Міністерство освіти і науки України

Харківський національний університет імені В. Н. Каразіна

ПУШКАРЬОВА ЯРОСЛАВА МИКОЛАЇВНА

УДК 543.061:004.032.26

РОЗВ'ЯЗАННЯ ЗАДАЧ ЯКІСНОГО ХІМІЧНОГО АНАЛІЗУ ЗА ДОПОМОГОЮ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Спеціальність 02.00.02 – аналітична хімія

Автореферат дисертації на здобуття наукового ступеня
кандидата хімічних наук

Харків – 2013

Дисертацією є рукопис

Робота виконана в Харківському національному університеті імені В. Н. Каразіна Міністерства освіти і науки України

Науковий керівник:

доктор хімічних наук, професор
ХОЛІН ЮРІЙ ВАЛЕНТИНОВИЧ,
Харківський національний університет
імені В. Н. Каразіна Міністерства освіти і науки України,
проректор з науково-педагогічної роботи,
завідувач кафедри хімічного матеріалознавства

Офіційні опоненти:

доктор хімічних наук, професор
АНТОНОВИЧ ВАЛЕРІЙ ПАВЛОВИЧ,
Фізико-хімічний інститут ім. О. В. Богатського
НАН України (м. Одеса),
завідувач відділу аналітичної хімії
та фізико-хімії координаційних сполук

кандидат хімічних наук
БЄЛІКОВ КОСТЯНТИН МИКОЛАЙОВИЧ,
Державна наукова установа «Науково-технологічний комплекс
«Інститут монокристалів» НАН України» (м. Харків),
виконуючий обов'язки завідувача відділу аналітичної хімії
функціональних матеріалів та об'єктів
навколишнього середовища

Захист відбудеться «_____» _____ 2013 р. о ____ годині на засіданні спеціалізованої вченої ради Д 64.051.14 Харківського національного університету імені В. Н. Каразіна Міністерства освіти і науки України (Україна, 61022, м. Харків, майдан Свободи, 4, ауд. 7-79).

З дисертацією можна ознайомитись у Центральній науковій бібліотеці Харківського національного університету імені В. Н. Каразіна Міністерства освіти і науки України (Україна, 61022, м. Харків, майдан Свободи, 4).

Автореферат розісланий «_____» _____ 2013 р.

Учений секретар
спеціалізованої вченої ради,
кандидат хімічних наук, доцент _____ В. Г. Панченко

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. За останні два десятиліття суттєво зросла роль якісного хімічного аналізу. Це обумовлено зростаючою потребою в масовому аналізі складних сумішей у таких галузях, як аналіз об'єктів навколишнього середовища, перевірка автентичності медико-біологічних препаратів, продуктів харчування, харчової сировини, виявлення токсикантів, наркотиків, вибухонебезпечних речовин тощо. Змінилося і розуміння змісту якісного хімічного аналізу. В навчальній та науковій літературі все більш чіткою стає тенденція до трактовки якісного хімічного аналізу як *процедури класифікації об'єктів* за їх ознаками. В такому підході природно поєднуються задачі виявлення (встановлення присутності певного аналіту в пробі); ідентифікації (дискримінації) (ототожнення аналіту з відомою індивідуальною сполукою чи групою сполук, віднесення зразка до одного з наперед визначених класів) та кластеризації (виявлення сукупностей зразків з близькими характеристиками за відсутності навчальних вибірок). Результати аналізу розглядаються як рекомендації для прийняття управлінських рішень. Алгоритми класифікації поділяються на дві групи: алгоритми «без навчання» та «з навчанням». Алгоритми «без навчання» застосовують для знаходження однорідних груп об'єктів (проведення кластеризації), алгоритми «з навчанням» застосовують для визначення приналежності об'єктів до наперед визначених класів (проведення дискримінації, ідентифікації).

Сучасний якісний аналіз ґрунтується на обробці багатовимірних масивів експериментальних даних, одержаних інструментальними методами (хроматографічними, спектроскопічними, системами «електронний язик», «електронний ніс» тощо). Для забезпечення високої надійності класифікації аналітів обробляти такі масиви слід з використанням ефективних методів аналізу даних, зокрема, хемометричних.

Серед хемометричних методів, призначених для розв'язання задач класифікації, все більшу увагу привертають штучні нейронні мережі (ШНМ), що характеризуються адаптивною архітектурою та здатністю до навчання. Доцільність використання штучних нейронних мереж як процедури аналізу хіміко-аналітичних даних для розв'язання задач якісного аналізу обумовлена тим, що нейронні мережі, на відміну від більшості інших алгоритмів, дозволяють ефективно працювати з даними будь-якої складності і структури, а також вільні від апріорних припущень щодо статистичних характеристик вихідних даних. Крім того, алгоритми класифікації за допомогою ШНМ мають підвищену стійкість до наявності у масивах даних пропусків та «промахів».

Незважаючи на широке застосування нейронних мереж у хімії, питання про їх характеристики (архітектура, типи функцій активації, число прихованих нейронів, метод навчання), необхідні і достатні для надійного розв'язання задач якісного хімічного аналізу, а також про найбільш ефективні процедури використання ШНМ у залежності від особливостей вихідних даних не розв'язані.

Таким чином, актуальність роботи обумовлено потребою в рекомендаціях щодо вибору архітектури та управляючих параметрів штучних нейронних мереж, а також процедур їх застосування при розв'язанні задач класифікації в якісному хімічному аналізі. Особливої уваги потребують такі галузі, як контроль продуктів споживання, харчової сировини і об'єктів довкілля.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційну роботу виконано відповідно до досліджень за темою НДР кафедри хімічного матеріалознавства Харківського національного університету імені В. Н. Каразіна «Хемометричні засоби для розв'язання задач якісного хімічного аналізу» (№ держреєстрації 0112U003024).

Мета роботи: розробка та випробування рекомендацій щодо вибору архітектури, параметрів і процедур використання штучних нейронних мереж, що забезпечують високу надійність і робастність розв'язання задач ідентифікації та кластеризації в якісному хімічному аналізі.

Для досягнення поставленої мети необхідно було розв'язати такі **задачі**: розробити рекомендації щодо формування оптимального об'єму навчальної вибірки для правильного навчання алгоритмів класифікації; запропонувати спосіб оцінки числа прихованих нейронів, обґрунтувати набір функцій активації і метод навчання, що забезпечують надійну та робастну класифікацію багатовимірних хіміко-аналітичних даних; порівняти надійність і робастність алгоритмів нейронних мереж і традиційних класифікаційних процедур при обробці тестових і експериментальних хіміко-аналітичних даних різного типу; розробити алгоритм класифікації об'єктів за багатовимірними масивами їх фізико-хімічних характеристик без апріорної інформації про число класів та про еталони для навчання алгоритму; дослідити ефективність застосування алгоритмів нейронних мереж до ідентифікації географічного походження зразків вод і харчової сировини.

Об'єкт дослідження: ідентифікація та кластеризація об'єктів за даними багатовідгукового експерименту в якісному хімічному аналізі.

Предмет дослідження: параметри алгоритмів штучних нейронних мереж; закономірності функціонування алгоритмів штучних нейронних мереж при варіабельності параметрів їх синтезу; стійкість алгоритмів до наявності у даних пропусків та «промахів».

Методи дослідження: алгоритми штучних нейронних мереж для встановлення класової приналежності об'єктів; експлораторний аналіз даних для знаходження однорідних груп об'єктів; методи параметричної і непараметричної статистики та хемометрії для встановлення залежностей між характеристиками зразків харчової сировини та їх географічним походженням.

Наукова новизна одержаних результатів. Встановлено, що найвищою надійністю та робастністю при розв'язанні задач ідентифікації в якісному хімічному аналізі серед вивчених алгоритмів володіють алгоритми імовірнісної та динамічної нейронних мереж. Розроблено рекомендації щодо вибору основних управляючих параметрів для синтезу штучних нейронних мереж (функцій активації, числа нейронів, методу навчання), що забезпечують високу надійність класифікації багатовимірних хіміко-аналітичних даних. Ефективність рекомендацій перевірено при класифікації даних різного типу, зокрема, при розв'язанні такої складної задачі, як ідентифікація розчинників за значеннями сольватохромних параметрів. Встановлено, що правильне навчання алгоритмів класифікації забезпечує навчальна вибірка, значення стандартного відхилення і розмаху якої найменш відмінні від зазначених параметрів тестової вибірки. На основі застосування мережі Кохонена та імовірнісної мережі розроблено й апробовано процедуру стійкої кластеризації (класифікація без апріорної інформації про

число класів та без наявності навчальної вибірки), що дозволяє отримувати з багатовимірних експериментальних масивів даних змістовну хімічну інформацію. Показано, що спільне використання непараметричних методів статистики та імовірнісної нейронної мережі забезпечує надійну ідентифікацію географічного походження овочів і фруктів (ідентифікацію типів ландшафтів). Побудовано класифікацію 76 розчинників за їх фізико-хімічними характеристиками, що відповідає хімічній природі речовин.

Практичне значення одержаних результатів. Рекомендації щодо вибору параметрів штучних нейронних мереж і формування оптимальної навчальної вибірки забезпечують розв'язання задач ідентифікації (дискримінації) в якісному хімічному аналізі навіть за наявності в даних пропусків і вимірювань, що різко виділяються.

Запропонована процедура оцінки робастності алгоритмів класифікації може використовуватися для дослідження ефективності нових хемометричних алгоритмів.

Отримані результати є корисними при розробці експертних систем для перевірки автентичності (ідентифікації) продуктів харчування, харчової сировини та об'єктів довкілля.

Розроблена в роботі процедура кластеризації дозволяє обробляти масиви експериментальних даних з метою виявлення однорідних груп зразків без залучення для розрахунків апріорної інформації щодо кількості груп та еталонів.

Результати роботи впроваджено в навчальний процес хімічного факультету Харківського національного університету імені В. Н. Каразіна при викладанні курсу «Хемометричні методи аналізу даних» за освітньо-професійною програмою магістрів спеціальності «хімія» (акт впровадження від 21.01.2013 р.).

Особистий внесок здобувача. Аналіз літературних даних, реалізація, апробація та застосування алгоритмів виконані здобувачем особисто. Постановка мети та задач дослідження, обговорення висновків виконані спільно з науковим керівником проф. Ю. В. Холіним. Експериментальні дані про вміст металів у зразках річкових та джерельних вод м. Харкова були одержані на кафедрі хімічної метрології Харківського національного університету імені В. Н. Каразіна (проф. О. І. Юрченко, с.н.с. Н. П. Тітова). Експериментальні дані про вміст металів у овочах і фруктах із різних районів м. Харкова та Харківської області були одержані на кафедрі екологічної безпеки та екологічної освіти Харківського національного університету імені В. Н. Каразіна (проф. А. Н. Некос, інж. А. Г. Гарбуз).

Апробація результатів дисертації. Основні результати роботи були оприлюднені на VI Всеукраїнській конференції молодих вчених, студентів і аспірантів з актуальних питань хімії (Харків, 2008); Другій та Третій Всеукраїнських наукових конференціях студентів та аспірантів «Хімічні Каразінські читання – 2010» та «Хімічні Каразінські читання – 2011» (Харків, 2010, 2011); Науковій конференції, присвяченій 100 річниці з дня народження професора І. В. П'ятницького (Київ, 2010); Сесіях Наукової Ради НАН України з проблеми «Аналітична хімія» (Гурзуф, 2011, 2012); XVIII Українській конференції з неорганічної хімії за участю закордонних вчених в рамках Міжнародного року хімії ООН (Харків, 2011); X Всеукраїнській конференції молодих вчених та студентів з актуальних питань хімії (Харків, 2012); 15th International symposium of students and young mechanical engineers «Advances in chemical and mechanical engineering» (Gdansk, 2012).

Публікації. За матеріалами дисертації опубліковано 16 наукових праць, з яких 7 статей у наукових фахових виданнях та 9 тез доповідей на наукових конференціях.

Структура та обсяг дисертації. Дисертація складається зі вступу, п'яти розділів, висновків, списку використаних літературних джерел та додатку. Дисертація викладена на 190 сторінках машинописного тексту та містить 30 рисунків, 70 таблиць і бібліографію з 225 найменувань.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету та завдання дослідження, вказано наукову новизну та практичну цінність одержаних результатів.

У розділі 1 «Актуальні задачі аналітичної хімії та можливості штучних нейронних мереж у їх розв'язанні (літературний огляд)» описано становлення сучасного якісного хімічного аналізу як процедури класифікації об'єктів за їх ознаками та обговорено проблеми метрології якісного аналізу; надаються відомості про штучні нейронні мережі; виділені задачі, розв'язання яких потребує застосування апарата штучних нейронних мереж та актуальні проблеми, пов'язані з розв'язанням задач ідентифікації об'єктів у якісному хімічному аналізі.

У розділі 2 «Робастні алгоритми класифікації багатовимірних хіміко-аналітичних даних за допомогою штучних нейронних мереж» обговорено алгоритми випробуваних нейронних мереж та наведено результати їх апробації на модельних та тестових наборах даних, описано оптимізацію параметрів нейронних мереж на прикладі класифікації розчинників за їх сольватохромними характеристиками.

У роботі вивчали ефективність роботи шести двошарових алгоритмів штучних нейронних мереж «з навчанням»: імовірнісної (Probabilistic Neural Network, PNN), каскадної (Cascade Neural Network, CNN), динамічної (Dynamic Neural Network, DNN), Елмана (Elman Neural Network, ENN), «класичної» мережі прямого поширення сигналу (Feed Forward Neural Network, FFNN), LVQ-мережі (Learning Vector Quantization Neural Network), а також мережі Кохонена (Kohonen Neural Network) «без навчання». Вибір алгоритмів обумовлений їх широким застосуванням для обробки хімічних даних. Для реалізації випробуваних алгоритмів нейронних мереж використали різні комбінації функцій активації (гіперболічний тангенс, лінійна, логістична), методи навчання (Левенберга-Марквардта, Пауела-Бієле, алгоритм зворотного поширення помилки) та число прихованих нейронів з метою формулювання рекомендацій щодо вибору їх параметрів. Для LVQ-мережі та мережі Кохонена визначали крок навчання. Вагові коефіцієнти та зсуви нейронів ініціалізували згідно з алгоритмом Нгуена-Відроу. Досягнення певного значення середнього квадратичного відхилення використали як критерій зупину навчання:

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

де d_k – класова приналежність зразка k навчальної вибірки, y_k – вихід мережі для зразка k навчальної вибірки, H – число зразків навчальної вибірки, Y – число виходів нейронної мережі.

Значення середнього квадратичного відхилення розраховується після пред'явлення мережі всіх зразків навчальної вибірки. Процес навчання закінчувався, коли відхилення досягало значення 0.001.

Результати, отримані при використанні алгоритмів ШНМ, порівнювали з результатами поширених методів класифікації, а саме: лінійного дискримінантного аналізу (Linear Discriminant Analysis, LDA), методу опорних векторів (Support Vector Machine, SVM), формального незалежного моделювання аналогій класів (Soft Independent Modeling of Class Analogy, SIMCA) та дискримінантного аналізу за допомогою регресії на латентні структури (Projection to Latent Structures Discriminant Analysis, PLS-DA). При реалізації методу SVM використали радіально-базисну функцію Гауса.

Метрологія якісного хімічного аналізу знаходиться на стадії формування, її основні засади відрізняються від метрології кількісного аналізу. Ключове поняття кількісного аналізу – невизначеність (Uncertainty) – не має аналогів у якісному аналізі. В метрології кількісного аналізу оцінювання невизначеності опрацьовано досить глибоко. Для оцінки результатів якісного аналізу учасниками проекту «Metrology of qualitative chemical analysis»[1] запропоновано використовувати такий показник як ненадійність (недостовірність, Unreliability) або надійність (достовірність, Reliability):

$$\text{Reliability} = 100 - \alpha - \beta, \tag{2}$$

де α та β – імовірності помилок I та II роду (%), відповідно.

Наведена формула справедлива для методик виявлення з бінарним відгуком, але неприйнятна для багатовідгукових процедур, зокрема, задач класифікації з числом класів понад два.

В нашій роботі ненадійність класифікації оцінювали як частку невірно класифікованих зразків тестової вибірки

$$P = \frac{n}{N} \cdot 100\ \% \tag{3}$$

де n – число невірно класифікованих зразків тестової вибірки, N – загальне число зразків тестової вибірки. Надійність класифікації

$$\text{Reliability} = 100 - P. \tag{4}$$

Особливу увагу приділяли перевірці змістовності класифікації. Контрольну вибірку для перевірки правильності класифікації річкових та джерельних вод складали зразки, відібрані у році, наступному за навчанням алгоритмів ШНМ. При цьому кількість визначених у цих зразках концентрацій металів відрізнялася від кількості визначених властивостей еталонів, що використовувалися для навчання алгоритмів. Можливість отримання змістовної ідентифікації також перевіряли, використовуючи запропоновану в роботі процедуру кластеризації за відсутності інформації про кількість класів та навчальну вибірку. Порівнювали знайдене при цьому розбиття об'єктів на класи з відомим з умов відбору зразків. Крім того, правильність класифікації оцінювали за допомогою процедури перехресної перевірки достовірності (крос-валідації, cross validation).

---

[1] Quality assurance of qualitative analysis in the framework of the European project «MEQUALAN» / A. Rios, D. Barcelo, L. Buydens [e. a.] // Accred. Qual. Assur. – 2003. – V. 8. – P. 68–77.

Оптимальний об'єм навчальної вибірки для навчання алгоритмів класифікації визначали за допомогою імовірнісної мережі, що характеризується простою архітектурою. Імовірнісна мережа – мережа, що містить прихований шар нейронів з радіально-симетричною функцією активації

$$F = e^{-\left(\frac{\sqrt{\sum_{i=1}^{N}(x_i - w_{ij})^2}}{2\delta}\right)^2}, \quad (5)$$

де x_i – числові характеристики об'єкту (вхідний вектор), δ – відхилення функції, w_{ij} – вагові коефіцієнти нейронів, кількість яких визначається кількістю зразків у навчальній вибірці, та конкуруючий вихідний шар нейронів. Для реалізації PNN необхідно визначити лише оптимальне значення δ.

Оптимальний об'єм навчальної вибірки, визначений за допомогою PNN, використовували для навчання інших типів нейронних мереж. Досліджувані масиви даних випадковим чином поділяли на навчальну та тестову вибірки при різній кількості зразків у їх складі.

Під оптимальним об'ємом навчальної вибірки розуміли таке число зразків, яке забезпечувало 100 %-ву надійність класифікації зразків тестової вибірки. Коефіцієнт T, % показує, яка частка зразків від їх загального числа знаходиться у навчальній вибірці:

$$T = \frac{H}{M} \cdot 100\,\%, \quad (6)$$

де M – загальне число зразків.

Для апробації алгоритмів ШНМ використали модельні дані, що характеризуються складними структурами (двоієрархічною та дугоподібною) (рис. 1) та еталонні (тестові) масиви даних з відомою структурою (масив даних, що містить відомості про 4 характеристики 150 зразків ірисів[2], та масив даних, що містить результати визначення 13 характеристик 178 зразків вин[3]).

Класифікація даних з дугоподібною структурою з надійністю 100 % із застосуванням PNN ($\delta = 0.1$) спостерігається при $T = 70\,\%$, класифікація зразків із двоієрархічною структурою – при $T = 60\,\%$, ідентифікація зразків ірисів – при $T = 70\,\%$, ідентифікація зразків вин – при $T = 87\,\%$. Ці значення оптимального об'єму навчальної вибірки використали для навчання інших типів ШНМ.

Для адекватної ідентифікації зразків вин, характеристикам яких притаманний великий розмах значень, попередньо провели автомасштабне перетворення даних:

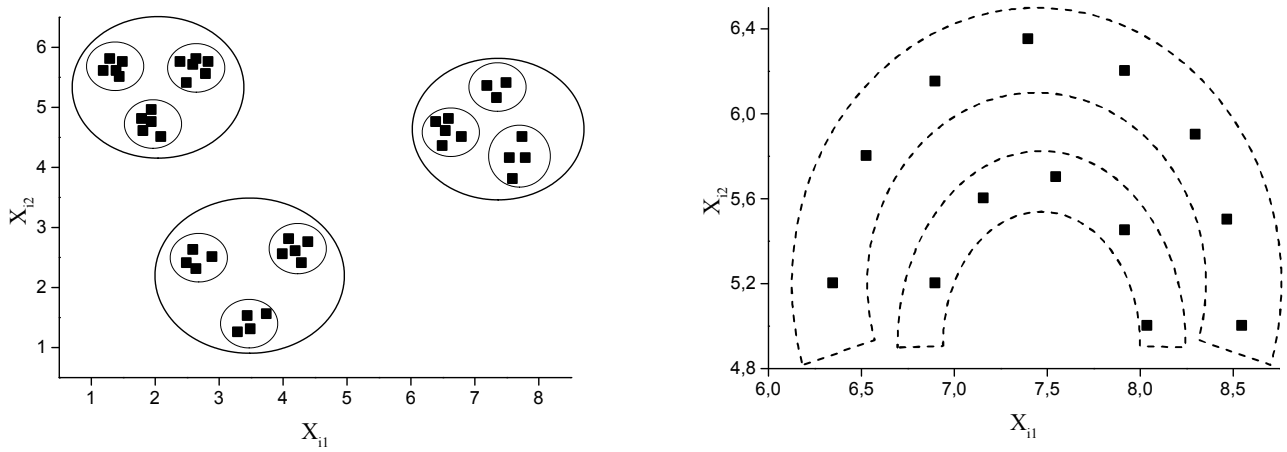
$$x_i^{norm} = \frac{x_i - \overline{x}}{S(x)}, i = 1, 2, ..., N, \quad (7)$$

де x_i^{norm} – безрозмірне значення характеристики для i-го зразка, отримане внаслідок автомасштабного перетворення, x_i – вихідне значення характеристики для i-го зразка,

[2] Iris Data Set (1988). UCI Machine Learning Repository [Електронний ресурс]. – Режим доступу : http://archive.ics.uci.edu/ml/datasets/Iris

[3] Wine Data Set (1991). UCI Machine Learning Repository [Електронний ресурс]. – Режим доступу : http://archive.ics.uci.edu/ml/datasets/wine

$\bar{x}$ – середнє значення характеристики в зразках, $S(x)$ – стандартне відхилення значень характеристики в зразках, N – число зразків.



а б

Рис. 1. Модельні дані (а) з двоієрархічною та (б) з дугоподібною структурами

Для алгоритмів ШНМ «з навчанням» визначили оптимальні методи навчання, комбінацію функцій активації та число прихованих нейронів. Під оптимальними розуміли такі значення параметрів, що забезпечують коректне навчання алгоритмів класифікації та правильне (надійність 100 %) віднесення до відповідних класів зразків тестової вибірки. Встановлено, що надійну класифікацію тестових і еталонних масивів даних у випадку FFNN, CNN та DNN забезпечує застосування методу навчання Левенберга-Марквардта, функції активації гіперболічний тангенс для прихованого шару та лінійної функції активації для вихідного шару, у випадку ENN – алгоритму зворотного поширення помилки та різних комбінацій функцій активації – гіперболічний тангенс та лінійної – в залежності від даних. Оптимальне число нейронів коливається від 6 до 14.

У роботі підтверджено гіпотезу, що у випадку двох можливих моделей нейронних мереж необхідно робити вибір на користь більш простої моделі.

Необхідно відзначити, що LVQ-мережа, мережа Кохонена та традиційні алгоритми класифікації не відображають реальну структуру модельних і тестових даних. Методи SIMCA, PLS-DA, а також LVQ-мережа проявили себе найменш ефективними.

Наступний досліджуваний масив містив значення трьох сольватохромних параметрів (параметр основності розчинників як акцепторів водневих зв’язків, параметр кислотності розчинників як донорів водневих зв’язків, параметр полярності та поляризовності) для 56 розчинників, що розподілені між 6 класами. Інтерес до обробки даного масиву викликаний декількома причинами. По-перше, сольватохромні параметри широко застосовуються в хімії та час від часу уточнюються. По-друге, в літературі зустрічаються випадки відсутності значень одного з параметрів для тих чи інших речовин. По-третє, складність обробки таких даних надає великий розмах значень сольватохромних параметрів усередині кожного класу та близькість значень сольватохромних характеристик багатьох розчинників, віднесених до різних класів. Обробка таких даних дає змогу оцінити стійкість алгоритмів до варіювання значень сольватохромних параметрів (наявності «промахів») та наявності в даних пропусків.

У вихідні дані характеристик зразків тестової вибірки вносили похибки, розподіл яких відповідає моделі «грубих промахів»:

$$\varepsilon = \left[(100 - q)\cdot \varepsilon_{Gauss}(0,\sigma) + q \cdot \varepsilon_{Laplace}(0,\sigma)\right]/100, \quad (8)$$

де $q, \%$ – інтенсивність «грубих промахів»; ε_{Gauss} – випадкова величина, що розподілена за законом Гауса з нульовим середнім значенням та стандартним відхиленням σ (приймали $\sigma = 0.1 \cdot \bar{x}$); $\varepsilon_{Laplace}$ – випадкова величина, що розподілена за законом Лапласа, хвости якого є більш довгими, ніж у нормального розподілу.

При вивченні стійкості алгоритмів класифікації до наявності в даних пропусків видаляли 10 и 20 % значень характеристик зразків тестової вибірки і заповнювали отримані пропуски середніми значеннями відповідних характеристик.

Встановлено, що PNN ($\delta = 0.1$) правильно класифікує розчинники, починаючи з $T = 60\,\%$. Це значення об'єму навчальної вибірки використали для навчання інших типів ШНМ. Для кожного алгоритму нейронної мережі при випадково обраному числі прихованих нейронів ($h = 9$) визначали метод навчання та комбінацію функцій активації, що забезпечують допустиму ненадійність (мінімальну з отриманих значень) ідентифікації розчинників тестової вибірки. Потім варіювали число прихованих нейронів до одержання 100 %-ої надійності класифікації розчинників (рис. 2).

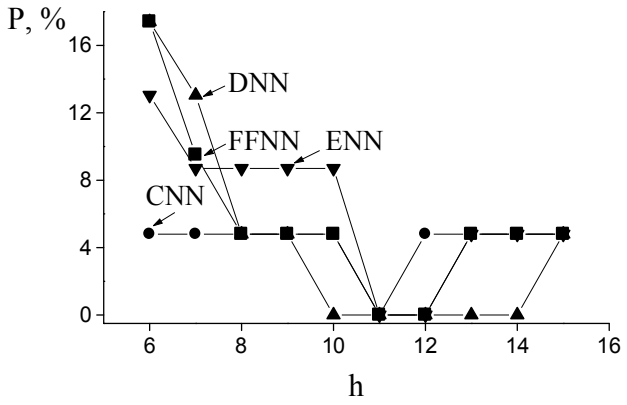


Рис. 2. Залежність ненадійності класифікації розчинників тестової вибірки від числа прихованих нейронів для 4 типів нейронних мереж

Надійну ідентифікацію розчинників за сольватохромними параметрами у випадку FFNN, CNN та DNN забезпечує застосування методу навчання Левенберга-Марквардта, у випадку ENN – алгоритму зворотного поширення помилки; ефективними є функції активації гіперболічний тангенс і лінійна. Відзначимо, що оптимальні методи навчання та функції активації, що забезпечують надійну ідентифікацію розчинників, співпадають з оптимальними параметрами для надійної класифікації модельних і тестових даних. При застосуванні методів SVM, PLS-DA ненадійність $P = 13\,\%$, методу SIMCA – $P = 26\,\%$. Стійкість алгоритмів класифікації до наявності у даних промахів зменшується у такому порядку:

$PNN > DNN > ENN \approx LDA \approx CNN > FFNN > PLS-DA > SVM > SIMCA$; стійкість до наявності пропусків – у такому порядку:

$CNN \approx PNN \approx DNN > LDA \approx ENN > FFNN > SVM > PLS-DA > SIMCA$.

Розділ 3 «Штучні нейронні мережі в розв'язанні задач ідентифікації географічного походження» містить результати ідентифікації за походженням 22 зразків річкових (річки Лопань, Уди, Немишля, Харків) і 24 зразків джерельних вод (6 джерел) м. Харкова за даними про вміст 8 металів та географічного походження (типів ландшафтів) овочів і фруктів (на прикладі 58 зразків картоплі та 22 зразків яблук) з різних районів м. Харкова і Харківської області за даними про вміст 10 металів,

відібраних протягом 2008–2010 рр., із застосуванням алгоритмів ШНМ і традиційних алгоритмів класифікації. Концентрації металів у досліджуваних зразках були визначені методом атомно-абсорбційної спектроскопії, відносні стандартні відхилення концентрацій металів у зразках вод не перевищували 3 %, сумарна відносна невизначеність результатів визначення вмісту металів у зразках рослинного походження – 10 %. Перед застосуванням алгоритмів класифікації виконали автомасштабне перетворення даних згідно з (7).

Встановлено, що PNN ($\delta = 0.1$) правильно ідентифікує зразки річкових і джерельних вод тестової вибірки, починаючи з $T = 75\,\%$. Це значення коефіцієнту T використали для навчання нейронних мереж інших типів. При виборі методів навчання керувалися встановленою оптимальною архітектурою мереж для класифікації модельних / тестових даних та ідентифікації розчинників: для FFNN, CNN та DNN застосували метод навчання Левенберга-Марквардта, для ENN – алгоритм зворотного поширення помилки. Параметри навчання та тестування реалізованих нейронних мереж (оптимальні комбінації функцій активації та число прихованих нейронів), а також ненадійність ідентифікації зразків вод наведені у табл. 1.

Таблиця 1

Параметри ШНМ і значення ненадійності ідентифікації зразків вод

Нейронна мережа	Число прихованих нейронів для ідентифікації зразків річкових / джерельних вод	Функції активації для прихованого / вихідного шарів	P, % для зразків річкових / джерельних вод
CNN	9 / 10	гіперболічний тангенс / гіперболічний тангенс	17 / 0
FFNN	9 / 11	гіперболічний тангенс / лінійна	17 / 0
ENN	11 / 12	гіперболічний тангенс / лінійна	17 / 17
DNN	11 / 11	гіперболічний тангенс / лінійна	0 / 0
PNN	16 / 18	радіальна базисна / конкуруючий шар	0 / 0

При застосуванні методів LDA, SIMCA, SVM та PLS-DA до ідентифікації зразків вод значення ненадійності P є неприйнятно високими (17–67 %).

Незалежну контрольну вибірку для перевірки правильності класифікації складали 4 зразки річкових вод, відібраних у 2011 році. У цих зразках були визначені концентрації тих самих металів (крім нікелю), що і у пробах, відібраних в 2008–2010 роках. Надійність ідентифікації зразків алгоритмами DNN та PNN за даними про вміст 7 металів склала 100 %. Крім того, перевірили стійкість алгоритмів ШНМ до наявності пропусків у даних: при підстановці замість невизначених концентрацій нікелю середньої концентрації цього елемента у зразках річкових вод, відібраних і проаналізованих у 2008–2010 роках, надійність ідентифікації за допомогою динамічної мережі становила

100 %, за допомогою інших ШНМ – 75 %, тоді як традиційні алгоритми класифікації виявилися нестійкими (табл. 2).

Таблиця 2

Результати ідентифікації зразків річкових вод при наявності в даних пропусків

Алгоритм	Число прихованих нейронів	P, %
CNN	14	25
FFNN	15	25
ENN	15	25
DNN	14	0
PNN	22	25
LDA, SVM, SIMCA, PLS-DA	–	75

Дані про вміст металів у зразках картоплі характеризуються наявністю вимірювань, що різко виділяються, та невідповідністю розподілу концентрацій п’яти металів нормальному. Імовірнісну мережу ($\delta = 0.1$) застосували як класифікаційний інструмент для встановлення географічного походження зразків картоплі та яблук, оскільки вона зарекомендувала себе алгоритмом, стійким до наявності у даних похибок, розподіл яких відрізняється від нормального. В результаті обробки масивів даних про вміст металів у зразках картоплі та яблук PNN зіткнулися з проблемою ідентифікації деяких груп зразків, що можна пояснити недостатньою кількістю зразків, відібраних із відповідних типів ландшафтів. Для знаходження залежностей між характеристиками зразків картоплі та яблук та їх географічним походженням до масивів даних застосували ряд статистичних процедур: непараметричні (розрахунок коефіцієнтів рангової кореляції Спірмена, критеріїв Уілкоксона-Манна-Уітні та Краскела-Уолліса) та параметричні методи (розрахунок коефіцієнтів кореляції Пірсона), а також метод головних компонент. В результаті розрахунку критерію Краскела-Уолліса встановлено, що на ідентифікацію зразків картоплі найбільше впливає вміст кобальту, а на ідентифікацію зразків яблук – свинцю. Результати розрахунку критерію Уілкоксона-Манна-Уітні дозволили об’єднати в одну групу зразки як яблук, так і картоплі, відібрані з певних типів ландшафтів, що узгоджується з близькістю характеристик відповідних типів ландшафтів.

Вищевказані дії дозволили правильно ідентифікувати зразки харчової сировини тестових вибірок за допомогою PNN, при цьому для зразків картоплі $T = 91\,\%$, для зразків яблук $T = 82\,\%$, а також за допомогою PNN та методу головних компонент, при цьому для зразків картоплі $T = 78\,\%$, для зразків яблук $T = 77\,\%$.

З метою з’ясування можливості скорочення числа металів, концентрації яких необхідно визначати у зразках картоплі та яблук для надійної ідентифікації їх географічного походження, розрахували коефіцієнти рангової кореляції Спірмена та коефіцієнти кореляції Пірсона, відповідно. Із кожної пари корельованих вмістів металів у зразках картоплі виключили по одному металу, концентрація якого характеризується найбільшою дисперсією. Таким чином, для зразків картоплі число металів скоротили вдвічі. Застосувавши до скороченого масиву даних PNN, отримали високий об’єм тестової вибірки – 22 % (13 зразків), усі зразки якого ідентифікуються вірно. Для зразків

яблук виявили лише одну пару корельованих вмістів металів, тому значне скорочення числа металів, що визначаються, є неможливим.

Розділ 4 «Процедура класифікації об'єктів без апріорної інформації про число класів та про приналежність об'єктів до того чи іншого класу» присвячений визначенню числа класів і знаходженню стійкої класифікації за допомогою поєднання мережі Кохонена «без навчання» та імовірнісної мережі «з навчанням». Алгоритми класифікації «з навчанням» застосовують навчальний набір зразків з відомою класовою приналежністю для вироблення класифікаційних правил; алгоритми класифікації «без навчання» вимагають *a priori* задавати число класів. У цій роботі відмовилися від будь-якої апріорної інформації для розрахунків.

Алгоритм запропонованої процедури складається з таких кроків: 1) класифікація «без навчання» за допомогою мережі Кохонена при різних значеннях числа нейронів (відповідно, і числа класів); 2) визначення груп зразків, що незалежно від числа заданих нейронів віднесені мережею Кохонена до одного й того ж класу; використання цих зразків у якості першої навчальної вибірки для навчання імовірнісної мережі; 3) формування випадковим чином невеликих вибірок із зразків, що не увійшли до першої навчальної вибірки, та їх послідовне пред'явлення на вхід імовірнісної мережі ($\delta = 0.1$) як тестових вибірок; 4) включення зразків кожної тестової вибірки до навчальної вибірки після їх класифікації імовірнісною мережею (навчальна вибірка збільшується, що забезпечує адекватну класифікацію наступних тестових вибірок); 5) проведення перехресної оцінки достовірності для перевірки та уточнення отриманої класифікації.

Процедуру класифікації апробували при класифікації 76 розчинників за набором з 9 фізико-хімічних характеристик (параметр розчинності, поверхневий натяг, дипольний момент, відносна діелектрична проникність, показник заломлення, емпіричний параметр кислотності розчинників як донорів водневих зв'язків, емпіричний параметр полярності та поляризовності, емпіричний параметр полярності Райхардта, структурованість) та при класифікації зразків вод та яблук (див. розділ 3). В результаті отримали розбиття розчинників на 11 класів (табл. 3), що узгоджується з їхньою хімічною природою та результатами класифікації інших масивів даних у низці попередніх робіт[4].

У випадку зразків річкових і джерельних вод м. Харкова отримано по 4 класи об'єктів. Порівняння знайденого розбиття зразків вод на групи з їх походженням свідчить про правильність отриманої класифікації: зразки, відібрані з різних рік та джерел, не перемішані між собою; спостерігається лише об'єднання в одну групу деяких зразків, відібраних з різних рік та джерел (наприклад, рік Харків і Лопань), в силу близькості їх характеристик; у випадку джерельних вод виділено окремий клас, що включає найменш забруднені зразки вод, а у випадку річкових вод – клас, що включає найбільш забруднені зразки.

Необхідно відзначити, що отримана класифікація зразків яблук (2 класи) не відповідає їх географічному походженню. Це пов'язано з тим, що на відміну від масиву даних про вміст металів у зразках вод, масив даних про вміст металів у зразках яблук характеризується значним перекриванням класів (концентрації металів у зразках яблук

[4] P. Gramatica, N. Navas, R. Todeschini // Trends Anal. Chem. – 1999. – V. 18, N 7. – P. 461–471.
M. Chastrette, M. Rajzmann, M. Chanon, K. F. Purcell // J. Am. Chem. Soc. – 1985. – V. 107, N 1. – P. 1–11.
A. R. Katritzky, T. Tamm, Y. Wang, M. Karelson // J. Chem. Inf. Comput. Sci. – 1999. – V. 39. – P. 692–698.
A. R. Katritzky, D. C. Fara, M. Kuanar [e. a.] // J. Phys. Chem. A. – 2005. – V. 109, N 45. – P. 10323– 10341.

різних груп знаходяться майже в одних і тих же межах). В той же час, розподіл зразків яблук на групи відповідає їх рівню забрудненості. Критерієм забрудненості зразків яблук є значення показника

$$PC = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{x_i}{\text{ГДК}_i}, \qquad (9)$$

де N – число металів, що визначалися, x_i – концентрація *i*-го металу, ГДК_i – гранично допустима концентрація *i*-го металу.

Один клас містить зразки зі значеннями $PC \le 2.00$, інший – зразки зі значенням $PC \ge 2.07$. Застосування процедури класифікації на основі поєднання мережі Кохонена та імовірнісної мережі дозволило провести межу ($PC = 2.00$) між зразками з допустимим та високим рівнями забруднення металами. Отримані результати використали для побудови інтерполяційної карти забруднення зразків харчової сировини металами.

Таблиця 3

Запропонована класифікація розчинників

Клас	Розчинник
I	н-пентан, н-гексан, н-гептан, н-октан, ізооктан, н-декан, перфлуоробензен
II	н-тетрадекан, н-гексадекан, н-додекан, циклогексан, метилциклогексан, трихлороетен, тетрахлорометан, бензен, п-ксилен, мезитилен, п-цимен, толуен, о-ксилен, м-ксилен, етилбензен
III	цис-декалін, карбон дисульфід, тетрахлороетен
IV	флуоробензен, дихлорометан, 1,2-дихлороетан, 1,1,1-трихлороетан, 1-бромопропан, тетрагідрофуран, о-крезол, 1,4-діоксан
V	1,1,2,2-тетрахлороетан, 1,2,3-трихлоропропан, хлоробензен, о-дихлоробензен, 1-йодопропан, піридин, тетрагідротіофен
VI	бромобензен, м-дихлоробензен, бромоформ, йодобензен, хінолін, 1,2,4-трихлоробензен, дийодометан, 1,2-дибромоетан, дибромометан, анізол
VII	метанол, етанол, пропан-1-ол, бутан-1-ол, ацетон
VIII	пентан-1-ол, гексан-1-ол, октан-1-ол, декан-1-ол, 1-хлоропропан
IX	етан-1,2-діол, вода, N-метилформамід, N,N-диметилформамід, нітроетан, N,N-диметилацетамід, диметилсульфоксид, пропіленкарбонат, нітрометан, ацетонітрил
X	N-метилпіролідон, гексаметилфосфортриамід, нітробензен, бензонітрил
XI*	хлороформ, бутан-1-амін

* До класу XI увійшли розчинники, які не можна віднести до жодної з груп I–X.

У розділі 5 «Рекомендації щодо вибору параметрів для синтезу нейронних мереж» запропоновано процедуру формування представницької навчальної вибірки та формулу для розрахунку оптимального числа нейронів прихованого шару; наведені рекомендації щодо вибору управляючих параметрів для синтезу нейронних мереж.

Для опису навчальних і тестових вибірок об'єктів, класифікація яких досліджувалася у розділах 2 та 3, розрахували для кожної характеристики середнє значення ($\bar{x}$), медіану ($\hat{x}$), стандартне відхилення (s), розмах (r) та інтерквартильний

розмах ($\hat{r}$). Метою дослідження було визначення найбільш значимих параметрів. За значеннями середніх значень, медіан, стандартних відхилень, розмахів та інтерквартильних розмахів характеристик об'єктів навчальної та тестової вибірок розрахували відповідні суми

$$\overline{X}_{навч} = \sum_{i=1}^{M} \overline{x}_i^{навч},\ \overline{X}_{тест} = \sum_{i=1}^{M} \overline{x}_i^{тест},\ \hat{X}_{навч} = \sum_{i=1}^{M} \hat{x}_i^{навч},\ \hat{X}_{тест} = \sum_{i=1}^{M} \hat{x}_i^{тест},\ S_{навч} = \sum_{i=1}^{M} s_i^{навч},$$

$$S_{тест} = \sum_{i=1}^{M} s_i^{тест},\ R_{навч} = \sum_{i=1}^{M} r_i^{навч},\ R_{тест} = \sum_{i=1}^{M} r_i^{тест},\ \hat{R}_{навч} = \sum_{i=1}^{M} \hat{r}_i^{навч},\ \hat{R}_{тест} = \sum_{i=1}^{M} \hat{r}_i^{тест}, \qquad (10)$$

де M – число характеристик, та модулі різниці між відповідними статистичними характеристиками навчальної та тестової вибірок:

$$\Delta\overline{X} = \left|\overline{X}_{навч} - \overline{X}_{тест}\right|,\ \Delta\hat{X} = \left|\hat{X}_{навч} - \hat{X}_{тест}\right|,\ \Delta S = \left|S_{навч} - S_{тест}\right|,\ \Delta R = \left|R_{навч} - R_{тест}\right|,$$

$$\Delta\hat{R} = \left|R_{навч} - R_{тест}\right|. \qquad (11)$$

Приклад розрахунку цих параметрів для зразків ірисів представлено в табл. 4. Оптимальний об'єм навчальної вибірки для ідентифікації зразків ірисів складає 70 %, і саме для цього значення коефіцієнту T значення параметрів ΔS та ΔR є найменшими. Подібні залежності спостерігалися і для інших пар навчальної / тестової вибірок інших об'єктів. Це дало змогу зробити висновок, що надійну класифікацію зразків тестової вибірки забезпечує навчальна вибірка, розмах та стандартне відхилення якої найменш відмінні від значень зазначених параметрів тестової вибірки.

Таблиця 4

Статистичні параметри для даних про зразки ірисів

Параметр	Об'єм навчальної вибірки, T			
	60 %	65 %	70 %	75 %
P, %	3	4	0	3
$\Delta\overline{X}$	0.30	0.08	0.43	0.27
$\Delta\hat{X}$	0.30	0.10	0.70	0.30
ΔS	0.50	0.13	0.09	0.29
ΔR	0.40	1.10	0.40	2.30
$\Delta\hat{R}$	0.73	0.10	0.30	0.20

Для розрахунку оптимального числа нейронів прихованого шару запропонували формулу, що використовує лише початкові відомості про задачу:

$$n_{neurons} = \left[\frac{n_{training}}{n_{testing}} \cdot k_3 + n_{properties} \cdot k_1\right] \pm n_{classes} \cdot k_2, \qquad (12)$$

де $n_{training}$ – число зразків у навчальній вибірці, $n_{testing}$ – число зразків у тестовій вибірці, $n_{classes}$ – число класів, $n_{properties}$ – число характеристик, k_1, k_2, k_3 – параметри.

Параметри k_1, k_2, k_3 приймають різні значення в залежності від особливостей задачі: $k_1 = 2$, якщо число характеристик не більше 3; $k_1 = 0.5$, якщо число характеристик більше 8; $k_1 = 1$, якщо число характеристик коливається від 3 до 8; $k_2 = 1$, якщо число

класів не більше 3; $k_2 = 0.5$, якщо число класів більше 3; $k_3 = 2$, якщо $\frac{n_{training}}{n_{testing}} \approx 1$, в інших випадках $k_3 = 1$. Знайдені емпірично оптимальні значення числа прихованих нейронів для досліджуваних у цій роботі об'єктів знаходяться в інтервалі, розрахованому згідно з (12).

Надійне розв'язання задач ідентифікації в якісному хімічному аналізі забезпечують двошарові нейронні мережі. Рекомендовано використовувати наступні параметри: алгоритм Левенберга-Марквардта для навчання мережі прямого поширення сигналу, динамічної та каскадної мереж, алгоритм зворотного поширення помилки для навчання мережі Елмана; функцію активації гіперболічний тангенс для прихованого шару та лінійну функцію активації для вихідного шару; значення відхилення радіально-базисної функції активації 0.1 для імовірнісної мережі.

ВИСНОВКИ

В роботі розв'язано актуальну наукову задачу адаптації апарату штучних нейронних мереж та розробки рекомендацій щодо вибору їх архітектури та параметрів для надійної ідентифікації та кластеризації об'єктів в якісному хімічному аналізі.

1. На основі обробки багатовимірних модельних, еталонних та експериментальних даних, що характеризуються наявністю пропусків та спостережень, що різко виділяються, встановлено, що найвищу надійність результатів якісного хімічного аналізу при класифікації «з навчанням» (ідентифікації) серед вивчених алгоритмів забезпечують імовірнісна та динамічна нейронні мережі.
2. Для формування представницької навчальної вибірки для правильного навчання алгоритмів класифікації достатньо керуватися значеннями стандартних відхилень та розмахів. Для високої надійності класифікації необхідні мінімальна відмінність значень стандартного відхилення та розмаху (за абсолютною величиною) навчальної вибірки від значень цих характеристик тестової вибірки.
3. На основі детального вивчення алгоритмів штучних нейронних мереж сформовані рекомендації щодо вибору архітектури та параметрів для синтезу нейронних мереж, що забезпечують надійну класифікацію хіміко-аналітичних даних: доцільно використовувати алгоритм навчання Левенберга-Марквардта, сигмоїдальну (гіперболічний тангенс) і лінійну функції активації; оптимальне число нейронів прихованого шару знаходиться в інтервалі, для розрахунку якого запропоновано формулу, до якої входять лише початкові відомості про задачу (число зразків в навчальній та тестовій вибірках, число характеристик та число класів).
4. Дослідження стійкості алгоритмів ідентифікації об'єктів до наявності в масивах даних пропусків та «промахів» доцільно проводити за розробленою в роботі простою процедурою, що дозволяє вносити до масивів даних похибки, розподіл яких відрізняється від нормального.
5. Знайдено оптимальні архітектуру та параметри п'яти типів нейронних мереж для надійного розв'язання складної задачі класифікації розчинників за сольватохромними параметрами (лише три сольватохромні параметри, великий розмах значень параметрів у класах, близькість параметрів для різних середовищ). Найвищу стійкість до наявності у вихідних даних промахів має імовірнісна нейронна мережа.

6. Обґрунтовані в роботі архітектура та параметри нейронних мереж забезпечують 100 %-ву надійність і робастність ідентифікації зразків річкових і джерельних вод м. Харкова, відібраних у різні сезони та роки, за вмістом 8 перехідних і важких металів (класифікація «з навчанням»). Алгоритми динамічної та імовірнісної нейронних мереж ефективно визначають походження зразків вод навіть за відсутності інформації про вміст одного з металів.

7. Процедура кластеризації на основі комбінації алгоритмів мережі Кохонена, імовірнісної мережі та перехресної оцінки достовірності (крос-валідації) забезпечує стійку класифікацію зразків без апріорної інформації про число класів і навчальну вибірку. Процедура виявилася ефективною для знаходження такого розбиття на групи 76 розчинників за набором з 9 фізико-хімічних характеристик, що має фізичний зміст; для ідентифікації за походженням зразків річкових і джерельних вод за даними про вміст металів та для розподілення зразків яблук між групами, що відрізняються рівнем забрудненості.

8. Задача ідентифікації за географічним походженням (типами ландшафтів) харчової сировини за вмістом у зразках металів характеризується особливою складністю: наявні спостереження, що різко виділяються, розподіл концентрацій металів у зразках відрізняється від нормального. Поєднання непараметричних статистичних методів (критеріїв Краскела-Уолліса та Уілкоксона-Манна-Уітні, розрахунок коефіцієнтів рангової кореляції Спірмена) з методом головних компонент та алгоритмом імовірнісної нейронної мережі дозволяє визначити найбільш інформативні характеристики зразків, об’єднувати подібні типи ландшафтів, скорочувати кількість металів, концентрації яких треба вимірювати, та забезпечує надійну ідентифікацію зразків за походженням.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Краснянчин Я. Н. Надежность идентификации аналитов с помощью искусственных нейронных сетей / **Я. Н. Краснянчин**, А. В. Пантелеймонов, Ю. В. Холин // Вісник Харківського національного університету. – 2010. – № 895. Хімія. Вип. 18 (41). – С. 39–46.

Здобувач апробувала алгоритми штучних нейронних мереж на модельних і тестових наборах даних, брала участь в обговоренні результатів та підготовці публікації.

2. Краснянчин Я. Н. Некоторые аспекты параметризации искусственных нейронных сетей в задачах качественного химического анализа / **Я. Н. Краснянчин**, А. В. Пантелеймонов, Ю. В. Холин // Вісник Харківського національного університету. – 2010. – № 932. Хімія. Вип. 19 (42). – С. 170–181.

Здобувач дослідила закономірності функціонування різних алгоритмів штучних нейронних мереж при зміні управляючих параметрів, запропонувала процедуру формування навчальної вибірки, брала участь в обговоренні результатів та підготовці публікації.

3. Краснянчин Я. Н. Хемометрические методы в контроле подлинности продуктов питания и пищевого сырья / **Я. Н. Краснянчин**, А. В. Пантелеймонов, Ю. В. Холин // Методи та об’єкти хімічного аналізу. – 2010. – Т. 5, № 3. – С. 118–147.

Здобувач провела аналіз літературних даних, брала участь в обговоренні результатів та підготовці публікації.

4. Pushkarova Yaroslava. The classification of solvents based on solvatochromic characteristics: the choice of optimal parameters for artificial neural networks / **Yaroslava**

Pushkarova, Yuriy Kholin // Central European Journal of Chemistry. – 2012. – Vol. 10, N 4. – P. 1318–1327.

5. 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 національного університету. – 2012. – № 1026. Хімія. Вип. 21 (44). – С. 212–217.

Здобувач апробувала розроблену процедуру кластеризації без залучення апріорної інформації про число класів та про приналежність об'єктів до того чи іншого класу при класифікації зразків річкових і джерельних вод, брала участь в обговоренні результатів та підготовці публікації.

6. 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, № 4. – С. 184–191.

Здобувач реалізувала і застосувала до встановлення географічного походження зразків харчової сировини ряд статистичних методів, запропонувала процедуру скорочення числа металів, концентрації яких необхідно визначати для надійної ідентифікації географічного походження харчової сировини, брала участь в обговоренні результатів та підготовці публікації.

7. 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 // Вестник Московского университета. Серия 2. Химия. – 2012. – Т. 53, № 6. – С. 405–412.

Здобувач реалізувала та застосувала до ідентифікації походження зразків річкових і джерельних вод м. Харкова ряд алгоритмів штучних нейронних мереж, оцінила стійкість алгоритмів до наявності пропусків у вихідних даних, запропонувала спосіб оцінки оптимального числа прихованих нейронів, брала участь в обговоренні результатів та підготовці публікації.

8. Краснянчин Я. Н. Классификация соединений с помощью искусственных нейронных сетей / **Я. Н. Краснянчин**, А. В. Пантелеймонов, Ю. В. Холин // VI Всеукраїнська конференція молодих вчених, студентів та аспірантів з актуальних питань хімії, 3–6 червня 2008 р. : тези доп. – Харків, 2008. – С. 22.

Здобувач реалізувала та застосувала до класифікації різних наборів даних ряд алгоритмів штучних нейронних мереж, виступила з усною доповіддю.

9. Краснянчин Я. Н. Применение искусственных нейронных сетей к решению задач идентификации продуктов питания / **Я. Н. Краснянчин**, А. В. Пантелеймонов // «Хімічні Каразінські читання – 2010» : Друга Всеукраїнська наукова конференція студентів та аспірантів, 19–22 квітня 2010 р. : тези доп. – Харків, 2010. – С. 73.

Здобувач застосувала алгоритми штучних нейронних мереж до ідентифікації об'єктів за даними багатовідгукового експерименту, виступила з усною доповіддю.

10. Холин Ю. В. Качественный химический анализ как задача классификации объектов / Ю. В. Холин, А. В. Пантелеймонов, **Я. Н. Краснянчин** // Наукова конференція присвячена 100 річниці з дня народження проф. І. В. П'ятницького, 10–13 жовтня 2010 р. : тези доп. – Київ, 2010. – С. 92.

Здобувач провела аналіз літературних даних, брала участь в обговоренні результатів.

11. Следзевская А. Б. Хемометрические методы исследования образцов питьевой воды из различных источников г. Харькова / А. Б. Следзевская, **Я. Н. Краснянчин**, А. В. Пантелеймонов // «Хімічні Каразінські читання – 2011» : Третя Всеукраїнська наукова конференція студентів та аспірантів, 18–21 квітня 2011 р. : тези доп. – Харків, 2011. – С. 63.

Здобувач визначила оптимальний об'єм навчальної вибірки для навчання алгоритмів класифікації, брала участь в обговоренні результатів та підготовці публікації.

12. Искусственные нейронные сети в решении задач качественного химического анализа / **Я. Н. Краснянчин**, А. В. Пантелеймонов, О. И. Юрченко, Ю. В. Холин // Річна Сесія Наукової Ради НАН України з проблеми «Аналітична хімія», 16–20 травня 2011 р. : тези доп. – Гурзуф, 2011. – С. 38.

Здобувач обґрунтувала вибір архітектури та параметрів штучних нейронних мереж для надійної ідентифікації розчинників, оцінила стійкість алгоритмів нейронних мереж до наявності пропусків і похибок у вихідних даних характеристик, виступила з усною доповіддю.

13. Краснянчин Я. Н. Классификация растворителей с применением хемометрических методов / **Я. Н. Краснянчин**, А. В. Пантелеймонов, Ю. В. Холин // XVIII Українська конференція з неорганічної хімії за участю закордонних учених в рамках Міжнародного року хімії ООН, 27 червня – 1 липня 2011 р. : тези доп. – Харків, 2011. – С. 254.

Здобувач запропонувала процедуру класифікації без залучення апріорної інформації про число класів та про приналежність об'єктів до того чи іншого класу та апробувала її при класифікації розчинників, виступила із стендовою доповіддю.

14. Пушкарева Я. Применение хемометрических и статистических методов для установления географического происхождения овощей / **Я. Пушкарева**, А. Следзевская, П. Семибратова // X Всеукраїнська конференція молодих вчених та студентів з актуальних питань хімії, 17–19 квітня 2012 р. : тези доп. – Харків, 2012. – С. 88.

Здобувач реалізувала процедуру оцінки якості зразків харчової сировини, виступила з усною доповіддю.

15. Identification of geographic origin of vegetables and fruits with the use of statistical and chemometric methods / **Yaroslava Pushkarova**, Anastasiia Sliedzevska, Polina Semybratova, Alla Nekos, Yuriy Kholin // 15th International symposium of students and young mechanical engineers «Advances in Chemical and Mechanical Engineering», 16–19 May 2012 : abstracts. – Gdansk, 2012. – P. 201.

Здобувач реалізувала алгоритм імовірнісної нейронної мережі, визначила оптимальне значення відхилення функції активації, брала участь в обговоренні результатів та підготовці публікації.

16. Искусственные нейронные сети в решении задач классификации, дискриминации и идентификации / **Я. Н. Пушкарева**, А. Б. Следзевская, О. И. Юрченко, А. Н. Некос, Ю. В. Холин // Річна Сесія Наукової Ради з проблеми «Аналітична хімія» НАН України, 3–10 червня 2012 р. : тези доп. – Гурзуф, 2012. – С. 66.

Здобувач сформулювала рекомендації, що спрощують процедуру підбору параметрів нейронних мереж, виступила з усною доповіддю.

АНОТАЦІЯ

Пушкарьова Я. М. Розв'язання задач якісного хімічного аналізу за допомогою штучних нейронних мереж. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата хімічних наук за спеціальністю 02.00.02 – аналітична хімія. – Харківський національний університет імені В. Н. Каразіна Міністерства освіти і науки України, Харків, 2013.

У роботі розв'язано актуальну наукову задачу адаптації апарату штучних нейронних мереж для розв'язання задач ідентифікації та кластеризації у якісному хімічному аналізі. Надано рекомендації щодо вибору архітектури та параметрів (методу навчання, функцій активації) штучних нейронних мереж. Запропоновано формулу для

розрахунку оптимального числа нейронів прихованого шару та процедуру формування представницької навчальної вибірки для правильного навчання алгоритмів класифікації. Вивчено стійкість алгоритмів до варіювання вихідних значень характеристик об'єктів.

Показано, що використання алгоритмів нейронних мереж забезпечує надійну ідентифікацію походження зразків вод і харчової сировини за даними про вміст металів. Алгоритми імовірнісної та динамічної нейронних мереж – ефективні методи класифікації об'єктів за даними багатовідгукового експерименту.

Розроблено процедуру стійкої класифікації об'єктів без апріорної інформації про число класів і навчальну вибірку на основі комбінації алгоритмів мережі Кохонена та імовірнісної мережі. У результаті її застосування запропоновано нову класифікацію розчинників (11 груп) за набором з 9 фізико-хімічних характеристик, що відповідає їх хімічній природі.

Ключові слова: класифікація «з навчанням», класифікація «без навчання», хемометрія, штучна нейронна мережа, якісний хімічний аналіз.

АННОТАЦИЯ

Пушкарева Я. Н. Решение задач качественного химического анализа с помощью искусственных нейронных сетей. – На правах рукописи.

Диссертация на соискание ученой степени кандидата химических наук по специальности 02.00.02 – аналитическая химия. – Харьковский национальный университет имени В. Н. Каразина Министерства образования и науки Украины, Харьков, 2013.

В работе решена актуальная научная задача адаптации аппарата искусственных нейронных сетей к решению задач идентификации и кластеризации в качественном химическом анализе.

Изучена эффективность ряда алгоритмов нейронных сетей. Алгоритмы нейронных сетей испытали при классификации модельных и эталонных наборов данных, а также при классификации растворителей по трем сольватохромным параметрам. Результаты, полученные с применением нейронных сетей, сравнивали с результатами работы традиционных методов классификации (линейный дискриминантный анализ, метод опорных векторов, формальное независимое моделирование аналогий классов, линейный дискриминантный анализ с помощью регрессии на латентные структуры). Предложена процедура внесения в исходные данные характеристик пропусков и погрешностей, распределение которых отличается от нормального. Установлено, что алгоритмы нейронных сетей обладают повышенной устойчивостью к наличию в исходных данных характеристик пропусков и «промахов».

Показана эффективность нейронных сетей в решении таких актуальных задач качественного химического анализа, как контроль подлинности пищевого сырья и объектов окружающей среды.

Обоснованы архитектура и параметры нейронных сетей, которые обеспечивают 100 %-ную надежность и робастность идентификации образцов речных и родниковых вод г. Харькова, отобранных в разные сезоны в течение 2008–2010 гг., по содержанию 8 переходных и тяжелых металлов (классификация «с обучением»). Особое внимание уделено проверке содержательности классификации. Контрольную выборку для проверки правильности классификации речных и родниковых вод составляли образцы,

отобранные в году, следующем за обучением алгоритмов искусственных нейронных сетей. При этом количество определенных в этих образцах концентраций металлов отличалось от количества определенных свойств эталонов, которые использовались для обучения алгоритмов. Установлено, что алгоритмы динамической и вероятностной нейронных сетей эффективно определяют происхождение образцов вод даже в случае отсутствия информации о содержании одного из металлов, описывающих образцы вод обучающей выборки.

Использование непараметрических статистических методов (критериев Краскела-Уоллиса и Вилкоксона-Манна-Уитни, расчет коэффициентов ранговой корреляции Спирмена) совместно с методом главных компонент и алгоритмом вероятностной нейронной сети позволяет установить происхождение образцов яблок и картофеля (идентифицировать типы ландшафтов) по содержанию в них 10 металлов, определить наиболее информативные характеристики образцов, объединить подобные типы ландшафтов и сократить количество металлов, концентрации которых необходимо определять. Образцы были отобраны из разных районов г. Харькова и Харьковской области в течение 2008–2010 гг. Необходимо отметить, что задача установления происхождения образцов растительного материала отличается особенной сложностью: присутствуют резко выделяющиеся наблюдения, распределение концентраций металлов в образцах отличается от нормального.

Изложены рекомендации по выбору архитектуры, параметров и процедур использования искусственных нейронных сетей, обеспечивающих высокую надежность и робастность решения задач классификации в качественном химическом анализе. Целесообразно использовать метод обучения Левенберга-Марквардта и две функции активации – гиперболический тангенс для скрытого слоя и линейную для выходного. Для расчета оптимального числа скрытых нейронов предложено формулу, в которую входят число образцов в обучающей и тестовой выборках, число характеристик и число классов. Для реализации вероятностной сети достаточно задавать значение отклонения радиальной базисной функции активации 0.1.

Наибольшую надежность результатов качественного химического анализа обеспечивают алгоритмы динамической и вероятностной нейронных сетей.

Установлено, что для формирования представительной обучающей выборки для правильного обучения алгоритмов необходимы минимальное отличие значений стандартного отклонения и размаха обучающей выборки от указанных характеристик тестовой выборки.

Разработана процедура нахождения числа классов и устойчивой классификации объектов по данным об их химико-аналитических характеристиках на основе объединения сети Кохонена и вероятностной сети. В отличие от существующих алгоритмов, предложенная процедура не требует привлекать априорную информацию ни о числе классов, ни о составе обучающей выборки. Процедура испытана при классификации 76 растворителей по набору из 9 физико-химических параметров, в результате чего получено 11 химически интерпретируемых групп, а затем верифицирована при классификации образцов вод и яблок по данным о содержании в них металлов. Показано, что отнесение образцов вод к классам соответствует их географическому происхождению, а разделение образцов яблок на группы соответствует различным уровням содержания металлов, что создает возможность

создания интерполяционных карт для определения очагов максимального / минимального риска в контроле качества пищевого сырья. Для проверки правильности и уточнения полученных классификаций использовали процедуру перекрестной проверки достоверности (процедуру кросс-валидации). Разработанный алгоритм можно рекомендовать для эксплораторного анализа химико-аналитических экспериментальных данных.

Ключевые слова: классификация «с обучением», классификация «без обучения», хемометрия, искусственная нейронная сеть, качественный химический анализ.

SUMMARY

Pushkarova Ya. N. Solving tasks of qualitative chemical analysis with the use of artificial neural networks. – Manuscript copyright.

The thesis for the candidate's degree, speciality 02.00.02 – analytical chemistry. – V. N. Karazin Kharkiv National University Ministry of Education and Science of Ukraine, Kharkiv, 2013.

Within the thesis the urgent scientific problem of adapting artificial neural networks for the reliable identification and clustering of objects in qualitative chemical analysis has been solved. The recommendations for the choice of optimal architecture and parameters (types of transfer functions, training method) of artificial neural networks have been developed. The rule for the determination of the optimal number of hidden neurons and the procedure for forming the representative training set have been proposed. Also, the influence of data variation on the stability of classification was examined.

Artificial neural networks were shown to be the efficient tool for the identification of the geographical origin of waters and foodstuff. The initial experimental data consisted of metal concentrations in samples. The probabilistic and dynamic neural networks are recommended as the most robust algorithms for the classification of objects.

The novel classification procedure based on a combination of the Kohonen and the probabilistic neural networks has been developed. The procedure does not require any a priori information about the number of classes and the patterns in the training set. Its capability has been demonstrated for the set of solvents, which have been classified into eleven classes based on nine physical-chemical characteristics.

Keywords: supervised classification, unsupervised classification, chemometrics, artificial neural network, qualitative chemical analysis.