої молекулярних орбіталей) на 9 груп: 1) апротонні полярні (14 розчинників); 2) апротонні сильнополярні (9 розчинників); 3) апротонні сильнополярні з високою поляризовністю (2 розчинники); 4) ароматичні неполярні (8 розчинників); 5) ароматичні відносно полярні (12 розчинників); 6) електронодонорні (10 розчинників); 7) здатні до утворення водневих зв'язків (19 розчинників); 8) сильно асоційовані розчинники, здатні до утворення водневих зв'язків (5 розчинників); 9) розчинники з невизначеною функцією (4 розчинники).

Таблиця 4.8

Запропонована класифікація розчинників. Числа в дужках − логарифми мольних часток (*x*) фулерену C_{60} в насичених розчинах при 298 K. Значення *x*, якщо не вказано інше джерело, запозичені з роботи [17]

Клас		Розчинник
I	Неполярні та слабкополярні	н-пентан (-6.1), н-гексан (-5.1), н-гептан (?), н-октан (-5.2), і-октан (-5.2), н-декан (-4.7), гексафторбензол (?)
II		н-тетрадекан (-4.3), н-гексадекан (?), н-додекан (-3.5), циклогексан (-5.3), метилциклогексан (-4.5), трихлоретилен (-3.8), тетрахлорметан (?), бензол (-4.0), п-ксилол (-3.3), мезитилен (-3.5), п-цимол (-3.6), толуол (-3.4), о-ксилол (-2.9), м-ксилол (3.3), етилбензол (-3.4)
III		цис-декалін (-3.3 [18]), дисульфід карбону (-3.2 [18]), тетрахлоретилен (-3.8)
IV	Слабкополярні	фторбензол (-4.1 [18]), дихлорметан (-4.6), 1,2-дихлоретан (-5.0), 1,1,1-трихлоретан (-4.7), 1-бромпропан (-5.2), тетрагідрофуран (?), **о-крезол (-5.7)**, 1,4-діоксан (-5.3 [19])
V		1,1,2,2-тетрахлоретан (-3.1), 1,2,3-трихлорпропан (-4.0), хлорбензол (-3.0), о-дихлорбензол (-2.4), 1-йодпропан (-4.6), піридин (-4.0), тетрагідротіофен (-5.4)
VI		бромбензол (-3.3), м-дихлорбензол (-3.4), бромоформ (-3.2), йодбензол (-3.5), хінолін (-2.9), 1,2,4-трихлорбензол (-2.8), **дийодметан (-4.8)**, 1,2-дибромметан (-4.2), дибромметан (-4.5), анізол (-3.1)
VII	Донори водневого зв'язку та інші	метанол (практично не розчиняється [20]), етанол (-7.1), 1-пропанол (-6.4), 1-бутанол (-5.9), ацетон (-7.0)
VIII		1-пентанол (-5.3), 1-гексанол (-5.1), 1-октанол (-5.0), 1-деканол (?), 1-хлорпропан (-5.6)
IX	Полярні	N-метилпіролідон (-3.9 [18]), гексаметилфосфорамід (?), нітробензол (-3.9), бензонітрил (-4.2)

Продовження табл. 4.8

X	Полярні	1,2-етандіол (?), вода (практично не розчиняється [20]), N-метилформамід (?), N,N-диметилформамід (-5.3), нітроетан (-6.7), N,N-диметилацетамід (?), диметилсульфоксид (?), пропіленкарбонат (?), нітрометан (практично не розчиняється [20]), ацетонітрил (практично не розчиняється [20])
XI	Інші	хлороформ (-4.8), **1-амінобутан (-3.3)**

* Розчинники, віднесені в роботі [17] до викидів, виділені жирним шрифтом.

П. Граматика та співавтори [4] поділили на групи 152 розчинники, кожен з яких характеризується набором зі 174 молекулярних дескрипторів. Було сформовано п'ять груп: 1) апротонні полярні (58 розчинників); 2) ароматичні неполярні або слабкополярні (25 розчинників); 3) електронодонорні (20 розчинників); 4) здатні до утворення водневих зв'язків (37 розчинників); 5) аліфатичні апротонні неполярні (12 розчинників).

Виходячи з 40 характеристик полярності, А. Катрицький і співавтори [21] запропонували класифікацію 40 розчинників на 5 груп: 1) N-формамід; 2) здатні до утворення водневих зв'язків (9 розчинників); 3) полярні апротонні (12 розчинників); 4) прості і складні ефіри, аміни, ароматичні вуглеводні, галогенопохідні аліфатичних і ароматичних вуглеводнів (15 розчинників); 5) н-гексан, циклогексан і тетрахлорметан.

Трохи пізніше А. Катрицький і співавтори [22] обробили великий масив даних про властивості 703 розчинників, які описували 100 дескрипторами; дескриптори характеризували різноманітні властивості: утворення порожнин в розчиннику, дисперсійну й електростатичну взаємодії, силу водневого зв'язку тощо. Було сформовано 11 груп речовин: 1) вуглеводні (81 розчинник); 2) галогенопохідні вуглеводнів (80 розчинників); 3) етери (58 розчинників); 4) естери (67 розчинників); 5) альдегіди, кетони та аміди (84 розчинники); 6) нітрили та нітросполуки (36 розчинників); 7) розчинники, здатні до утворення водневих зв'язків (125 розчинників); 8) аміни і піридини (100 розчинників); 9) тіоли, сульфіди, сульфоксиди, тіосполуки (49 розчинників); 10) сполуки, що містять фосфор (12 розчинників); 11) розчинники з невизначеною функцією (11 розчинників).

Табл. 4.9 містить порівняльну характеристику запропонованої нами класифікації розчинників із переліченими класифікаціями. Наведено лише розчинники, присутні в масиві даних, дослідженому нами.

Таблиця 4.9

Порівняння запропонованої класифікації розчинників з іншими класифікаціями

Розчинник	Запропонована класифікація	Згідно з [15]	Згідно з [21]	Згідно з [22]	Згідно з [4]
н-пентан	1	–*	–	1	1
н-гексан	1	1	5	1	1
н-гептан	1	–	–	1	1
н-октан	1	–	–	1	1
і-октан	1	–	–	1	1
н-декан	1	–	–	1	1
н-додекан	2	–	–	1	–
н-тетрадекан	2	–	–	1	–
н-гексадекан	2	–	–	1	–
циклогексан	2	1	–	1	1
метилциклогексан	2	–	–	1	–
бензол	2	2	4	1	2
п-ксилол	2	2	–	1	2
мезитилен	2	2	–	1	2
тетрахлорметан	2	5	5	2	3
цис-декалін	3				1
трихлоретилен	2	2	–	2	5
тетрахлоретилен	3	–	–	2	1
толуол	2	2	4	1	2
о-ксилол	2	2	–	1	2
м-ксилол	2	–	–	1	2
етилбензол	2	–	–	1	–
п-цимол	2	–	–	1	–
1,4-діоксан	4	1	4	3	4
фторбензол	4	2	–	2	2
дихлорметан	4	7	4	2	5
хлороформ	11	4	4	2	4
1,2-дихлоретан	4	–	4	2	5
1,1,1-трихлоретан	4	–	–	2	5
1,1,2,2-тетрахлоретан	5	–	–	2	5
1,2,3-трихлорпропан	5	–	–	2	–
хлорбензол	5	5	4	2	2
о-дихлорбензол	5	5	–	2	2
м-дихлорбензол	6	5	–	2	2
1,2,4-трихлорбензол	6	–	–	2	–

Продовження табл. 4.9

дибромметан	6	–	–	2	–
бромоформ	6	–	–	2	–
1,2-диброметан	6	–	–	2	–
1-бромпропан	4	–	–	2	–
бромбензол	6	5	4	2	2
дийодметан	6	–	–	2	–
1-йодпропан	5	–	–	2	–
йодбензол	6	5	–	2	2
тетрагідрофуран	4	3	4	3	4
анізол	6	5	4	3	2
піридин	5	6	3	8	5
хінолін	6	–	–	8	5
дисульфід карбону	3	4	–	–	1
1-хлорпропан	8				4
метанол	7	3	2	7	3
етанол	7	3	2	7	3
1-пропанол	7	3	2	7	3
1-бутанол	7	3	2	7	3
1-пентанол	8	3	–	7	3
1-гексанол	8	–	–	7	3
1-октанол	8	5	–	7	3
1-деканол	8	–	–	7	–
1,2-етандіол	9	–	2	7	3
вода	9	8	2	7	3
N-метилформамід	9	8	–	5	5
N,N-диметилформамід	9	6	3	5	5
N,N-диметилацетамід	9	6	3	5	5
N,N-метилпіролідон	10	6	–	–	5
гексаметилфосфорамід	10	9	–	–	5
диметилсульфоксид	9	6	3	9	5
пропіленкарбонат	9				5
о-крезол	4	–	–	7	–
ацетон	7	7	3	5	5
нітрометан	9	7	3	6	5
нітроетан	9	–	–	6	5
нітробензол	10	6	3	6	5
ацетонітрил	9	7	3	6	5
бензонітрил	10	6	3	6	5

* Розчинник відсутній у масиві даних.

Аналіз відомостей, представлених в табл. 4.9, дозволяє зробити висновок, що запропонована нами класифікація розчинників у багатьох аспектах схожа з іншими класифікаційними підходами. Виділимо основні моменти:

- віднесення трихлоретилену до окремої від інших галогенозаміщених вуглеводнів групи узгоджується з класифікацією розчинників за Шастретом [15];
- склад класів IX і X отриманої нами класифікації розчинників, а також виділення спиртів в одну групу узгоджується з іншими класифікаціями;
- віднесення етерів (тетрагідрофуран, анізол, 1,4-діоксан) до однієї групи з галогенопохідними вуглеводнів, а також виділення тетрахлорметану в групу, окрему від інших галогенозаміщених вуглеводнів, узгоджується з класифікацією А. Катрицького [21];
- склад класу III, віднесення тетрахлорметану і 1-хлорпропану в окремі від інших галогенозаміщених вуглеводнів групи, а також віднесення в нашій класифікації аліфатичних і ароматичних вуглеводнів до різних груп узгоджується з класифікацією П. Граматики [4];
- подібно до класифікацій, запропонованих М. Шастретом та А. Катрицьким [15, 21, 22], наша класифікація також містить клас сполук із невизначеною функцією, що складається з хлороформу та 1-амінобутану.

Класифікацію, засновану на певному наборі дескрипторів, можна визнати раціональною, якщо вдається отримати змістовні результати при її поширенні на нові об'єкти і дані. Використані нами дескриптори включали фізико-хімічні характеристики, що використовуються для дослідження феноменів розчинності та трактування взаємодій розчинник–розчинена речовина [9, 13].

Проблема розчинності фулеренів у розчинниках різної природи привертає велику увагу дослідників (див., наприклад, недавні ґрунтовні огляди [20, 23]). На сьогодні доступні дані про розчинність фулеренів C_{60} за кімнатної температури у приблизно 150 чистих розчинниках [20]. Тлумачення цих даних ускладнене значними розбіжностями у значеннях розчинності, що наводяться у різних джерелах [20, 23]. Крім того, вважають, що розчинення C_{60} в полярних розчинниках приводить до утворення колоїдних, а не істинних розчинів [23], а в деяких випадках відбуваються хімічні реакції C_{60} з розчинниками [17, 24]. Як

результат, важко дійти однозначних тверджень щодо залежності розчинності фулерену від параметрів розчинників. Для прогнозування розчинності фулерену розроблено багато моделей QSPR, причому успішні моделі є досить складними і використовують, щонайменше, чотири або п'ять квантово-хімічних і топологічних параметрів [17, 25–27].

За цих обставин видається природним з'ясувати: чи відповідає віднесення розчинника до тієї чи іншої групи певному рівню розчинності C_{60}, чи ні? Основна ідея проста: можна припустити, що розчинники, віднесені до одного класу, мають схожі властивості, включаючи здатність розчиняти C_{60}. Значення розчинності C_{60} у різних розчинниках, виражені в мольних частках фулерену (x) у насичених при 298 K розчинах, представлені в табл. 4.8. Цей набір даних, вперше сформований М. Беком і Г. Менді [19], неодноразово використовували при дослідженні розчинності фулеренів [24, 25–27]. Судячи з даних, наведених у табл. 4.8, фулерен C_{60} *дуже слабко розчиняється* в розчинниках, віднесених до класу VII нашої класифікації ($\lg x \leq -5.9$), класу VIII ($-5.6 \leq \lg x \leq -5.0$), класу I ($-6.1 \leq \lg x \leq -4.7$) та класу X ($\lg x \approx -5 - -7$). Фулерен C_{60} є *слабкорозчинним* у розчинниках, що належать до класу IV ($-5.3 \leq \lg x \leq -4.1$ (о-крезол було виключено з розгляду як випадаючий розчинник [17]). Фулерен C_{60} є *помірно розчинним* у розчинниках, що утворюють класи II, III, VI та IX ($-4.6 \leq \lg x \leq -2.8$, циклогексан та тетрагідротіофен демонструють аномально високу здатність до розчинення; дийодметан виключено з розгляду як випадаючий розчинник [17]).

Таким чином, із 58 розчинників, для яких була доступна кількісна інформація про розчинність фулерену, здатність до розчинення лише двох розчинників не відповідає їх віднесенню до класів. Це вказує на змістовність знайденої класифікації і надає можливість напівкількісно оцінювати розчинність фулеренів у розчинниках, для яких відсутні експериментальні дані. Наприклад, можна очікувати, що мольні частки C_{60} у насичених розчинах у 1,2-етандіолу, N-метилформаміду, N,N-диметилацетаміду або пропіленкарбонату (клас X), не перевищують 10^{-5}.

Запропонована процедура класифікації поєднує привабливі риси класифікації «з навчанням» та класифікації «без навчання». Вона не вимагає апріорної інформації про кількість класів та формування навчальних вибірок. Крім того, процедура дозволяє обробляти неповні дані (набори даних з пропусками), її можна прикладати до багатопараметричних даних, уникаючи стиснення даних. Вона є перспек-

тивним засобом розв'язання некоректно поставлених задач класифікації хімічних даних.

4.3. Кластеризація зразків річкових і джерельних вод

У розділі 3.1 було проведено ідентифікацію походження зразків річкових та джерельних вод за даними про вміст у водах важких і перехідних металів. Ураховуючи, що нам відоме походження зразків вод, видавалося доцільним перевірити, наскільки процедура кластеризації об'єктів без апріорної інформації про кількість класів (розділ 4.1) здатна надати інформацію про походження зразків вод.

Відповідно до алгоритму кластеризації на основі поєднання мережі Кохонена та ймовірнісної мережі, на першому етапі застосовували мережу Кохонена, змінюючи число нейронів від трьох до семи. Для 22 зразків річкових вод виявили чотири групи (10 зразків, табл. 4.10), сформовані зі зразків, що потрапляли до відповідної групи незалежно від кількості нейронів. Для 24 зразків джерельних вод виявили п'ять таких груп (14 зразків, табл. 4.11).

Таблиця 4.10

Результати класифікації зразків річкових вод мережею Кохонена при різних кількостях нейронів (*h*)

№ зразка	Річка, рік відбору й аналізу проби	Клас					Група першої навчальної вибірки
		h = 3	*h* = 4	*h* = 5	*h* = 6	*h* = 7	
1	Немишля, 2008 р.	***3***	***3***	***3***	***1***	***6***	I
2		***3***	***3***	***3***	***1***	***6***	I
3		1	2	5	6	1	
4		3	1	2	6	6	
5		***2***	***4***	***4***	***2***	***3***	II
6	Харків, 2009 р.	1	2	5	6	6	
7		***2***	***1***	***4***	***5***	***2***	III
8		1	2	2	4	5	
9		1	1	2	4	7	
10		1	2	2	4	7	
11	Лопань, 2009 р.	2	1	2	5	2	
12		***2***	***1***	***4***	***5***	***2***	III
13		***2***	***1***	***4***	***5***	***2***	III
14		2	1	5	6	1	
15		2	4	5	6	1	
16	Уди, 2010 р.	***3***	***3***	***1***	***3***	***4***	IV
17		3	3	3	1	4	

Продовження табл. 4.10

18	Уди, 2010 р.	3	2	1	3	5	
19		***2***	***4***	***4***	***2***	***3***	II
20		2	4	3	2	6	
21		***3***	***3***	***1***	***3***	***4***	IV
22		***3***	***3***	***1***	***3***	***4***	IV

Таблиця 4.11

Результати класифікації зразків джерельних вод мережею Кохонена при різних кількостях нейронів (*h*)

№ зразка	Джерело, рік відбору й аналізу проби	Клас					Група першої навчальної вибірки
		h = 3	*h* = 4	*h* = 5	*h* = 6	*h* = 7	
1	Саржин Яр «Харківська-1», 2010 р.	***2***	***2***	***3***	***5***	***5***	I
2		3	2	3	5	5	
3		***2***	***2***	***3***	***5***	***5***	I
4		2	1	4	2	6	
5	«Харківська-2», 2010 р.	***2***	***2***	***3***	***5***	***5***	I
6		2	2	3	2	7	
7		***1***	***1***	***1***	***1***	***7***	II
8		***2***	***2***	***3***	***5***	***5***	I
9		3	3	5	4	1	
10	«Харківська-2», 2010 р.	***3***	***3***	***5***	***3***	***4***	III
11		***3***	***3***	***5***	***3***	***4***	III
12	Пантелеймонівська церква, 2010 р.	***3***	***3***	***5***	***3***	***4***	III
13		***3***	***3***	***5***	***3***	***4***	III
14		***2***	***1***	***4***	***2***	***2***	IV
15		***2***	***1***	***4***	***2***	***2***	IV
16	Завод харчових кислот, 2010 р.	***1***	***1***	***1***	***1***	***7***	II
17		3	3	4	2	6	
18		***1***	***4***	***1***	***1***	***3***	V
19	вул. Уборевича, 2009 р.	1	4	2	6	3	
20		***1***	***4***	***1***	***1***	***3***	V
21		1	4	2	4	1	
22		1	1	2	4	1	
23	Парк «Юність», 2009 р.	3	4	2	6	3	
24		1	4	1	6	3	

Із зразків, що залишилися, сформували три тестові вибірки для зразків річкових вод і дві тестові вибірки для зразків джерельних вод. Остаточну класифікацію зразків вод, одержану після виконання етапів 3–5 алгоритму кластеризації, представлено в табл. 4.12, 4.13 [28].

Таблиця 4.12

Класифікація зразків річкових вод м. Харкова

№ зразка	Річка, рік відбору і аналізу проби	Клас
1	Немишля, 2008 р.	1
2		1
3		1
4		1
5		2
6	Харків, 2009 р.	3
7		3
8		3
9		3
10		3
11	Лопань, 2009 р.	3
12		3
13		3
14		3
15		3
16	Уди, 2010 р.	4
17		4
18		4
19		2
20		2
21		4
22		4

Слід зазначити, що дві групи зразків джерельних вод були об'єднані (джерела в районі вул. Уборевича і джерела в районі парку «Юність»), оскільки остання група включала лише два зразки, що не дозволило підтвердити їх виділення в окремий клас.

Отримана класифікація зразків річкових і джерельних вод відповідає їх походженню. Зразки, відібрані з різних річок або джерел, не змішані між собою; спостерігається лише об'єднання деяких зразків, відібраних із різних річок і джерел (річки Харків і Лопань; джерела «Харківська-1», «Харківська-2» і джерело в районі Пантелеймонівської церкви), що обумовлено близькістю їх характеристик.

У випадку зразків річкових вод клас 2 включає найбільш забруднені зразки, що характеризуються найбільшим вмістом мангану, плюмбуму, кобальту, нікелю (концентрації металів у зразках значно перевищують концентрації в інших зразках).

Таблиця 4.13

Класифікація зразків джерельних вод м. Харкова

№ зразка	Джерело, рік відбору і аналізу проби	Клас
1	Саржин Яр «Харківська-1», 2010 р.	1
2		1
3		1
4		1
5	«Харківська-2», 2010 р.	1
6		1
7		1
8		1
9		1
10		1
11		1
12	Пантелеймонівська церква, 2010 р.	1
13		1
14		2
15		2
16	Завод харчових кислот, 2010 р.	3
17		3
18		3
19	вул. Уборевича, 2009 р.	4
20		4
21		4
22		2
23	Парк «Юність», 2009 р.	4
24		4

Для джерельних вод клас 2 включає найменш забруднені зразки (аналіз джерельних вод був спрямований на виявлення джерел, вода з яких найбільш придатна для вживання) з найменшими концентраціями цинку, купруму, плюмбуму і кобальту.

4.4. Кластеризація зразків харчової сировини

В розділі 3.2 описано класифікацію з навчанням зразків овочів та фруктів за вмістом у них важких і перехідних металів. Було показано, що сукупність статистичних і хемометричних процедур, включаючи класифікацію з використанням імовірнісної нейронної

мережі, дозволяє з високою надійністю пов'язати рівень забруднення харчової сировини металами з типами географічних ландшафтів.

Описану в розділі 4.1 процедуру розбиття масиву даних на групи за відсутності навчальної вибірки і без інформації про кількість груп застосували для знаходження однорідних груп зразків яблук [29].

На першому етапі процедури кількість нейронів мережі Кохонена варіювала від двох до п'яти (табл. 4.14). Для 22 зразків виявили дві групи (9 зразків, табл. 4.14), сформовані із зразків, що потрапляли у відповідну групу незалежно від кількості нейронів. Із зразків, що залишилися, сформували чотири тестові вибірки. Остаточна класифікація зразків яблук, одержана після виконання етапів 3–5 алгоритму кластеризації, представлена в табл. 4.15. Відзначимо, що на п'ятому етапі (крос-валідація) два зразки змінили свою групову приналежність: зразки 12 і 19 із другої групи перемістилися до першої.

Таблиця 4.14

Результати класифікації зразків яблук мережею Кохонена при різних кількостях заданих нейронів (*h*)

№ зразка	Тип ландшафту (рис. 3.6)	Місце відбору зразка	Клас				Група першої навчальної вибірки
			$h = 2$	$h = 3$	$h = 4$	$h = 5$	
1	10	м. Харків, Основа	2	2	1	3	
2	10	м. Харків, Жовтневий район	1	2	3	5	
3	10		1	1	2	4	
4	10	м. Харків, Ленінський район	***2***	***3***	***4***	***2***	I
5	10	м. Харків, Орджонікідзевський район	***2***	***3***	***4***	***2***	I
6	10	м. Харків, Орджонікідзевський район	2	2	1	3	
7	10	м. Харків, Жовтневий район	***1***	***1***	***2***	***1***	II
8	10		***1***	***1***	***2***	***1***	II
9	10	м. Харків, Дзержинський район	***1***	***1***	***2***	***1***	II
10	10		1	1	2	4	
11	10		1	2	1	5	

Продовження табл. 4.14

12	10	м. Харків, Жовтневий район	1	1	3	4	
13	10	м. Харків, Московський район	***2***	***3***	***4***	***2***	I
14	1	Харківський район, смт Манченки	***1***	***1***	***2***	***1***	II
15	1	Харківський район, с. Вільхівка	2	3	4	3	
16	1	Харківський район, м. Мерефа	1	1	3	4	
17	9	Харківський район, смт Кулиничі	***2***	***3***	***4***	***2***	I
18	9		1	2	3	5	
19	4	м. Чугуїв	1	1	3	4	
20	4		1	2	1	3	
21	4	Харківський район, смт Рогань	1	2	1	5	
22	4	Чугуївський район, смт Світанок	***2***	***3***	***4***	***2***	I

Таблиця 4.15

Результати класифікації тестових вибірок за допомогою ймовірнісної мережі

Тестова вибірка	№ зразка	Група
Перша	2	1
	10	2
	6	1
Друга	18	1
	12	1
	3	2
	11	1
Третя	15	1
	16	1
	19	1
Четверта	1	1
	20	1
	21	2

Отримана класифікація зразків яблук не відповідає типам ландшафтів, з яких походили зразки (рис. 4.2).

Рис. 4.2. Локалізація місць збирання зразків яблук

Ландшафт 10

Ландшафти 1+9

Ландшафт 4

Кластеризацію масиву даних про вміст металів у зразках яблук проводити значно складніше, ніж обробляти масиви даних про вміст металів у водах. Рис. 4.3–4.5 ілюструють розподіл концентрацій металів у досліджуваних зразках після автомасштабного перетворення початкових даних. Легко бачити, що концентрації металів у різних групах зразків річкових і джерельних вод змінюються в широких інтервалах, що і дозволяє правильно ідентифікувати їх географічне походження, а вміст металів у різних групах зразків яблук знаходиться практично в одних і тих самих межах, що робить будь-яке розбиття масиву даних на групи дещо штучним і значно ускладнює ідентифікацію зразків за місцем походження.

В той же час і отримана класифікація зразків яблук не позбавлена змісту: розподіл зразків яблук між групами відповідає рівню забруднення металами.

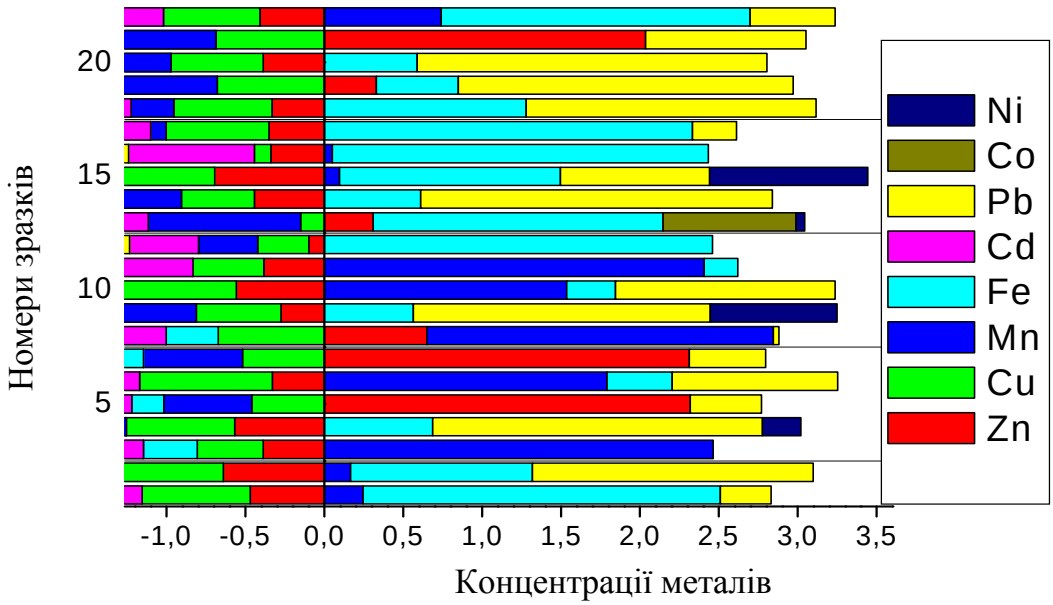

Рис. 4.3. Вміст металів у зразках річкових вод

Забруднення зразків яблук кількісно характеризували значенням критерію

$$PC = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{x_i}{\text{ГДК}_i}, \qquad (4.1)$$

де N – кількість металів, x_i – концентрація *i*-го металу, ГДК_i – гранично допустима концентрація *i*-го металу (див. табл. 3.11).

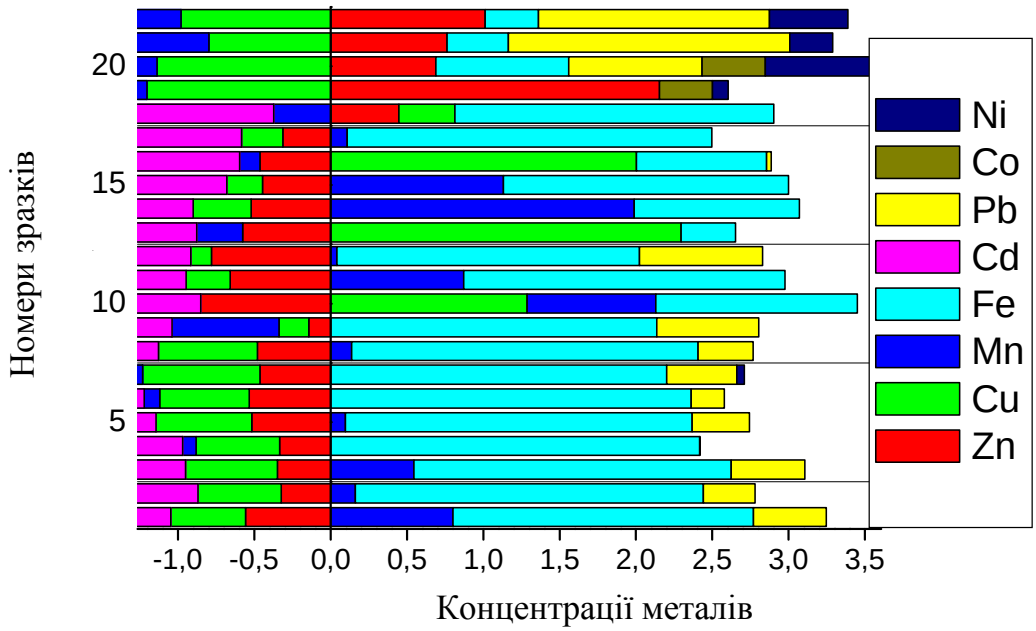

Рис. 4.4. Вміст металів у зразках джерельних вод

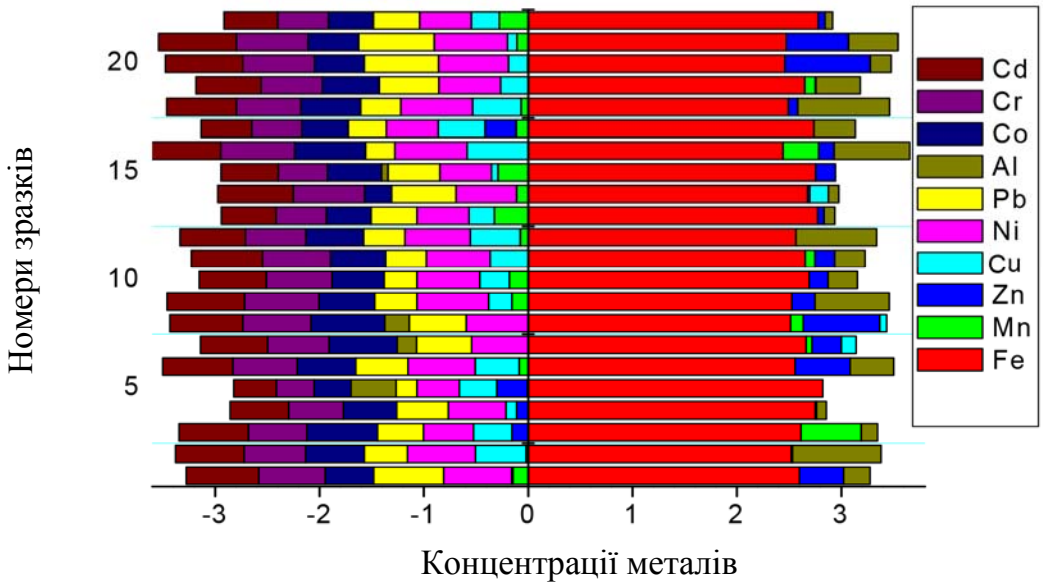

Рис. 4.5. Вміст металів у зразках яблук

Значення ГДК у знаменнику формули (4.1) для металів із невідомим значенням ГДК у фруктах (табл. 3.11) приймали за одиницю.

Перший клас містить зразки зі значеннями $PC \geq 2.07$, другий – зразки з $PC \leq 2.00$ (табл. 4.16). Значення $PC = 2$ можна розглядати як границю між зразками з прийнятним і високим рівнем забруднення металами.

Отримані результати допомагають побудувати інтерполяційну карту забруднення зразків яблук металами (рис. 4.6). Інтерполяційні карти використовують при контролі якості харчової сировини для оцінки ділянок максимального / мінімального ризику.

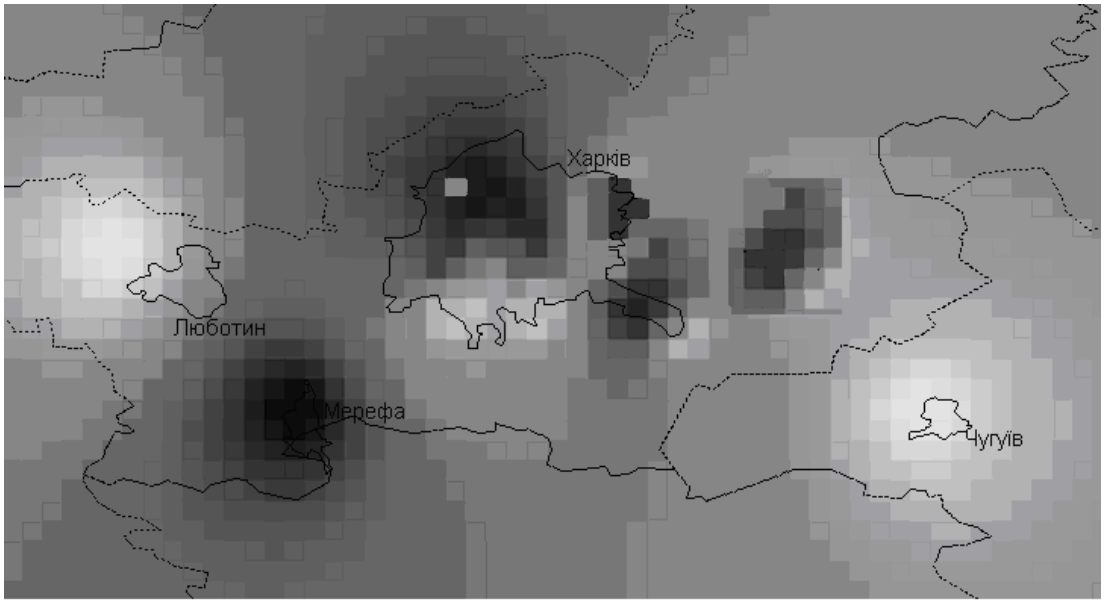

Рис. 4.6. Інтерполяційна карта забруднення зразків яблук металами (світлий колір відповідає найменш забрудненим районам, темний колір – найбільш забрудненим)

Таблиця 4.16

Остаточна класифікація зразків яблук

Тип ландшафту (рис. 3.6)	Місце відбору зразка	Клас	*PC*
10	м. Харків, Основа	1	2.21
10	м. Харків, Жовтневий район	1	2.27
10		2	1.31
10	м. Харків, Ленінський район	1	3.26
10	м. Харків, Орджонікідзевський район	1	2.72
10	м. Харків, Орджонікідзевський район	1	2.56
10	м. Харків, Жовтневий район	2	1.22
10		2	1.04
10	м. Харків, Дзержинський район	2	1.46
10		2	1.99
10		1	2.87
10	м. Харків, Жовтневий район	2	2.12
10	м. Харків, Московський район	1	3.40
1	Харківський район, смт Манченки	2	1.37
1	Харківський район, с. Вільхівка	1	2.39
1	Харківський район, м. Мерефа	1	2.28
9	Харківський район, смт Кулиничі	1	3.47
9		1	2.86
4	м. Чугуїв	2	2.00
4		1	2.07
4	Харківський район, смт Рогань	2	1.86
4	Чугуївський район, смт Світанок	1	3.26

***

Використання процедури кластеризації на основі об'єднання мережі Кохонена та ймовірнісної мережі дозволило отримати змістовні результати при класифікації зразків річкових і джерельних вод відомого походження; віднесення зразків вод до класів відповідає їх географічному походженню.

Згадана процедура поділила зразки яблук на групи, що відповідають різним вмістам важких і перехідних металів, що дозволило побудувати інтерполяційну карту забруднення яблук, корисну при контролі якості харчової сировини.

Дані, наведені в главі 4, дозволяють рекомендувати запропонований алгоритм кластеризації як ефективну процедуру експлораторного аналізу багатовимірних хімічних даних.

Література до глави 4

1. Shivaswamy P. K. Maximum relative margin and data-dependent regularization / P. K. Shivaswamy, T. Jebara, J. Mach // J. of Machine Learning Research. – 2010. – Vol. 11. – P. 747-788.
2. Wehrens R. Chemometrics with R: multivariate data analysis in the natural sciences and life sciences / R. Wehrens. – New York : Springer, 2011. – 283 p.
3. de Juan A. Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach / A. de Juan, G. Fonrodona, E. Casassas // Trends Anal. Chem. – 1997. – Vol. 16, No 1. – P. 52-62.
4. Gramatica P. Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of volecular descriptors / P. Gramatica, N. Navas, R. Todeschini // Trends Anal. Chem. – 1999. – Vol. 18, No 7. – P. 461-471.
5. Simeonov V. Lake water monitoring data assessment by multivariate statistics / V. Simeonov, P. Simeonova, S. Tsakovskii, V. Lovchinov // J. Water Resource and Protection. – 2010. – Vol. 2. – P. 353-361.
6. Skorek R. Application of ICP-MS and various computational methods for drinking water quality assessment from the Silesian District (Southern Poland) / R. Skorek, M. Jablonska, M. Polowniak, A. Kita, P. Janoska, F. Buhl // Cent. Eur. J. Chem. – 2010. – Vol. 10. – P. 71-84.
7. dos Santos J. S. Honey classification from semi-arid, atlantic and transitional forest zones in Bahia, Brazil / J. S. dos Santos, N. S. dos Santos, M. L. P. dos Santos, S. N. dos Santos, J. J. de Jesus Lacerda // J. Braz. Chem. Soc. – 2008. – Vol. 19, No 3. – P. 502-508.
8. Marcus Y. Solubilities of buckminsterfullerene and sulfur hexafluoride in various solvents / Y. Marcus // J. Phys. Chem. B. – 1997. – Vol. 101, No 42. – P. 8617–8623.
9. Reichardt C. Solvents and solvent effects in organic chemistry, 4th ed. / C. Reichardt, T. Welton – Wiley, 2011. – 692 p.
10. Barwick V. J. Strategies for solvent selection – A literature review / V. J. Barwick // Trends Anal. Chem. – 1998. – Vol. 16, No 6. – P. 293-309.
11. Dong M. Adaptive network-based fuzzy inference system with leave-one-out cross-validation approach for prediction of surface roughness / M. Dong, N. Wang // Appl. Math. Model. – 2011. – Vol. 35. – P. 1024-1035.
12. Wong T.-T. Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation / T.-T. Wong // Pattern Recognition. – 2015. – Vol. 48, No 9. – P. 2839-2846.

13. Marcus Y. The properties of solvents / Y. Marcus. – Wiley, 1999. – 399 p.
14. Snyder L. R. Classification of the solvent properties of common liquids / L. R. Snyder // J. Chromatogr. – 1974. – Vol. 92. – P. 223-230.
15. Chastrette M. Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters / M. Chastrette, M. Rajzmann, M. Chanon, K. F. Purcell, // J. Am. Chem. Soc. – 1985. – Vol. 107, No 1. – P. 1-11.
16. Pushkarova Y. A procedure for meaningful unsupervised clustering and its application for solvent classification / Y. Pushkarova, Y. Kholin // Cent. Eur. J. Chem. – 2014. – Vol. 12, No 5. – P. 594-603.
17. Liu H. Accurate quantitative structure−property relationship model to predict the solubility of C_{60} in various solvents based on a novel approach using a least-squares support vector machine / H. Liu, X. Yao, R. Zhang, M. Liu, Z. Hu, B. Fan // J. Phys. Chem. B. – 2005. – Vol. 109, No 43. – P. 20565-20571.
18. Ruoff R. S. Solubility of fullerene (C_{60}) in a variety of solvents / R. S. Ruoff, D. S. Tse, R. Malhotra, D. C. Lorents // J. Phys. Chem. – 1993. – Vol. 97, No 13. – P. 3379-3383.
19. Beck M. T. Solubility of C_{60} / M. T. Beck, G. Mandi // Fullerenes, Nanotubes, and Carbon Nanostructures. – 1997. – Vol. 5, No 2. – P. 291-310.
20. Semenov K. N. Temperature dependence of solubility of individual light fullerenes and industrial fullerene mixture in 1-chloronaphthalene and 1-bromonaphthalene / K. N. Semenov, N. A. Charykov, V. A. Keskinov, A. K. Piartman, A. A. Blokhin, A. A. Kopyrin // J. Chem. Eng. Data. – 2010. – Vol. 55, No 7. – P. 2373-2378.
21. Katritzky A. R. A unified treatment of solvent properties / A. R. Katritzky, T. Tamm, Y. Wang, M. Karelson // J. Chem. Inf. Comput. Sci. – 1999. – Vol. 39. – P. 692-698.
22. Katritzky A. R. The classification of solvents by combining classical QSPR methodology with principal component analysis / A. R. Katritzky, D. C. Fara, M. Kuanar, E. Hur, M. Karelson // J. Phys. Chem. A. – 2005. – Vol. 109, No 45. – P. 10323-10341.
23. Mchedlov-Petrossyan N. O. Fullerenes in liquid media: an unsettling intrusion into the solution chemistry / N. O. Mchedlov-Petrossyan // Chem. Rev. – 2013. – Vol. 113, No 7. – P. 5149-5193.
24. Kiss I. Z. Artificial neural network approach to predict the solubility of C-60 in various solvents / I. Z. Kiss, G. Mandi, M. T. Beck // J. Phys. Chem. A. – 2000. – Vol. 104, No 34. – P. 8081-8088.
25. Gharagheizi F. A molecular-based model for prediction of solubility of C(60) fullerene in various solvents / F. Gharagheizi, R. F. Alamdari // Fullerenes, Nanotubes and Carbon Nanostructures. – 2008. – Vol. 16, No 1. – P. 40-57.

26. Petrova T. Improved model for fullerene C_{60} solubility in organic solvents based on quantum-chemical and topological descriptors / T. Petrova, B. F. Rasulev, A. A. Toropov, D. Leszczynska, J. Leszczynski // J. Nanopart. Res. – 2011. – Vol. 13, No 8. – P. 3235-3247.
27. Yousefinejad S. New LSER model based on solvent empirical parameters for the prediction and description of the solubility of buckminsterfullerene in various solvents / S. Yousefinejad, F. Honarasa, F. Abbasitabar, Z. Arianezhad // J. Sol. Chem. – 2013. – Vol. 42, No 8. – P. 1620-1632.
28. Пушкарева Я. Н. Классификация химико-аналитических данных на основе объединения нейронной сети Кохонена и вероятностной нейронной сети / Я. Н. Пушкарева, Н. П. Титова, О. И. Юрченко, Ю. В. Холин // Вісник Харківського нац. ун-ту. – 2012. – № 1026. Хімія. Вип. 21 (44). – С. 212-217.
29. Пушкарева Я. Н. Особенности идентификации географического происхождения овощей и фруктов с помощью хемометрических и статистических методов / Я. Н. Пушкарева, А. Б. Следзевская, П. В. Семибратова, А. Г. Гарбуз, А. Н. Некос, Ю. В. Холин // Методы и объекты химического анализа. – 2012. – Т. 7, No 4. – С. 184-191.

Глава 5

ІДЕНТИФІКАЦІЯ ОБ'ЄКТІВ В ЯКІСНОМУ ХІМІЧНОМУ АНАЛІЗІ. ПІДХІД НА ОСНОВІ ТЕОРІЇ НЕЧІТКИХ МНОЖИН

Один з основних підходів до ідентифікації заснований на оцінюванні близькості характеристик аналіту *a* й еталона *e* [1–7]. Характеристиками можуть бути, наприклад, час утримування в хроматографії, положення або максимуми смуг поглинання, хімічні зсуви в спектрах ЯМР, результати роботи мультисенсорних систем «електронний ніс» і «електронний язик» [8–13]. Рішення «аналіт збігається з еталоном» або «аналіт відрізняється від еталона» приймають, порівнюючи кількісну міру подібності еталона й аналіту з її критичним значенням. Мірами подібності можуть виступати, наприклад, різні види відстаней між аналітом й еталоном або індекси подібності (відстані Мінковського при різних метриках, Махаланобіса, індекси Танімото) [14, 15].

У припущенні, що характеристики незалежні, а розподіл їх похибок описує закон Гауса з нульовим математичним очікуванням і відомими дисперсіями, найпростішим способом перевірки гіпотези про збіг (в межах похибок вимірювань) характеристик еталона й аналіту є дослідження статистики

$$\chi^2_{\text{експ}} = \sum_{i=1}^{N} \xi_i^2 , \qquad (5.1)$$

де N – загальне число характеристик, ξ_i – зважені нев'язки

$$\xi_i = w_i^{1/2}(a_i - e_i),\ i = 1, 2, \ldots N, \tag{5.2}$$

статистичні ваги

$$w_i = \frac{1}{s^2(a_i)},\ i = 1, 2, \ldots N, \tag{5.3}$$

$s^2(a_i)$ – дисперсія *i*-ї характеристики аналіту, яку потрібно або експериментально визначити по розсіянню результатів повторних спостережень, або задати, виходячи з особливостей проведення експерименту і характеристик вимірювальної системи. При рівні значущості α аналіт ототожнюють з еталоном, якщо виконується нерівність

$$\chi^2_{\text{експ}} < \chi^2_{N,\alpha}, \tag{5.4}$$

де $\chi^2_{N,\alpha}$ – 100α-відсоткова точка розподілу χ^2 для N ступенів свободи.

При ідентифікації можливі два типи помилок – помилковий висновок про відмінність аналіту від еталона або їх неправомірне ототожнення. Зазвичай більш небезпечні за наслідками помилки першого типу (наприклад, неототожнення аналіту з небезпечним токсикантом), і їх за традицією розглядають як помилки I роду [1]. Надійність, ключову метрологічну характеристику процедури ідентифікації [16, 17], доречно пов'язувати з малою ймовірністю помилок I роду.

Як і при традиційній перевірці статистичних гіпотез, розрахувати ймовірності помилок першого і другого роду можна лише тоді, коли відомі густини розподілу похибок характеристик аналіту й еталона. Така інформація доступна тільки при обробці модельних (імітованих) даних. Тому надійність ідентифікації оцінюють за допомогою двох підходів – статистичного та апріорного [3]. У першому знаходять відсоток помилок при ідентифікації аналітів, про які наперед відомо, що вони або збігаються з еталонами, або відрізняються від них. У другому для розрахунку ймовірності помилок першого і другого роду використовують апріорні гіпотези про характеристики аналіту й еталона. Вимогам до підвищення надійності ідентифікації приділено особливу увагу у відомій монографії [3].

З відсутністю розгорнутої інформації про похибки характеристик аналіту й еталона пов'язані не лише труднощі в оцінці

надійності ідентифікації. Важливо також, що при всьому різноманітті використовуваних критеріїв подібності за відсутності згаданої інформації теоретично обґрунтувати вибір того чи іншого критерію подібності як найбільш потужного для вирішення конкретного ідентифікаційного завдання неможливо. В силу цього особливий інтерес викликають критерії, які припускають використання розгорнутої інформації про густину розподілу похибок експериментальних характеристик аналіту й еталона.

Перспективною видається розробка критеріїв, заснованих на застосуванні теорії нечітких множин (fuzzy sets theory). В основі теорії нечітких множин лежить поняття суб'єктивної ймовірності («possibility»), відмінне від статистичної ймовірності («probability»).

Підхід до ідентифікації, заснований на теорії нечітких множин, можна проілюструвати так. Нехай функціональну групу в молекулі аналіту ідентифікують, порівнюючи експериментально знайдений максимум смуги поглинання з еталонним значенням, наведеним в атласі спектрів (базі даних). Оскільки ці величини в точності майже ніколи не збігаються, замість точкового значення еталонної характеристики задають деякий інтервал допустимих значень. Групу вважають ідентифікованою, якщо експериментальне значення потрапляє в зазначений інтервал, і відсутньою, якщо воно лежить поза інтервалом. Цей дихотомічний результат викликає деяке замішання, адже виявляється неважливим, наскільки велика різниця між еталонним і виміряним значеннями, якщо останнє потрапляє в допустимий інтервал. Більш змістовну інформацію дозволяє отримати залучення ключового поняття теорії нечітких множин – функції приналежності. Функцію приналежності характеристик аналіту й еталона визначають так, щоб у разі повного збігу виміряного значення з еталонним вона приймала значення 1, а в міру наближення експериментального значення до межі допустимого інтервалу знижувалася до 0. Ймовірність правильної ідентифікації (надійність) тим більше, чим вище значення функції приналежності.

Починаючи з піонерських робіт [18, 19], теорія нечітких множин успішно застосовується для обробки результатів хімічного аналізу, в тому числі якісного [20–26]. Навіть наявність в даних «грубих промахів» (outliers) [27] не перешкоджає побудові на основі теорії нечітких множин робастних алгоритмів, стійких до відхилення закону розподілу експериментальних похибок від нормального [28–31].

У даній частині роботи запропоновано та протестовано алгоритм ідентифікації аналітів, який заснований на аналізі багатовимірних масивів даних і використовує підходи теорії нечітких множин. Переваги алгоритму – низька чутливість до наявності в даних грубих промахів і мінімальні вимоги до апріорної інформації про властивості результатів вимірювань.

5.1. Нечіткі критерії подібності

Значення характеристик аналіту й еталона розглядали як одномірні нечіткі числа [32]. Уявити результат вимірювання у вигляді нечіткого числа можна за допомогою процедури розмиття (фазифікації) [26]. Фазифікувати результат вимірювання p – значить визначити на інтервалі таку безперервну функцію приналежності $\mu(x)$, яка приймає максимальне значення в точці p ($\mu(p) = 1$) і рівномірно убуває до $\mu = 0$ в точках $p \pm 0.5b$ (зауважимо, що вимога нормування до функції $\mu(x)$ не пред'являється). Параметр b будемо називати допустимим розмахом даних. Його значення слід обирати, виходячи з особливостей проведення експерименту і характеристик вимірювальної системи. Параметр b не можна ототожнювати з параметрами розмаху статистичних розподілів. Найпростіше уявлення результату вимірювання p у вигляді нечіткого числа представлено на рис. 5.1.

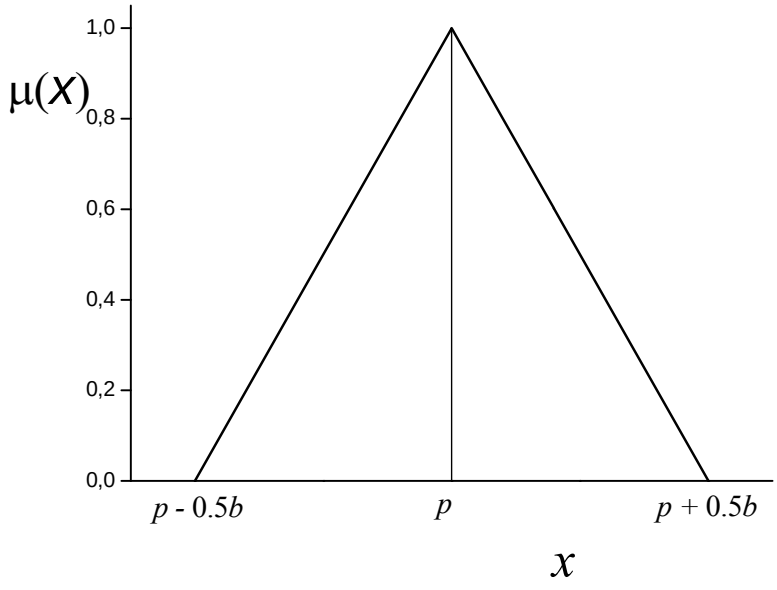

Рис. 5.1. Можливе представлення результату вимірювання p у вигляді нечіткого числа з допустимим розмахом даних b

Провівши процедуру фазифікації характеристик аналіту й еталона, можна оцінити ступінь їх приналежності одній множині, використовуючи таку характеристику, як перетин двох нечітких чисел

[18, 33]. На рис. 5.2 вимірювання p_1 і p_2 належать одній множині зі ступенем приналежності 0.25.

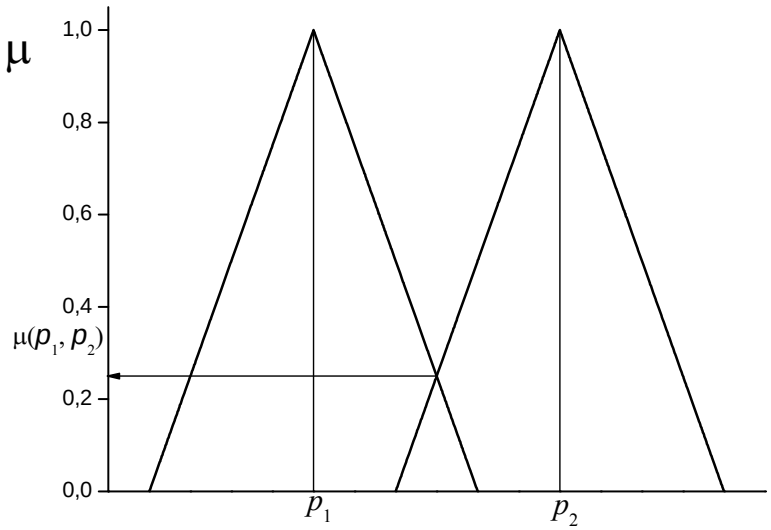

Рис. 5.2. Приналежність вимірювань p_1 і p_2 одній множині

Подання характеристик аналіту й еталона як нечітких чисел можна використовувати для виконання процедури ідентифікації.

Пропонується наступний алгоритм ідентифікації аналітів.

1. Задати допустимий розмах даних b.

2. Фазифікувати значення характеристик аналіту й еталона (p), задавши функції приналежності μ_i, $i = 1, 2, ..., N$. Визначали функції приналежності чотирьох типів (табл. 5.1).

3. Для i-ї характеристики еталона і відповідної характеристики аналіту обчислити функції їх приналежності одній множині (μ_{ae}^{i}) як перетин нечітких чисел.

Таблиця 5.1

Типи функцій приналежності, які були використані в роботі

Тип функції приналежності	Формула для розрахунку
Сімпсона	$\mu^S(x) = 1 - \frac{2}{b}\lvert p - x\rvert$
Квадратична	$\mu^Q(x) = a_0(x-p)^2 + a_1\lvert x-p\rvert + a_2$
Гауса	$\mu^G(x) = \exp\left(-0.5\left[\frac{x-p}{\sigma^G}\right]^2\right)$
Лапласа	$\mu^L(x) = \exp\left(-\left\lvert\frac{x-p}{\sigma^L}\right\rvert\right)$

4. Знайти сумарну потужність множини функцій приналежності:

$$\mu_{\text{sum}} = \frac{1}{N}\sum_{i=1}^{N}\mu_{ae}^{i}. \qquad (5.5)$$

5. Порівняти значення критерію μ_{sum} з критичним значенням μ_α і зробити висновок про ідентифікацію аналіту або про його відмінність від еталона.

5.2. Розрахунок параметрів функцій приналежності

Параметри трикутної (Сімпсона) і параболічної (квадратичної) функцій приналежності розраховували за допомогою елементарних алгебраїчних перетворень.

Використовували також дві дзвоноподібні функції приналежності – Гаусова і Лапласова типів. Визначаючи параметр розмаху σ^G функції приналежності Гаусова типу, ґрунтувалися на таких міркуваннях. Нехай вимірювання *p* має функцію приналежності, описувану густиною розподілу Гауса (без накладення умови нормування). Для визначення реперної точки приймемо, що можливість появи результату вимірювання *x* поза інтервалом [*p*-0.5*b*; *p*+0.5*b*] дорівнює 0.003. Якщо результати вимірювань насправді підкоряються закону Гауса, це значення відповідає шести стандартним відхиленням (ширина інтервалу [*p*-0.5*b*; *p*+0.5*b*] дорівнює $6\sigma^G$). Таким чином,

$$\sigma^G = \frac{b}{6}. \qquad (5.6)$$

При розрахунку параметра розмаху функції приналежності Лапласова типу знаходили такі значення σ^L, при яких значення ординат функції приналежності μ^L збігаються з координатами функції μ^G в точках *p*-0.5*b* і *p*+0.5*b*; параметр

$$\sigma^L = \frac{b}{9}. \qquad (5.7)$$

Різниця параметрів розмаху функцій приналежності Гаусова і Лапласова типів пояснюється особливостями функцій – використанням квадрата різниці «математичне очікування – результат вимірювання» і модуля зазначеної різниці. Загальний вигляд функцій приналежності показаний на рис. 5.3.

Надійність ототожнення аналіту й еталона приймали рівною 100·(1–α)% при виконанні нерівності

$$\mu_{sum} > \mu_{\alpha}, \qquad (5.8)$$

де α – рівень значущості (ймовірність помилкового висновку про відмінність еталона й аналіту). Очевидно, що $\mu_0 = 0$, $\mu_1 = 1$. Інші значення μ_α оцінювали евристично. Розглядали унімодальні симетричні функції приналежності μ (*x*) і використовували аналогію між ними і густинами ймовірності. Для заданої ймовірності α знаходили таке значення *r*, для якого для випадкової величини *x* ймовірність $p\left(x \in \left[p - 0.5b, p\ -\ r\right]\right) = \alpha$, де *p* – математичне очікування *x* (при цьому значенні функція приналежності дорівнює 1). Вказану ймовірність оцінювали як відношення A_r / A, де A_r – площа кривої функції приналежності на інтервалі, A – площа під всією кривою функції приналежності. Як μ_α брали значення ординати функції приналежності в точці *(p - r)* (табл. 5.2). Процедуру визначення μ_α ілюструє рис. 5.4.

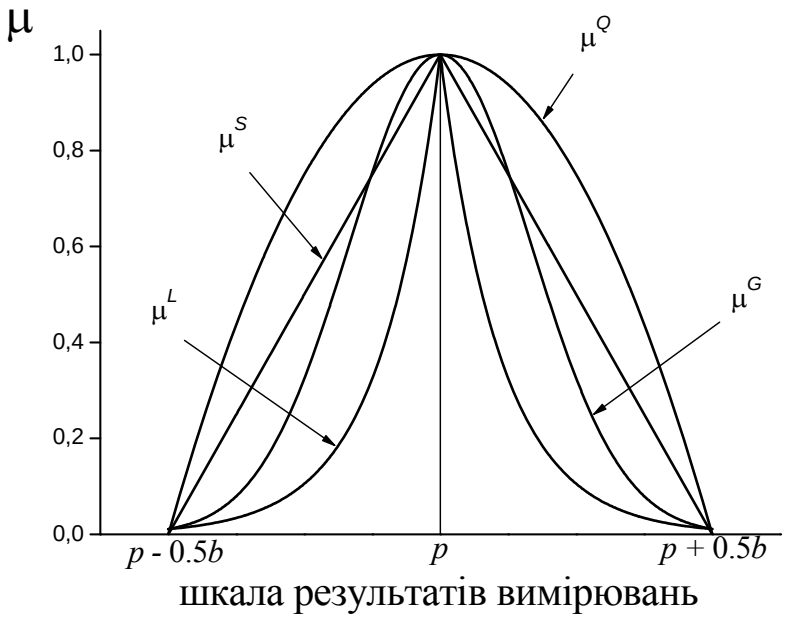

Рис. 5.3. Загальний вигляд функцій приналежності

Рис. 5.4. До визначення критичного значення нечіткого критерію ідентифікації

Таблиця 5.2

Критичні значення нечітких критеріїв ідентифікації

α	Функція приналежності			
	S	*Q*	*G*	*L*
0.10	0.46	0.64	0.45	0.22
0.05	0.32	0.48	0.27	0.12
0.01	0.16	0.23	0.09	0.03

5.3. Випробування алгоритму ідентифікації

На першому етапі властивості пропонованого алгоритму ідентифікації досліджували за допомогою імітаційного моделювання. Перевіряли стійкість результатів до наявності у вихідних даних похибок, що мають густину розподілу з хвостами, довшими, ніж у нормального розподілу. В роботі [34] наведені положення максимумів 14 смуг поглинання трет-бутил-2-[[(2,2,2-трихлоретанімідоїл)окси]метил]акрилату в ІЧ-області. Їх розглядали як характеристики еталона *e* (табл. 5.3). В ці характеристики вносили похибки, згенеровані за моделлю грубих промахів [35]:

$$\varepsilon = \left[\left(100 - \Delta\right) \cdot \varepsilon_G + \Delta \cdot \varepsilon_L \right] / 100, \qquad (5.9)$$

де Δ – інтенсивність грубих промахів (Δ міняли від 0 до 100% з кроком 25%), ε_G – похибки, розподілені за законом Гауса з нульовим середнім і стандартним відхиленням σ^G, ε_L – похибки, розподілені за законом Лапласа з нульовим середнім і стандартним відхиленням σ^L. Виходячи з даних табл. 5.3, припускали, що допустимий розмах положення смуг поглинання можна прийняти рівним 1 см$^{-1}$ (смуги 7, 10 і 11 визначено найменш точно), тоді σ^G = (1/6) см$^{-1}$, а σ^L = (1/9) см$^{-1}$. Генерацію випадкових чисел, розподілених відповідно до законів Гауса і Лапласа, виконували за допомогою методу зворотних функцій [36].

Характеристики аналітів розраховували як $a_i = e_i + \varepsilon_i$. Отримали a^Δ, Δ = 0, 25, 50, 75, 100%. Всім значенням *a* приписали однакові стандартні відхилення $s(a_i)$ = (1/6) см$^{-1}$. Результати ідентифікації аналітів за наборами a^Δ (табл. 5.4) показують, що підхід, заснований на використанні статистики χ^2, не ефективний для ототожнення аналітів з еталоном при наявності в результатах вимірювань «грубих промахів», тоді як нечіткі критерії подібності стійкі до відхилення закону розподілу експериментальних похибок від нормального.

Таблиця 5.3

Положення максимумів смуг поглинання в ІЧ-спектрі трет-бутил-2-[[(2,2,2-трихлоретанімідоїл)окси]метил]акрилату

№ смуги поглинання	Максимум смуги поглинання, $см^{-1}$	Віднесення
1	3348.7	ν_{NH}
2	1738.5	$\nu_{C=O}$
3	1732.2	
4	1727.9	
5	1724	
6	1719.7	
7	1712	
8	1707.3	$\nu_{C=O}$
9	1702.3	
10	1697	
11	1674	$\nu_{C=N}$
12	1669.7	
13	1664.7	
14	1645.2	$\nu_{C=C}$

Таблиця 5.4

Ототожнення характеристик аналітів (*a*) і еталона (*e*). Жирним шрифтом виділені результати, що задовольняють умовам (5.4) або (5.8)

Критерій	Критичне значення*	Набори характеристик				
		$(a^0; e)$	$(a^{25}; e)$	$(a^{50}; e)$	$(a^{75}; e)$	$(a^{100}; e)$
$\chi^2_{\text{експ}}$	23.7	**13.3**	**19.8**	27.5	36.7	47.1
μ^S_{sum}	0.32	**0.87**	**0.84**	**0.82**	**0.79**	**0.77**
μ^Q_{sum}	0.48	**0.97**	**0.96**	**0.95**	**0.93**	**0.91**
μ^G_{sum}	0.27	**0.90**	**0.86**	**0.82**	**0.78**	**0.75**
μ^L_{sum}	0.12	**0.60**	**0.56**	**0.53**	**0.50**	**0.48**

* Тут і в табл. 5.5, 5.6 для статистики $\chi^2_{\text{експ}}$ наводяться критичні значення $\chi^2_{N,0.05}$ і $\mu_{0.05}$ в решті випадків.

В роботі [37] виміряні спектри комбінаційного розсіювання (КР) сумішей води і етанолу при різних температурах. З масиву отриманих даних [38] виділили результати вимірювання інтенсивностей поглинання розчинів (I) етанолу при 30, 50 і 70 °C (E30, E50, E70) і води

при 30 °С (W30). Масив даних містив вимірювання при 200 довжинах хвиль – від 850 до 1049 нм – з кроком 1 нм. Значення I змінювалися в широкому інтервалі – від $2 \cdot 10^6$ до $2 \cdot 10^1$.

Таблиця 5.5

Результати ототожнення етанолу з водою за даними спектрів КР

Критерій	Критичне значення	Набори характеристик		
		(E30; W30)	(E50; W30)	(E70; W30)
$\chi^2_{\text{експ}}$	234	$1.1 \cdot 10^7$	$9.6 \cdot 10^6$	$4.4 \cdot 10^6$
μ^S_{sum}	0.32	0.02	0.02	0.02
μ^Q_{sum}	0.48	0.03	0.03	0.03
μ^G_{sum}	0.27	0	0	0
μ^L_{sum}	0.12	0	0	0

Таблиця 5.6

Результати ототожнення етанолу за спектрами КР, виміряними за різних температур. Жирним шрифтом виділені результати, що задовольняють умові (5.8)

Критерій	Критичне значення	Набори характеристик		
		(E30; E50)	(E30; E70)	(E50; E70)
$\chi^2_{\text{експ}}$	234	$3.0 \cdot 10^4$	$9.4 \cdot 10^4$	$1.1 \cdot 10^4$
μ^S_{sum}	0.32	**0.48**	**0.41**	**0.43**
μ^Q_{sum}	0.48	**0.57**	**0.54**	**0.57**
μ^G_{sum}	0.27	0	0	0
μ^L_{sum}	0.12	0	0	0

Розв'язували два завдання ідентифікації. У першому характеристики води вважали еталонними і перевіряли можливість ототожнення води з етанолом на основі порівняння масиву даних W30 з масивами E30, E50 і E70. У другому характеристики E30 розглядали як еталонні і перевіряли можливість ототожнити з ними дані E50 і E70. Для кожного виміряного значення інтенсивності I_i, $i = 1, 2, ..., N$, брали допустимий розмах даних $b_i = 0.1 \cdot I_i$, а при обчисленні статистики χ^2

стандартні відхилення задавали як $s(I_i) = 0.1 \cdot I_i / 6$. Результати розрахунків показують, що всі використані в роботі підходи до ідентифікації забезпечують розрізнення води і етанолу за характеристиками спектрів комбінаційного розсіювання (табл. 5.5). Ототожнення спектрів етанолу, виміряних при різних температурах, забезпечує використання нечітких критеріїв з трикутною і параболічною функціями приналежності (табл. 5.6). Вбачається, що саме цими критеріями слід користуватися в разі, якщо помилка I роду особливо небажана.

***

Запропонований алгоритм ідентифікації аналітів заснований на аналізі багатовідгукових масивів даних, що використовує підходи теорії нечітких множин. Переваги алгоритму – стійкість до наявності в даних грубих промахів і мінімальні вимоги до апріорної інформації про статистичні властивості результатів вимірювань (необхідно вказувати лише допустимий розмах даних для аналіту й еталона). Використання трикутної і квадратичної функцій приналежності в процедурі ідентифікації найбільш доцільно у випадках, коли помилковий висновок про відмінність аналіту від еталона особливо небезпечний.

Література до глави 5

1. Milman B.L. Identification of chemical compounds / B. L. Milman // Trends Anal. Chem. – 2005. – Vol. 24, No 6. – P. 493-508.
2. Cárdenas S. Analytical features in qualitative analysis / S. Cárdenas, M. Valcárcel // Trends Anal. Chem. – 2005. – Vol. 24, No 6. – P. 477-487.
3. Вершинин В. И. Компьютерная идентификация органических соединений / В. И. Вершинин, Б. Г. Дерендяев, К. С. Лебедев. – М. : Академкнига, – 2002. – 197 с.
4. Vlasov Yu. Nonspecific sensor arrays («electronic tongue») for chemical analysis of liquids (IUPAC Technical Report) / Yu. Vlasov, A. Legin, A. Rudnitskaya, C. Di Natale, A. D'Amico // Pure Appl. Chem. – 2005. – Vol. 77, No 11. – P. 1965-1983.
5. Вершинин В. И. Критерии совпадения пиков в качественном хроматографическом анализе / В. И. Вершинин, В. А. Топчий, И. И. Медведовская // Журнал аналитической химии. – 2001. – Т. 56, №4. – С. 367-373.

6. Соколова О. В. Достоверность компьютерной идентификации углеводородов при хроматографическом анализе бензинов / О. В. Соколова, Н. Б. Ильичева, В. И. Вершинин // Аналитика и контроль. – 2000. – Т. 4. – С. 363-369.
7. Vershinin V. I. Methodology of computer-assisted identification of substances using information retrieval systems / V. I. Vershinin // Journal of Analytical Chemistry. – 2000. – Vol. 55, No 5. – P. 417-425.
8. Oliveri P. Development of a voltammetric electronic tongue for discrimination of edible oils / P. Oliveri, M. A. Baldo, S. Daniele, M. Forina // Anal. Bioanal. Chem. – 2009. – Vol. 395. – P. 1135-1143.
9. Cotte J. F. Chromatographic analysis of sugars applied to the characterisation of monofloral honey / J. F. Cotte, H. Casabianca, S. Chardon, J. Lheritier, M. F. Grenier-Loustalot // Anal. Bioanal. Chem. – 2004. – Vol. 380. – P. 698-705.
10. Boffo E. F. Classification of Brazilian vinegars according to their 1H NMR spectra by pattern recognition analysis / E. F. Boffo, L. A. Tavares, M. M. C. Ferreira, A. G. Ferreira // LWT – Food Sci. and Technol. – 2009. – Vol. 42, No 9. – P. 1455-1460.
11. Bellorini S. Discriminating animal fats and their origins: assessing the potentials of Fourier transform infrared spectroscopy, gas chromatography, immunoassay and polymerase chain reaction techniques / S. Strathmann, V. Baeten, O. Fumiere, G. Berben, S. Tirendi, C. von Holst // Anal. Bioanal. Chem. – 2005. – Vol. 382. – P. 1073-1083.
12. Li H. A chemometrics approach for distinguishing between beers using near infrared spectroscopy / H. Li, Y. Takahashi, M. Kumagai, K. Fujiwara, R. Kikuchi, N. Yoshimura, T. Amano, N. Ogawa // J. of Near Infrared Spectrosc. – 2009. – Vol. 17, No 2. – P. 69-76.
13. Shin Y.-S. Fingerprinting analysis of fresh ginseng roots of different ages using1H-NMR spectroscopy and principal components analysis / Y.-S. Shin, K.-H. Bang, D.-S. In, O.-T. Kim, D.-Y. Hyun, I.-O. Ahn, C. K. Bon, H. K. Choi // Archives of Pharm. Research. – 2007. – Vol. 30, No 12. – P. 1625-1628.
14. Urbano Cuadrado M. Study of spectral analytical data using fingerprints and scaled similarity measurements / M. Urbano Cuadrado, M. Luque de Castro, M. D. Gómez-Nieto / Anal. Bioanal. Chem. – 2005. – Vol. 381. – P. 953-963.
15. Saito S. Discovery of Chemical Compound Groups with Common Structures by a Network Analysis Approach (Affinity Prediction Method) / S. Saito, T. Hirokawa, K. Horimoto // J. Chem. Inf. Model. – 2011. – Vol. 51. – P. 61-68.
16. Руководство ЕВРАХИМ / СИТАК. Количественное описание неопределенности в аналитических измерениях. 2-е издание. Пер. с англ.

Р. Л. Кадиса, Г. Р. Нежиховского, В. Б. Симина / Под ред. Л. А. Конопелько. – СПб. : ВНИИМ им. Д. И. Менделеева, 2002. – 149 с.
17. Мильман Б. Л. Неопределенность результатов качественного химического анализа. Общие положения и бинарные тест-методы / Б. Л. Мильман, Л. А. Конопелько // Журн. аналит. химии. – 2004. – Т. 59, № 12. – С. 1244-1258.
18. Bandemer H. Fuzzy theory in analytical chemistry / H. Bandemer, M. Otto // Mikrochim. Acta. –1986. – Vol. 89. – P. 93-124.
19. Назаренко А. Ю. Применение теории нечетких множеств для обработки результатов анализа / А. Ю. Назаренко, В. В. Сухан, Н. А. Назаренко / Заводская лаборатория. – 1991. – Т. 57, № 10. – С. 63-65.
20. Linusson A. Statistical molecular design of peptoid libraries / A. Linusson, S. Wold, B. Norden // Chemom. Intell. Lab. Sys. – 1998. – Vol. 44. – P. 213-227.
21. Barbieri P. Robust cluster analysis for detecting physico-chemical typologies of freshwater from wells of the plain of Friuli (northeastern Italy) / P. Barbieri, G. Adami, A. Favretto, A. Lutman, W. Avoscan, E. Reisenhofer // Anal. Chim. Acta. – 2001. – Vol. 440, No 2. – P. 161-170.
22. Sârbu C. Principal component analysis versus fuzzy principal component analysis: A case study: the quality of danube water (1985–1996) / C. Sârbu, H. F. Pop // Talanta. – 2005. – Vol. 65. – P. 1215-1220.
23. Sârbu C. Fuzzy clustering analysis of the first 10 MEIC chemicals / C. Sârbu, H. F. Pop // Chemosphere. – 2000. – Vol. 40. – P. 513-520.
24. Iliev B. A fuzzy technique for food- and water quality assessment with an electronic tongue / B. Iliev, M. Lindquist, L. Robertsson, P. Wide // Fuzzy Sets and Systems. – 2006. – Vol. 157. – P. 1155-1168.
25. Musee N. New methodology for hazardous waste classification using fuzzy set theory: Part I. Knowledge acquisition / N. Musee, L. Lorenzen, C. Aldrich / J. Hazard. Materials. – 2008. – Vol. 154. – P. 1040-1051.
26. Холин Ю. В. Метрологические характеристики методик обнаружения с бинарным откликом / Ю. В. Холин, Н. А. Никитина, А. В. Пантелеймонов, Е. А. Решетняк, А. А. Бугаевский, Л. П. Логинова. – Х. : Тимченко, 2008. – 128 с.
27. Huber P. J. Robust statistics / P. J. Huber. – New York : John Wiley and Sons, 1981. – P. 308.
28. Pop H. F. A New Fuzzy Regression Algorithm / H. F. Pop, C. Sârbu // Anal. Chem. – 1996. – Vol. 68, No 5. – P. 771-778.
29. Sârbu C. Fuzzy robust estimation of central location / C. Sârbu, H. F. Pop // Talanta. – 2001. – Vol. 54. – P. 125-130.
30. Changa Y.-H. O. Fuzzy regression methods – a comparative assessment / Y.-H. O. Changa, B. M. Ayyub // Fuzzy Sets and Systems. – 2001. – Vol. 119. – P. 187-203.

31. Nasrabadi E. An LP-Based Approach to Outliers Detection in Fuzzy Regression Analysis / E. Nasrabadi, S. Mehdi Hashemi, A. N. Mehdi Ghatee / Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems. – 2007. – Vol. 15, No 4. – P. 441-456.
32. Заде Л. А. Размытые множества и их применение в распознавании образов и кластер-анализе / Л. А. Заде // Классификация и кластер. – М. : Мир, 1980. – 389 с. (С. 208-247)
33. Орловский С. А. Проблемы принятия решений при нечеткой исходной информации / С. А. Орловский. – М. : Наука. Главная редакция физ.-мат. литературы, 1981. – 208 с.
34. Conti C. FT-IR of trichloroacetoimidates in different solvent systems / C. Conti, R. Galeazzi, E. Giorgini, G. Tosi // J. Mol. Struct. – 2005. – Vol. 744–747. – P. 417-423.

Додатки

Таблиця Д.1

Характеристики зразків квітів ірису [54]

№ зразка	Довжина чашолистка, см	Ширина чашолистка, см	Довжина пелюстки, см	Ширина пелюстки, см	Вид
1	5.1	3.5	1.4	0.2	Ірис щетинистий
2	4.9	3.0	1.4	0.2	-//-
3	4.7	3.2	1.3	0.2	-//-
4	4.6	3.1	1.5	0.2	-//-
5	5.0	3.6	1.4	0.2	-//-
6	5.4	3.9	1.7	0.4	-//-
7	4.6	3.4	1.4	0.3	-//-
8	5.0	3.4	1.5	0.2	-//-
9	4.4	2.9	1.4	0.2	-//-
10	4.9	3.1	1.5	0.1	-//-
11	5.4	3.7	1.5	0.2	-//-
12	4.8	3.4	1.6	0.2	-//-
13	4.8	3.0	1.4	0.1	-//-
14	4.3	3.0	1.1	0.1	-//-
15	5.8	4.0	1.2	0.2	-//-
16	5.7	4.4	1.5	0.4	-//-
17	5.4	3.9	1.3	0.4	-//-
18	5.1	3.5	1.4	0.3	-//-
19	5.7	3.8	1.7	0.3	-//-
20	5.1	3.8	1.5	0.3	-//-
21	5.4	3.4	1.7	0.2	-//-

Продовження табл. Д.1

22	5.1	3.7	1.5	0.4	-//-
23	4.6	3.6	1.0	0.2	-//-
24	5.1	3.3	1.7	0.5	-//-
25	4.8	3.4	1.9	0.2	-//-
26	5.0	3.0	1.6	0.2	-//-
27	5.0	3.4	1.6	0.4	-//-
28	5.2	3.5	1.5	0.2	-//-
29	5.2	3.4	1.4	0.2	-//-
30	4.7	3.2	1.6	0.2	-//-
31	4.8	3.1	1.6	0.2	-//-
32	5.4	3.4	1.5	0.4	-//-
33	5.2	4.1	1.5	0.1	-//-
34	5.5	4.2	1.4	0.2	-//-
35	4.9	3.1	1.5	0.1	-//-
36	5.0	3.2	1.2	0.2	-//-
37	5.5	3.5	1.3	0.2	-//-
38	4.9	3.1	1.5	0.1	-//-
39	4.4	3.0	1.3	0.2	-//-
40	5.1	3.4	1.5	0.2	-//-
41	5.0	3.5	1.3	0.3	-//-
42	4.5	2.3	1.3	0.3	-//-
43	4.4	3.2	1.3	0.2	-//-
44	5.0	3.5	1.6	0.6	-//-
45	5.1	3.8	1.6	0.2	-//-
46	4.8	3.0	1.4	0.3	-//-
47	5.1	3.8	1.6	0.2	-//-
48	4.6	3.2	1.4	0.2	-//-
49	5.3	3.7	1.5	0.2	-//-
50	5.0	3.3	1.4	0.2	-//-
51	5.0	3.0	1.6	0.2	Ірис різноколірний
52	5.0	3.4	1.6	0.4	-//-
53	5.2	3.5	1.5	0.2	-//-
54	5.2	3.4	1.4	0.2	-//-
55	4.7	3.2	1.6	0.2	-//-
56	4.8	3.1	1.6	0.2	-//-

Продовження табл. Д.1

57	5.4	3.4	1.5	0.4	-//-
58	5.2	4.1	1.5	0.1	-//-
59	5.5	4.2	1.4	0.2	-//-
60	4.9	3.1	1.5	0.1	-//-
61	5.0	3.2	1.2	0.2	-//-
62	5.5	3.5	1.3	0.2	-//-
63	6.0	2.2	4.0	1.0	-//-
64	6.1	2.9	4.7	1.4	-//-
65	5.6	2.9	3.6	1.3	-//-
66	6.7	3.1	4.4	1.4	-//-
67	5.6	3.0	4.5	1.5	-//-
68	5.8	2.7	4.1	1.0	-//-
69	6.2	2.2	4.5	1.5	-//-
70	5.6	2.5	3.9	1.1	-//-
71	5.9	3.2	4.8	1.8	-//-
72	6.1	2.8	4.0	1.3	-//-
73	6.3	2.5	4.9	1.5	-//-
74	6.1	2.8	4.7	1.2	-//-
75	6.4	2.9	4.3	1.3	-//-
76	6.6	3.0	4.4	1.4	-//-
77	6.8	2.8	4.8	1.4	-//-
78	6.7	3.0	5.0	1.7	-//-
79	6.0	2.9	4.5	1.5	-//-
80	5.7	2.6	3.5	1.0	-//-
81	5.5	2.4	3.8	1.1	-//-
82	5.5	2.4	3.7	1.0	-//-
83	5.8	2.7	3.9	1.2	-//-
84	6.0	2.7	5.1	1.6	-//-
85	5.4	3.0	4.5	1.5	-//-
86	6.0	3.4	4.5	1.6	-//-
87	6.7	3.1	4.7	1.5	-//-
88	6.0	2.2	4.0	1.0	-//-
89	6.1	2.9	4.7	1.4	-//-
90	5.6	2.9	3.6	1.3	-//-
91	6.7	3.1	4.4	1.4	-//-
92	5.6	3.0	4.5	1.5	-//-

Продовження табл. Д.1

93	5.8	2.7	4.1	1.0	-//-
94	6.2	2.2	4.5	1.5	-//-
95	5.6	2.5	3.9	1.1	-//-
96	5.9	3.2	4.8	1.8	-//-
97	6.1	2.8	4.0	1.3	-//-
98	6.3	2.5	4.9	1.5	-//-
99	6.1	2.8	4.7	1.2	-//-
100	6.4	2.9	4.3	1.3	-//-
101	6.3	3.3	6.0	2.5	Ірис віргінський
102	5.8	2.7	5.1	1.9	-//-
103	7.1	3.0	5.9	2.1	-//-
104	6.3	2.9	5.6	1.8	-//-
105	6.5	3.0	5.8	2.2	-//-
106	7.6	3.0	6.6	2.1	-//-
107	4.9	2.5	4.5	1.7	-//-
108	7.3	2.9	6.3	1.8	-//-
109	6.7	2.5	5.8	1.8	-//-
110	7.2	3.6	6.1	2.5	-//-
111	6.5	3.2	5.1	2.0	-//-
112	6.4	2.7	5.3	1.9	-//-
113	6.8	3.0	5.5	2.1	-//-
114	5.7	2.5	5.0	2.0	-//-
115	5.8	2.8	5.1	2.4	-//-
116	6.4	3.2	5.3	2.3	-//-
117	6.5	3.0	5.5	1.8	-//-
118	7.7	3.8	6.7	2.2	-//-
119	7.7	2.6	6.9	2.3	-//-
120	6.0	2.2	5.0	1.5	-//-
121	6.9	3.2	5.7	2.3	-//-
122	5.6	2.8	4.9	2.0	-//-
123	7.7	2.8	6.7	2.0	-//-
124	6.3	2.7	4.9	1.8	-//-
125	6.7	3.3	5.7	2.1	-//-
126	6.3	3.3	6.0	2.5	-//-
127	5.8	2.7	5.1	1.9	-//-

Продовження табл. Д.1

128	7.1	3.0	5.9	2.1	-//-
129	6.3	2.9	5.6	1.8	-//-
130	6.5	3.0	5.8	2.2	-//-
131	7.6	3.0	6.6	2.1	-//-
132	4.9	2.5	4.5	1.7	-//-
133	7.3	2.9	6.3	1.8	-//-
134	6.7	2.5	5.8	1.8	-//-
135	7.2	3.6	6.1	2.5	-//-
136	6.5	3.2	5.1	2.0	-//-
137	6.4	2.7	5.3	1.9	-//-
138	6.8	3.0	5.5	2.1	-//-
139	6.0	3.0	4.8	1.8	-//-
140	6.9	3.1	5.4	2.1	-//-
141	6.7	3.1	5.6	2.4	-//-
142	6.9	3.1	5.1	2.3	-//-
143	5.8	2.7	5.1	1.9	-//-
144	6.8	3.2	5.9	2.3	-//-
145	6.7	3.3	5.7	2.5	-//-
146	6.7	3.0	5.2	2.3	-//-
147	6.3	2.5	5.0	1.9	-//-
148	6.5	3.0	5.2	2.0	-//-
149	6.2	3.4	5.4	2.3	-//-
150	5.9	3.0	5.1	1.8	-//-

Таблиця Д.2

Характеристики зразків вин [55]

№ зразка	Етанол, ppm*	Яблучна кислота, ppm	Зола	Лужність золи	Магній, ppm	Феноли, ppm	Флавоноїди, ppm	Нефлавоноїдні феноли, ppm	Проантоціанідини, ppm	Інтенсивність кольору	Колірний тон	OD280/OD315** розбавленого вина	Пролін, ppm	Сорт
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065	1
2	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050	1
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185	1
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480	1
5	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735	1
6	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450	1
7	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.25	1.02	3.58	1290	1
8	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.05	1.06	3.58	1295	1
9	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.20	1.08	2.85	1045	1
10	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045	1
11	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510	1
12	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.00	1.17	2.82	1280	1
13	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.60	1.15	2.90	1320	1
14	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.40	1.25	2.73	1150	1

Продовження табл. Д.2

34	13.76	1.53	2.70	19.5	132	2.95	2.74	0.50	1.35	5.40	1.25	3		
35	13.51	1.80	2.65	19.0	110	2.35	2.53	0.29	1.54	4.20	1.10	2.00	1547	1
												.88	1310	1
15	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.50	1.20	3.65	1280	1
16	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.30	1.28	2.57	1130	1
17	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.20	1.07	2.82	1680	1
18	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.60	1.13	2.36	845	1
19	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.70	1.23	2.71	780	1
20	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.10	0.96	3.52	770	1
21	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10	5.65	1.09	3.00	1035	1
22	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98	4.50	1.03	3.63	1015	1
23	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.80	1.11	4.82	845	1
24	12.85	1.60	2.52	17.8	95	2.48	2.37	0.26	1.46	3.93	1.09	3.20	830	1
25	13.50	1.81	2.61	20.0	96	2.53	2.61	0.28	1.66	3.52	1.12	3.22	1195	1
26	13.05	2.05	3.22	25.0	124	2.63	2.68	0.47	1.92	3.58	1.13	3.77	1285	1
27	13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45	4.80	0.92	3.40	915	1
28	13.30	1.72	2.14	17.0	94	2.40	2.19	0.27	1.35	3.95	1.02	2.59	1035	1
29	13.87	1.90	2.80	19.4	107	2.95	2.97	0.37	1.76	4.50	1.25	3.71	1285	1
30	14.02	1.68	2.21	16.0	96	2.65	2.33	0.26	1.98	4.70	1.04	3.88	1515	1
31	13.73	1.50	2.70	22.5	101	3.00	3.25	0.29	2.38	5.70	1.19	2.87	990	1
32	13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95	6.90	1.09	2.00	1235	1
33	13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97	3.84	1.23	2.87	1095	1
34	13.76	1.53	2.70	19.5	132	2.95	2.74	0.50	1.35	5.40	1.25	3		

Продовження табл. Д.2

56	13	48	1	2.41	20.5	100	2.70	2.98	0.26	1.86	5.10	1.04	3.47	920	1
		28	1.81	2.84	15.5	110	2.60	2.68	0.34	1.36	4.60	1.09	2.78	880	1
36	13	05	1.64	2.55	18.0	98	2.45	2.43	0.29	1.44	4.25	1.12	2.51	1105	1
37	13	07	1.65	2.10	15.5	98	2.40	2.64	0.28	1.37	3.70	1.18	2.69	1020	1
38	13	22	3.50	2.51	13.2	128	3.00	3.04	0.20	2.08	5.10	0.89	3.53	760	1
39	13	56	1.99	2.31	16.2	117	3.15	3.29	0.34	2.34	6.13	0.95	3.38	795	1
40	14	41	3.71	2.12	18.8	90	2.45	2.68	0.27	1.48	4.28	0.91	3.00	1035	1
41	13	88	1.84	2.59	15.0	101	3.25	3.56	0.17	1.70	5.43	0.88	3.56	1095	1
42	13	24	3.89	2.29	17.5	103	2.64	2.63	0.32	1.66	4.36	0.82	3.00	680	1
43	13	05	1.98	2.10	17.0	107	3.00	3.00	0.28	2.03	5.04	0.88	3.35	885	1
44	13	21	4.77	2.44	18.9	111	2.85	2.65	0.30	1.25	5.24	0.87	3.33	1080	1
45	13	38	3.04	2.28	16.0	102	3.25	3.17	0.27	2.19	4.90	1.04	3.44	1065	1
46	14	90	1.59	2.12	16.0	101	3.10	3.39	0.21	2.14	6.10	0.91	3.33	985	1
47	14	10	2.68	2.40	18.8	103	2.75	2.92	0.32	2.38	6.20	1.07	2.75	1060	1
48	13	94	1.02	2.27	17.4	108	2.88	3.54	0.32	2.08	8.90	1.12	3.10	1260	1
49	14	05	1.73	2.04	12.4	92	2.72	3.27	0.17	2.91	7.20	1.12	2.91	1150	1
50	13	83	1.73	2.60	17.2	94	2.45	2.99	0.22	2.29	5.60	1.24	3.37	1265	1
51	13	82	1.65	2.42	14.0	111	3.88	3.74	0.32	1.87	7.05	1.01	3.26	1190	1
52	13	77	1.75	2.68	17.1	115	3.00	2.79	0.39	1.68	6.30	1.13	2.93	1375	1
53	13	74	1.90	2.25	16.4	118	2.60	2.90	0.21	1.62	5.85	0.92	3.20	1060	1
54	13	56	1.67	2.46	20.5	116	2.96	2.78	0.20	2.45	6.25	0.98	3.03	1120	1
55	13		73												

Продовження табл. Д.2

74	12.99	1.67	2.60	30.0	139	3.30	2.89	0.21	1.96	3.35	0.94	3.31	970	1
75	11.96	1.09	2.30	21.0	101	3.38	2.14	0.13	1.65	3.21	1.07	2.84	1270	1
76	11.66	1.88	1.92	16.0	97	1.61	1.57	0.34	1.15	3.80	0.89	2.87	1285	1
77	13.03	0.90	1.71	16.0	86	1.95	2.03	0.24	1.46	4.60	1.05	1.82	520	2
57	14.22	1.70	2.30	16.3	118	3.20	3.00	0.25	2.03	6.38	1.25	1.67	680	2
58	13.29	1.97	2.68	16.8	102	3.00	3.23	0.31	1.66	6.00	0.98	1.59	450	2
59	13.72	1.43	2.50	16.7	108	3.40	3.67	0.19	2.04	6.80	1.23	2.46	630	2
60	12.37	0.94	1.36	10.6	88	1.98	0.57	0.28	0.42	1.95	1.22	2.87	420	2
61	12.33	1.10	2.28	16.0	101	2.05	1.09	0.63	0.41	3.27	1.45	2.23	355	2
62	12.64	1.36	2.02	16.8	100	2.02	1.41	0.53	0.62	5.75	1.19	2.30	678	2
63	13.67	1.25	1.92	18.0	94	2.10	1.79	0.32	0.73	3.80	1.12	3.18	502	2
64	12.37	1.13	2.16	19.0	87	3.50	3.10	0.19	1.87	4.45	1.12	3.48	510	2
65	12.17	1.45	2.53	19.0	104	1.89	1.75	0.45	1.03	2.95	1.02	1.93	750	2
66	12.37	1.21	2.56	18.1	98	2.42	2.65	0.37	2.08	4.60	1.28	3.07	718	2
67	13.11	1.01	1.70	15.0	78	2.98	3.18	0.25	2.28	5.30	0.91	1.82	870	2
68	12.37	1.17	1.92	19.6	78	2.11	2.00	0.27	1.04	4.68	1.36	3.16	410	2
69	13.34	0.94	2.36	17.0	110	2.53	1.30	0.85	0.42	3.17	0.98	2.78	472	2
70	12.21	1.19	1.75	16.8	151	1.85	1.28	0.14	2.50	2.85	1.31	3.50	985	2
71	12.29	1.61	2.21	20.4	103	1.10	1.02	0.37	1.46	3.05	0.99	3.13	886	2
72	13.86	1.51	2.67	25.0	86	2.95	2.86	0.21	1.87	3.38	1.23	2.14	428	2
73	13.49	1.66	2.24	24.0	87	1.88	1.84	0.27	1.03	3.74	1.19	2.48	392	2

Продовження табл. Д.2

78	11.84	2.89	2.23	18.0	112	1.72	1.32	0.43	0.95	2.65	0.96	2.52	500	2
79	12.33	0.99	1.95	14.8	136	1.90	1.85	0.35	2.76	3.40	1.06	2.31	750	2
80	12.70	3.87	2.40	23.0	101	2.83	2.55	0.43	1.95	2.57	1.19	3.13	463	2
81	12.00	0.92	2.00	19.0	86	2.42	2.26	0.30	1.43	2.50	1.38	3.12	278	2
82	12.72	1.81	2.20	18.8	86	2.20	2.53	0.26	1.77	3.90	1.16	3.14	714	2
83	12.08	1.13	2.51	24.0	78	2.00	1.58	0.40	1.40	2.20	1.31	2.72	630	2
84	13.05	3.86	2.32	22.5	85	1.65	1.59	0.61	1.62	4.80	0.84	2.01	515	2
85	11.84	0.89	2.58	18.0	94	2.20	2.21	0.22	2.35	3.05	0.79	3.08	520	2
86	12.67	0.98	2.24	18.0	99	2.20	1.94	0.30	1.46	2.62	1.23	3.16	450	2
87	12.16	1.61	2.31	22.8	90	1.78	1.69	0.43	1.56	2.45	1.33	2.26	495	2
88	11.65	1.67	2.62	26.0	88	1.92	1.61	0.40	1.34	2.60	1.36	3.21	562	2
89	11.64	2.06	2.46	21.6	84	1.95	1.69	0.48	1.35	2.80	1.00	2.75	680	2
90	12.08	1.33	2.30	23.6	70	2.20	1.59	0.42	1.38	1.74	1.07	3.21	625	2
91	12.08	1.83	2.32	18.5	81	1.60	1.50	0.52	1.64	2.40	1.08	2.27	480	2
92	12.00	1.51	2.42	22.0	86	1.45	1.25	0.50	1.63	3.60	1.05	2.65	450	2
93	12.69	1.53	2.26	20.7	80	1.38	1.46	0.58	1.62	3.05	0.96	2.06	495	2
94	12.29	2.83	2.22	18.0	88	2.45	2.25	0.25	1.99	2.15	1.15	3.30	290	2
95	11.62	1.99	2.28	18.0	98	3.02	2.26	0.17	1.35	3.25	1.16	2.96	345	2
96	12.47	1.52	2.20	19.0	162	2.50	2.27	0.32	3.28	2.60	1.16	2.63	937	2
97	11.81	2.12	2.74	21.5	134	1.60	0.99	0.14	1.56	2.50	0.95	2.26	625	2
98	12.29	1.41	1.98	16.0	85	2.55	2.50	0.29	1.77	2.90	1.23	2.74	428	2

Продовження табл. Д.2

116	11.03	1.51	2.10	18.5	88	3.52	3.75	0.24	1.95	4.50	1.04	2.77	660	2
117	11.82	1.47	2.21	18.0	88	2.85	2.99	0.45	2.81	2.30	1.42	2.83	406	2
118	12.42	1.61	1.70	17.5	97	2.23	2.17	0.26	1.40	3.30	1.27	2.96	710	2
119	12.77	3.43	1.90	18.5	88	1.45	1.36	0.29	1.35	2.45	1.04	2.77	562	2
99	12.37	1.07	2.46	21.0	98	2.56	2.11	0.34	1.31	2.80	0.80	3.38	438	2
100	12.29	3.17	1.88	19.5	86	2.50	1.64	0.37	1.42	2.06	0.94	2.44	415	2
101	12.08	2.08	1.98	20.5	85	2.20	1.92	0.32	1.48	2.94	1.04	3.57	672	2
102	12.60	1.34	2.27	22.0	90	1.68	1.84	0.66	1.42	2.70	0.86	3.30	315	2
103	12.34	2.45	2.12	19.0	80	1.65	2.03	0.37	1.63	3.40	1.00	3.17	510	2
104	11.82	1.72	2.28	22.5	84	1.38	1.76	0.48	1.63	3.30	0.88	2.42	488	2
105	12.51	1.73	1.94	19.0	92	2.36	2.04	0.39	2.08	2.70	0.86	3.02	312	2
106	12.42	2.55	2.70	20.0	94	2.74	2.92	0.29	2.49	2.65	0.96	3.26	680	2
107	12.25	1.73	1.82	19.5	107	3.18	2.58	0.24	3.58	2.90	0.75	2.81	562	2
108	12.72	1.75	2.17	21.0	88	2.55	2.27	0.26	1.22	2.00	0.90	2.78	325	2
109	12.22	1.29	2.92	20.0	103	1.75	2.03	0.60	1.05	3.80	1.23	2.50	607	2
110	11.61	1.35	2.50	21.0	88	2.48	2.01	0.42	1.44	3.08	1.10	2.31	434	2
111	11.46	3.74	2.50	22.5	84	2.56	2.29	0.43	1.04	2.90	0.93	3.19	385	2
112	12.52	2.43	2.20	21.5	85	2.46	2.17	0.52	2.01	1.90	1.71	2.87	407	2
113	11.76	2.68	1.99	20.8	86	1.98	1.60	0.30	1.53	1.95	0.95	3.33	495	2
114	11.41	0.74	2.19	22.5	108	2.00	2.09	0.34	1.61	2.06	1.06	2.96	345	2
115	12.08	1.39	1.98	16.0	80	1.63	1.25	0.43	0.83	3.40	0.70	2.12	372	2

Продовження табл. Д.2

137	12.25	4.72	2.54	21.0	89	1.38	0.47	0.53	0.80	3.85	0.75	1.27	
138	12.53	5.51	2.64	25.0	96	1.79	0.60	0.63	1.10	5.00	0.82	1.69	564	2
139	13.49	3.59	2.19	19.5	88	1.62	0.48	0.58	0.88	5.70	0.81	1.82	625	2
140	12.84	2.96	2.61	24.0	101	2.32	0.60	0.53	0.81	4.92	0.89	2.15	465	2
													365	2
120	12.00	3.43	2.00	19.0	87	2.00	1.64	0.37	1.87	1.28	0.93	3.05	380	2
121	11.45	2.40	2.42	20.0	96	2.90	2.79	0.32	1.83	3.25	0.80	3.39	380	2
122	11.56	2.05	3.23	28.5	119	3.18	5.08	0.47	1.87	6.00	0.93	3.69	378	2
123	12.42	4.43	2.73	26.5	102	2.20	2.13	0.43	1.71	2.08	0.92	3.12	352	2
124	13.05	5.80	2.13	21.5	86	2.62	2.65	0.30	2.01	2.60	0.73	3.10	466	2
125	11.87	4.31	2.39	21.0	82	2.86	3.03	0.21	2.91	2.80	0.75	3.64	342	2
126	12.07	2.16	2.17	21.0	85	2.60	2.65	0.37	1.35	2.76	0.86	3.28	580	2
127	12.43	1.53	2.29	21.5	86	2.74	3.15	0.39	1.77	3.94	0.69	2.84	630	3
128	11.79	2.13	2.78	28.5	92	2.13	2.24	0.58	1.76	3.00	0.97	2.44	530	3
129	12.37	1.63	2.30	24.5	88	2.22	2.45	0.40	1.90	2.12	0.89	2.78	560	3
130	12.04	4.30	2.38	22.0	80	2.10	1.75	0.42	1.35	2.60	0.79	2.57	600	3
131	12.86	1.35	2.32	18.0	122	1.51	1.25	0.21	0.94	4.10	0.76	1.29	650	3
132	12.88	2.99	2.40	20.0	104	1.30	1.22	0.24	0.83	5.40	0.74	1.42	695	3
133	12.81	2.31	2.40	24.0	98	1.15	1.09	0.27	0.83	5.70	0.66	1.36	720	3
134	12.70	3.55	2.36	21.5	106	1.70	1.20	0.17	0.84	5.00	0.78	1.29	515	3
135	12.51	1.24	2.25	17.5	85	2.00	0.58	0.60	1.25	5.45	0.75	1.51	580	3
136	12.60	2.46	2.20	18.5	94	1.62	0.66	0.63	0.94	7.10	0.73	1.58	590	3
137	12.25	4.72	2.54	21.0	89	1.38								

Продовження табл. Д.2

141	12.93	2.81	2.70	21.0	96	1.64	0.50	0.53	0.75	4.60	0.77	2.31	600	3
142	13.36	2.56	2.35	20.0	89	1.40	0.50	0.37	0.64	5.60	0.70	2.47	780	3
143	13.52	3.17	2.72	23.5	97	1.55	0.52	0.50	0.55	4.35	0.89	2.06	520	3
144	13.62	4.95	2.35	20.0	92	2.00	0.80	0.47	1.02	4.40	0.91	2.05	550	3
145	12.25	3.88	2.20	18.5	112	1.38	0.78	0.29	1.14	8.21	0.65	2.00	855	3
146	13.16	3.57	2.15	21.0	102	1.50	0.55	0.43	1.30	4.00	0.60	1.68	830	3
147	13.88	5.04	2.23	20.0	80	0.98	0.34	0.40	0.68	4.90	0.58	1.33	415	3
148	12.87	4.61	2.48	21.5	86	1.70	0.65	0.47	0.86	7.65	0.54	1.86	625	3
149	13.32	3.24	2.38	21.5	92	1.93	0.76	0.45	1.25	8.42	0.55	1.62	650	3
150	13.08	3.90	2.36	21.5	113	1.41	1.39	0.34	1.14	9.40	0.57	1.33	550	3
151	13.50	3.12	2.62	24.0	123	1.40	1.57	0.22	1.25	8.60	0.59	1.30	500	3
152	12.79	2.67	2.48	22.0	112	1.48	1.36	0.24	1.26	10.80	0.48	1.47	480	3
153	13.11	1.90	2.75	25.5	116	2.20	1.28	0.26	1.56	7.10	0.61	1.33	425	3
154	13.23	3.30	2.28	18.5	98	1.80	0.83	0.61	1.87	10.52	0.56	1.51	675	3
155	12.58	1.29	2.10	20.0	103	1.48	0.58	0.53	1.40	7.60	0.58	1.55	640	3
156	13.17	5.19	2.32	22.0	93	1.74	0.63	0.61	1.55	7.90	0.60	1.48	725	3
157	13.84	4.12	2.38	19.5	89	1.80	0.83	0.48	1.56	9.01	0.57	1.64	480	3
158	12.45	3.03	2.64	27.0	97	1.90	0.58	0.63	1.14	7.50	0.67	1.73	880	3
159	14.34	1.68	2.70	25.0	98	2.80	1.31	0.53	2.70	13.00	0.57	1.96	660	3
160	13.48	1.67	2.64	22.5	89	2.60	1.10	0.52	2.29	11.75	0.57	1.78	620	3
161	12.36	3.83	2.38	21.0	88	2.30	0.92	0.50	1.04	7.65	0.56	1.58	520	3

Продовження табл. Д.2

Співвідношення світлопоглинання розведеного вина при 280 і 315 нм

162	13.69	3.26	2.74	20.0	107	1.83	0.56	0.50	0.80	5.88	0.96	1.82	680	3
163	12.85	3.27	2.[illegible]	22.0	106	1.65	0.60	0.60	0.96	5.58	0.87	2.11	570	3
164	12.96	3.45	2.[illegible]	18.5	106	1.39	0.70	0.40	0.94	5.28	0.68	1.75	675	3
165	13.78	2.76	2.54	22.0	90	1.35	0.68	0.41	1.03	9.58	0.70	1.68	615	3
166	13.73	4.36	2.68	22.5	88	1.28	0.47	0.52	1.15	6.62	0.78	1.75	520	3
167	13.45	3.70	2.35	23.0	111	1.70	0.92	0.43	1.46	0.68	0.85	1.56	695	3
168	12.82	3.37	2.30	19.5	88	1.48	0.66	0.40	0.97	0.26	0.72	1.75	685	3
169	13.58	2.58	2.66	24.5	105	1.55	0.84	0.39	1.54	8.66	0.74	1.80	750	3
170	13.40	4.60	2.60	25.0	112	1.98	0.96	0.27	1.11	8.50	0.67	1.92	630	3
171	12.20	3.03	2.30	19.0	96	1.25	0.49	0.40	0.73	5.50	0.66	1.83	510	3
172	12.77	2.39	2.69	19.5	86	1.39	0.51	0.48	0.64	9.90	0.57	1.63	470	3
173	14.16	2.51	2.86	20.0	91	1.68	0.70	0.44	1.24	9.70	0.62	1.71	660	3
174	13.71	5.65	2.32	20.5	95	1.68	0.61	0.52	1.06	7.70	0.64	1.74	740	3
175	13.40	3.91	2.28	23.0	102	1.80	0.75	0.43	1.41	7.30	0.70	1.56	750	3
176	13.27	4.28	2.48	20.0	120	1.59	0.69	0.43	1.35	0.20	0.59	1.56	835	3
177	13.17	2.59	2.45	20.0	120	1.65	0.68	0.53	1.46	9.30	0.60	1.62	840	3
178	14.13	4.10	2.48	24.5	96	2.05	0.76	0.56	1.35	9.20	0.61	1.60	560	3

* Мільйонна частка.

** Співвідношення світлопоглинання розведеного вина при 280 і 315 нм.

Таблиця Д.3

Сольватохромні параметри розчинників

№	Розчинник	Параметр			Клас
		α	β	π^*	
1	диізопропіловий етер	0.00	0.49	0.27	1
2	ди-н-бутиловий етер	0.00	0.46	0.24	1
3	діетиловий етер	0.00	0.47	0.27	1
4	діоксан	0.00	0.37	0.55	2
5	тетрагідрофуран	0.00	0.55	0.58	2
6	анізол	0.00	0.22	0.73	4
7	етер дибензилу	0.00	0.41	0.80	2
8	дифеніловий етер	0.00	0.13	0.66	4
9	етилфеніловий етер	0.00	0.20	0.69	4
10	2-бутанон	0.06	0.48	0.67	2
11	ацетон	0.08	0.48	0.71	2
12	етилацетат	0.00	0.45	0.55	2
13	етилбензоат	0.00	0.41	0.74	2
14	пропіленкарбонат	0.00	0.40	0.83	2
15	диметилацетамід	0.00	0.76	0.88	3
16	диметилформамід	0.00	0.69	0.88	3
17	N-метилпіролідон	0.00	0.77	0.92	3
18	тетраметилсечовина	0.00	0.80	0.83	3
19	триетиламін	0.00	0.71	0.14	1
20	диметилсульфоксид	0.00	0.76	1.00	3
21	гексаметилфосфортриамід	0.00	1.05	0.87	3
22	нітробензен	0.00	0.39	1.01	2
23	бензонітрил	0.00	0.41	0.90	2
24	ацетонітрил	0.19	0.31	0.75	2
25	піридин	0.00	0.64	0.87	3
26	2,6-діметилпіридин	0.00	0.76	0.80	3
27	хінолін	0.00	0.64	0.92	3
28	толуол	0.00	0.11	0.54	4
29	бензол	0.00	0.10	0.59	4
30	хлорбензол	0.00	0.07	0.71	4
31	бромбензол	0.00	0.06	0.79	4
32	тетрахлорметан	0.00	0.00	0.28	4

Продовження табл. Д.3

33	1,2-дихлоретан	0.00	0.00	0.81	4
34	дихлорметан	0.30	0.00	0.82	4
35	хлороформ	0.44	0.00	0.58	4
36	трет-бутанол	0.68	1.01	0.41	5
37	ізопропанол	0.76	0.95	0.48	5
38	н-бутанол	0.79	0.88	0.47	5
39	етанол	0.83	0.77	0.54	5
40	метанол	0.93	0.62	0.60	5
41	етиленгліколь	0.90	0.52	0.92	5
42	вода	1.17	0.18	1.09	5
43	пентан	0.00	0.00	-0.08	6
44	гексан	0.00	0.00	-0.04	6
45	гептан	0.00	0.00	-0.02	6
46	декан	0.00	0.00	0.03	6
47	гексадекан	0.00	0.00	0.08	6
48	ізооктан	0.00	0.00	-0.04	6
49	циклогексан	0.00	0.00	0.00	6
50	циклогексанон	0.00	0.53	0.76	2
51	γ-бутиролактон	0.00	0.49	0.87	2
52	нітрометан	0.22	0.25	0.85	2
53	п-ксилол	0.00	0.12	0.51	4
54	ацетофенон	0.00	0.49	0.90	2
55	хлоретан	0.00	0.00	0.81	4
56	фторбензол	0.00	0.07	0.62	4

Таблиця Д.4

Властивості 76 органічних розчинників

Розчинник	δ	γ	μ	ε	n	α	π^*	ET	Str.
н-пентан	14.4	15.5	0.00	1.84	1.843	0.00	-0.15	31.1	0.41
н-гексан	15.0	17.9	0.00	1.88	1.883	0.00	-0.11	31.0	0.40
н-гептан	15.2	19.7	0.00	1.92	1.925	0.00	-0.06	31.1	0.41
н-октан	15.5	21.2	0.00	1.95	1.946	0.00	0.01	31.1	0.41
і-октан	14.7	18.3	0.00	1.96	1.962	0.00	0.01	31.0	0.31
н-декан	15.8	23.4	0.00	1.99	1.987	0.00	0.03	31.0	0.42
н-додекан	16.0	24.9	0.00	2.00	2.015	0.00	-0.01	31.1	0.43
н-тетрадекан	16.2	26.0	0.00	2.03	2.034	0.00	0.06	31.0	0.44
н-гексадекан	16.3	27.1	0.00	2.05	2.052	0.00	0.08	31.0	0.45
циклогексан	16.8	24.6	0.00	2.02	2.026	0.00	0.00	30.9	0.46
метилциклогексан	16.0	23.3	0.00	2.02	2.018	0.00	0.00	31.0	0.39
цис-декалін	17.8	31.6	0.00	2.20	2.187	0.00	0.09	31.2	0.42
бензол	18.8	28.2	0.00	2.27	2.244	0.00	0.55	34.3	0.45
п-ксилол	18.1	27.8	0.00	2.27	2.230	0.00	0.45	33.1	0.45
мезитилен	18.1	28.3	0.00	2.28	2.240	0.00	0.45	32.9	0.34
гексафторбензол	16.9	21.6	0.00	1.89	1.890	0.00	0.27	34.2	0.56
тетрахлорметан	17.6	26.1	0.00	2.24	2.124	0.00	0.21	32.4	0.49
трихлоретилен	19.0	28.8	0.80	3.42	2.176	0.00	0.48	35.9	0.47
тетрахлоретилен	19.0	31.3	0.00	2.28	2.260	0.00	0.25	32.1	0.50
толуол	18.8	27.9	0.31	2.38	2.232	0.00	0.49	33.9	0.61
о-ксилол	18.0	29.5	0.45	2.57	2.259	0.00	0.51	34.7	0.50
м-ксилол	18.0	28.1	0.30	2.37	2.234	0.00	0.47	34.6	0.44
етилбензол	18.0	28.5	0.37	2.40	2.051	0.00	0.53	34.1	0.46
п-цимол	17.6	27.7	0.39	2.38	2.217	0.00	0.41	34.6	0.47
1,4-діоксан	19.7	32.8	0.45	2.21	2.017	0.00	0.49	36.0	0.66
фторбензол	18.1	27.1	1.48	5.42	2.156	0.00	0.62	37.0	0.55
дихлорметан	20.2	27.2	1.14	8.93	2.020	0.13	0.82	40.7	0.78
хлороформ	19.5	26.5	1.15	4.89	2.082	0.20	0.58	39.1	0.64
1,2-дихлоретан	20.0	31.5	1.83	10.36	2.080	0.00	0.73	41.3	0.65
1,1,1-трихлоретан	19.6	24.9	1.70	7.25	2.062	0.00	0.44	36.2	0.52

Продовження табл. Д.4

1,1,2,2-хлорпропан	20.2	35.4	1.71	8.20	2.224	0.00	0.95	39.4	0.63
1-хлорпропан	17.4	21.1	1.97	7.70	1.844	0.00	0.43	37.4	0.53
1,2,3-трихлор-пропан	20.6	37.1	1.85	7.45	2.194	0.00	0.78	40.4	0.56
хлорбензол	19.8	32.5	1.69	5.62	2.317	0.00	0.68	36.8	0.50
о-дихлорбензол	20.5	36.2	2.50	9.93	2.400	0.00	0.77	38.0	0.96
м-дихлорбензол	20.0	35.5	1.54	5.04	2.382	0.00	0.65	36.7	0.42
1,2,4-трихлор-бензол	20.7	44.7	1.26	4.15	2.468	0.00	0.66	36.2	0.44
дибромметан	22.4	40.1	1.43	6.68	2.377	0.00	0.92	39.4	0.70
бромоформ	21.9	45.0	0.99	4.39	2.546	0.00	0.62	37.7	0.70
1,2-диброметан	19.8	38.3	1.19	4.75	2.359	0.00	0.75	38.3	0.71
1-бромпропан	18.2	25.2	1.93	8.09	2.050	0.00	0.49	36.9	0.59
бромбензол	20.2	35.5	1.56	5.40	2.424	0.00	0.77	36.6	0.52
дийодметан	24.1	50.0	1.08	5.32	3.083	0.00	1.00	36.5	0.95
1-йодпропан	18.6	29.0	1.84	7.00	2.258	0.00	0.60	35.7	0.61
йодбензол	20.7	38.8	1.40	4.49	2.615	0.00	0.84	36.2	0.52
тетрагідрофуран	19.0	26.4	1.75	7.58	1.974	0.00	0.55	37.4	0.58
анізол	19.7	34.6	1.25	4.33	2.293	0.00	0.70	37.1	0.63
1-амінобутан	17.8	23.5	1.37	4.88	1.956	0.05	0.31	37.6	0.71
піридин	21.7	36.3	2.37	12.91	2.273	0.00	0.87	40.5	0.69
хінолін	22.8	45.2	2.18	8.95	2.640	0.00	0.93	39.4	0.58
дисульфід карбону	20.3	31.5	0.06	2.64	2.638	0.00	0.55	32.8	0.50
тетрагідротіофен	20.5	35.0	1.90	8.61	2.257	0.00	0.60	36.8	0.55
метанол	29.3	22.3	2.87	32.66	1.760	0.98	0.60	55.4	0.92
етанол	26.0	21.9	1.66	24.55	1.848	0.86	0.54	52.9	0.80
1-пропанол	24.4	23.1	3.09	20.45	1.915	0.84	0.52	50.7	0.76
1-бутанол	23.3	24.2	1.75	17.51	1.952	0.84	0.47	49.7	0.73
1-пентанол	22.4	25.2	1.70	13.90	1.982	0.84	0.40	49.1	0.70
1-гексанол	21.8	25.7	1.55	13.30	2.005	0.80	0.40	48.8	0.68
1-октанол	20.9	26.9	1.76	10.34	2.038	0.77	0.40	48.1	0.66

Продовження табл. Д.4

1-деканол	19.9	28.4	1.62	8.10	2.064	0.70	0.45	47.7	0.63
1,2-етандіол	32.4	47.4	2.31	37.70	2.047	0.90	0.92	56.3	0.92
вода	47.9	71.8	1.85	78.36	1.776	1.17	1.09	63.1	2.31
N-метил-формамід	31.1	39.5	3.86	182.4	2.045	0.62	0.90	54.1	0.88
N,N-диметил-формамід	24.1	36.4	3.92	36.71	2.024	0.00	0.88	43.2	0.74
N,N-диметил-ацетамід	23.3	31.7	3.72	37.78	2.061	0.00	0.85	42.9	0.72
N,N-метил-піролідон	23.6	40.7	4.09	32.20	2.155	0.00	0.92	42.2	0.64
гексаметил-фосфорамід	19.1	33.8	5.54	29.30	2.123	0.00	0.87	40.9	0.63
диметил-сульфооксид	26.6	43.0	4.06	46.45	2.182	0.00	1.00	45.1	0.89
о-крезол	21.9	35.0	1.45	11.50	2.344	1.65	0.68	51.9	0.96
пропіленкарбонат	21.8	41.4	4.94	64.92	2.019	0.00	0.83	46.0	0.63
ацетон	22.1	22.7	2.69	20.56	1.839	0.08	0.62	42.2	0.68
нітрометан	25.7	36.3	3.56	35.87	1.903	0.22	0.75	46.3	0.90
нітроетан	22.7	32.1	3.60	28.06	1.931	0.00	0.77	43.6	0.78
нітробензол	22.1	42.4	4.22	34.78	2.403	0.00	0.86	42.2	0.49
ацетонітрил	24.1	28.3	3.92	35.94	1.800	0.19	0.66	45.6	0.74
бензонітрил	22.7	38.5	4.18	25.20	2.328	0.00	0.88	41.5	0.79

Наукове видання

Холін Юрій Валентинович
Пушкарьова Ярослава Миколаївна
Пантелеймонов Антон Віталійович
Некос Алла Наумівна

ХЕМОМЕТРИЧНІ МЕТОДИ В РОЗВ'ЯЗАННІ ЗАДАЧ ЯКІСНОГО ХІМІЧНОГО АНАЛІЗУ ТА КЛАСИФІКАЦІЇ ФІЗИКО-ХІМІЧНИХ ДАНИХ

Монографія

Коректор *О. В. Анцибора*
Комп'ютерне верстання *Н. О. Ваніна*
Макет обкладинки *Ю. М. Рижова*

Формат 60x84/16. Ум. друк. арк. 10,47. Наклад 300 пр. Зам. № 163/16.

Видавець і виготовлювач
Харківський національний університет імені В. Н. Каразіна,
61022, м. Харків, майдан Свободи, 4.
Свідоцтво суб'єкта видавничої справи ДК № 3367 від 13.01.2009

Видавництво ХНУ імені В. Н. Каразіна
Тел. 705-24-32