УДК 628.316.6.094.3:637.1

APPLICATION OF MACHINE LEARNING ALGORITHMS IN STATISTICAL MODELS FOR PREDICTION AND CALCULATION OF OZONING PROCESSES

V. Zakharov, O. Ustinov, Yu. Zmievskii, V. Myronchuk
National University of Food Technologies

Key words:
Machine learning
Ozoning
Organic pollution
Naive Bayes classifier
Ozone-gas composition

Article history:
Received 02.07.2019
Received in revised form 25.07.2019
Accepted 14.08.2019

Corresponding author:
V. Zakharov
E-mail:
npnuht@ukr.net

ABSTRACT

The paper presents the probabilistic and statistical model developed by the authors for calculating and predicting the efficiency of ozonation processes using the technology of "machine learning". The model is implemented as a software. The basis of the calculation algorithm is Bayes' theorem. The program code is written in Python 2.6.8 and Bash.

There were proposed 4 classes that correspond to a certain percentage of dissolved ozone in the liquid phase, since this index is one of the main parameters in determining the effectiveness of the ozonation process. The principle of forming a training sample is to create a set of events in which the set of values of the selected parameters correspond to a certain class.

As a result of the statistical analysis of the probability distribution of various parameters in classes, it was found that ozone concentration in the ozone-gas mixture and temperature are most affected by the ozonation process. The more events are present in the training sample, the more precisely the classification takes place.

The result of the work is a probabilistic-statistical model using the technology of "machine learning" (expert system) and testing of this model for the ozonation process. The model allows to determine the efficiency of the ozonation process depending on the given values of temperature and ozone concentration in the ozone-gas mixture.

Within the limits of the temperature 0...35°C and the initial concentration of ozone 20...240 g/m^3, the accuracy of the forecast is 91%. Also, the program has implemented the function of machine learning on the principle of "Supervised Learning". It is implemented by an additional module, which, after the determination of the user of dissolved ozone, asks for confirmation of the correctness of the results. If the user confirms the correctness of the classification, then the given event is entered into the training sample.

The presented model can be adapted to simulate a wide range of tasks related to barometric processes.

DOI: 10.24263/2225-2924-2019-25-4-3

ЗАСТОСУВАННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ В ІМОВІРНІСНО-СТАТИСТИЧНИХ МОДЕЛЯХ ДЛЯ ПРОГНОЗУВАННЯ ТА РОЗРАХУНКУ ПРОЦЕСІВ ОЗОНУВАННЯ

В. В. Захаров, О. А. Устінов, Ю. Г. Змієвський, В. Г. Мирончук
Національний університет харчових технологій

У статті представлено розроблену авторами ймовірнісно-статистичну модель для розрахунку та прогнозування ефективності процесів озонування із застосуванням технології «машинного навчання». Модель реалізовано у вигляді програмного забезпечення. В основі розрахункового алгоритму лежить теорема Баєса. Код програми написано на мовах Python 2.6.8 та Bash.

Запропоновано чотири класи, які відповідають певному відсотку розчиненого озону в рідкій фазі, оскільки цей показник — один з головних параметрів при визначенні ефективності процесу озонування. Принцип формування навчальної вибірки полягає у створенні набору подій, у яких сукупність значень обраних параметрів відповідає певному класу. Після проведення статистичного аналізу розподілів імовірностей різних параметрів за класами з'ясовано, що на процес озонування найбільше впливають концентрація озону в озоно-газовій суміші і температура.

У результаті проведеного дослідження побудовано ймовірнісно-статистичну модель із застосуванням технології «машинного навчання» та проведено апробацію моделі для процесу озонування. Модель дає змогу визначити ефективність процесу озонування залежно від заданих значень температури та концентрації озону в озоно-газовій суміші.

У межах значень температури 0…35°С та початкової концентрації озону 20…240 г/м^3 точність прогнозу становить 91%. Також у програмі реалізовано функцію «машинного навчання» на принципі «Supervised Learning» додатковим модулем, який після визначення користувачем розчиненого озону видає запит на підтвердження правильності отриманих результатів. Якщо користувач підтверджує правильність класифікації, то задана подія вноситься до навчальної вибірки як достовірна.

Представлена модель може бути адаптована для моделювання широкого класу задач, пов'язаних з баромембранними процесами.

***Ключові слова**: машинне навчання, озонування, органічне забруднення, наївний Баєсів класифікатор, озоно-газова суміш.*

Постановка проблеми. З розвитком технологій збільшуються потоки інформації і виникає проблема обробки експериментальних даних. Для автоматизації процесів обробки експериментальних даних та обробки інформації в цілому застосовують різноманітні підходи. Одним із сучасних напрямків розвитку інформаційних технологій є «машинне навчання» (англ. Machine Learning) — клас штучного інтелекту, математична дисципліна, що вико-

ристовує методи математичної статистики, теорії ймовірностей, чисельні методи оптимізації для обробки й аналізу великих масивів даних [1].

Розділяють два напрямки: індуктивне навчання (виявлення закономірностей з емпіричних даних) та дедуктивне (формування баз знань). Формально дедуктивне навчання відносять до експертних систем (комп'ютерна система, здатна частково або повністю замінити спеціаліста-експерта у вирішенні проблемної ситуації), а під терміном «машинне навчання» часто мають на увазі індуктивні алгоритми навчання [1].

Більшість задач «машинного навчання», так чи інакше, зводяться до задач регресії (дослідження впливу незалежних змінних $\{X\_\{i\}\}$ на незалежну Y) або до задач класифікації. Класифікація алгоритмів за типом навчання [1—3]:

Supervised Learning. Навчання «з учителем», задаються приклади з «вірними відповідями». Так працює частина антиспамових фільтрів. Користувач відмічає пошту, як спам, алгоритм це запам'ятовує і порівнює вхідні з відміченими повідомленнями (навчальна вибірка) та сортує на «спам» і «не спам».

Unsupervised learning. Задається набір прикладів та алгоритм класифікації цих прикладів.

Semi-supervised learning. Задається набір прикладів, в якому лише частина прикладів «вірна» і алгоритм для аналізу.

Reinforcement learning. Певна функція аналізує роботу програми і ставить оцінку дій програми. Оцінки впливають на подальшу роботу програми.

Transduction. Використовуються певні логічні конструкції для визначення вірних відповідей.

Learning to learning. Задаються декілька суміжних задач, навчання полягає у пошуку спільних закономірностей.

В основу будь-якої моделі або програми, що використовує технологію «машинного навчання», покладено певні алгоритми. Найвідоміші з них: C4.5, метод К-середніх, метод опорних векторів, Apriori, EM-алгоритм, kNN, Баєсовський класифікатор, CART тощо. Кожен з перерахованих алгоритмів має свої переваги та недоліки при застосуванні у різних задачах.

У пропонованому дослідженні розроблено ймовірнісно-статистичну модель, в основу якої покладено алгоритм Баєсовський класифікатор із технологією «машинного навчання» за методикою «Supervised Learning».

Через складність і різноманіття перебігу процесів під час озонування існує проблема його розрахунку та моделювання. Особливо гостро це питання стоїть на підприємствах, де оброблювана озоном продукція має у своєму складі широкий спектр органічних домішок [4]. На практиці застосовують пілотні та експериментальні установки з використанням надмірної кількості озону, що часто нераціонально та затратно. За результатами таких випробувань формуються рекомендації щодо застосуванню озонаторів на підприємствах [5—8].

Апробацію представленої моделі проводили для визначення ефективності процесу озонування для окислення органічних домішок та їх подальшого видалення на сорбційних фільтрах, що має значну перспективу використання в харчовій промисловості. Застосовуючи цей процес, можна видаляти небажані органічні домішки з оброблюваних розчинів і забезпечувати мікробіоло-

гічну чистоту технологічного обладнання [4—7]. Значною перевагою озонування є його екологічність і безпечність для харчових виробництв [4; 7—8].

На сьогодні існує вкрай обмежена кількість методів, що дають змогу з великою точністю прогнозувати ефективність зазначених процесів, тому апробація представленої ймовірнісно-статистичної моделі на процесі озонування є доцільною.

Проте розвиток техніки та особливо інформаційних технологій відкриває нові можливості застосовування відомих методів математичної статистики [1—3], тому авторами розроблений метод визначення параметрів процесу озонування та прогнозування його ефективності у вигляді класифікатора, який заснований на теоремі Баєса із застосуванням технологій «машинного навчання».

Мета дослідження: розроблення ймовірнісно-статистичної моделі, що дає змогу прогнозувати ефективність процесів озонування. Апробація створеної моделі проводилась для озонування в процесі утилізації нанофільтраційного пермеату молочної сироватки.

Методи і обладнання. Програма для розрахунку процесу озонування була написана на мові програмування Python (Пайтон) версії 2.6.8. Python — потужна мова програмування, має ефективні структури даних високого рівня, простий та водночас ефективний підхід до об'єктно-орієнтовного програмування. Ця мова програмування має зручний та інтуїтивно зрозумілий синтаксис, динамічну обробку типів, вбудований інтерпретатор і велику кількість різноманітних модулів для широкого спектра задач, що робить її придатною для розробки прикладних програм [10; 11].

В основу розрахунку було покладено теорему Баєса. Припущення незалежності параметрів полягає в тому, що, незважаючи на незалежність або залежність параметрів один від одного та їхніх зв'язків між собою, вважається, що при визначенні класу кожний з цих параметрів вносить свій окремий і незалежний внесок [1; 2; 11; 12]. Хоча це припущення справедливе не завжди, але в певних випадках залежність ознак однакова для всіх класів і взаємно компенсується.

У задачах класифікації теорема Баєса дає змогу розрахувати ймовірність того, що певний набір значень параметрів (ознак) відноситься до певного класу. Тобто об'єкт або «подія» (як набір значень ознак) відноситься до класу з певною імовірністю. В задачі класифікації є матриця об'єкти-ознаки, $X = \{x_i\}$ — множина ознак, C_i — конкретний клас (а C — множина всіх класів) [3].

Розглянемо теоретичну основу та застосування цього алгоритму на прикладі: Нехай подія A може здійснитись лише при умові виконання (появи) однієї з несумісних подій (гіпотез) B_1, B_2, ..., B_n, що утворюють повну групу подій. Якщо подія A вже відбулась, то ймовірності гіпотез можна визначити за формулами Баєса:

$$P(B_i / A) = \frac{P(B_i) \cdot P(A / B_i)}{P(A)}, \quad i = 1,2,\ldots,n; \tag{1}$$

$$P(A) = P(B_1) \cdot P(A / B_1) + P(B_2) \cdot P(A / B_2) + \ldots + P(B_n) \cdot P(A / B_n), \tag{2}$$

де $P(A|B_i)$ — функція правдоподібності, що визначається нашою моделлю, тобто створюємо модель збору даних, що залежить від параметра, який і цікавить нас.

Для інтерполяції даних за допомогою прямої $y = ax + b$ (таким чином ми припускаємо, що всі дані мають лінійну залежність з накладеним на неї гаусовим шумом з відомою дисперсією). Тоді a і b — це необхідні параметри, тому потрібно дізнатися їхні найбільш імовірні значення, а функція правдоподібності — гаус із середнім, заданим рівнянням прямої і даної дисперсією [1—3].

$P(B_i)$ — апріорна імовірність включає в себе інформацію, відому до проведення аналізу. Наприклад, відомо, що пряма повинна мати позитивний нахил або що значення в точці перетину з віссю x має бути позитивним. Все це і не тільки можна втілити у нашому аналізі.

$P(B_i|A)$ — апостеріорна ймовірність певного класу, тобто значення цільової змінної при конкретному наборі значень ознак.

Якщо значення безперервних ознак не описується нормальним розподілом, то за допомогою відповідного перетворення необхідно привести їх до такого розподілу [2].

Якщо в тестовому наборі даних є певне значення категорійної ознаки, яке не зустрічалося в навчальному наборі даних, то модель присвоїть нульову імовірність цього значення і не зможе зробити прогноз. Це явище відоме під назвою «нульова частота» (zero frequency). Зазначену проблему можна вирішити за допомогою згладжування. Одним із найпростіших методів є згладжування Лапласа (Laplace smoothing) [1; 2].

У випадку, коли дві ознаки мають високу кореляцію, одну з них слід видалити, інакше вони будуть завищувати свою значимість [1—3; 11].

Результати та обговорення. У зв'язку із складністю прогнозування процесу озонування було вирішено провести апробацію розробленої ймовірнісно-статистичної моделі для визначення ефективності процесу озонування.

Ключовим показником ефективності озонування є кількість розчиненого озону в оброблюваному розчині *RO*. Це пов'язано з тим, що у реакцію з обраними для обробки речовинами в розчині вступає саме розчинений у рідкій фазі озон. Головним чином ця величина залежить від pH розчину, його температури (T), вмісту речовин, здатних окислитися озоном (M) та концентрації озону в озоно-газовій суміші (C_0):

$$RO = f(\text{pH}, T, M, C_{O_3}). \quad (3)$$

Для цього було сформовано вибірку даних і проведено її статистичний аналіз. Вхідними параметрами було обрано температуру T та концентрацію озону в озоно-газовій суміші x, шуканою змінною виступає розчинність озону *RO*.

Сукупність значень набору параметрів відповідає частці розчиненого озону в рідкій фазі Y%. Було обрано чотири класи: «C_1», «C_2», «C_3», «C_4» і сформовано їх відповідно до різних діапазонів значень Y (табл. 1).

На рис. 1 представлені емпіричні кореляції, з яких сформовано навчальну вибірку даних.

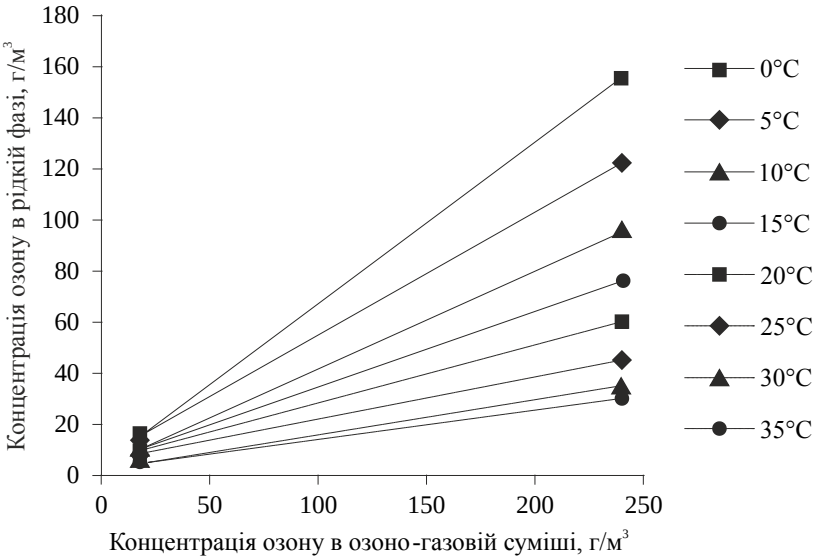

Рис. 1. Експериментальні залежності концентрації розчиненого озону для різних початкових концентрацій озону та різних температур

Таблиця 1. **Поділ діапазону значень розчинності озону на класи**

Клас	C_1	C_2	C_3	C_4
Діапазон значень, %	$60 \le Y < 80$	$40 \le Y < 60$	$20 \le Y < 40$	$0 \le Y < 20$

Розподіл класів зроблено з таких міркувань: за основу взято клас C_2 (40…60% розчиненого озону) — така область значень розчиненого озону в рідкій фазі вважається доцільною для використання з точки зору балансу між затратами на озонування й отриманим ефектом. Для інших класів було обрано крок у 20%, де C_1 клас — вищий за середній (доцільний); C_3 — доцільно лише у разі відсутності інших альтернатив, окрім застосування озонування; C_4 — недоцільно, оскільки економічні затрати на вироблення озону значно перевищуватимуть отриманий ефект. Випадки розчинності вище 80% не зафіксовані в експериментальних даних, тому такий клас не враховувався.

Принцип формування класів і параметрів навчальної вибірки представлено в табл. 2.

Таблиця 2. **Формування класів і параметрів навчальної вибірки**

№	t, °C	X, г/м^3	Y, %	Клас
1.	0	240	74	C_1
2.	5	120	47	C_2
3.	0	70	66	C_1
4.	10	200	40	C_3
…	…	…	…	…

На основі зазначеної навчальної вибірки програма прогнозує найбільш імовірний діапазон (клас) значень розчиненого озону.

З експериментальних даних було сформовано статистичну вибірку, на рис. 2, 3 зображено отримані полігони відносних частот.

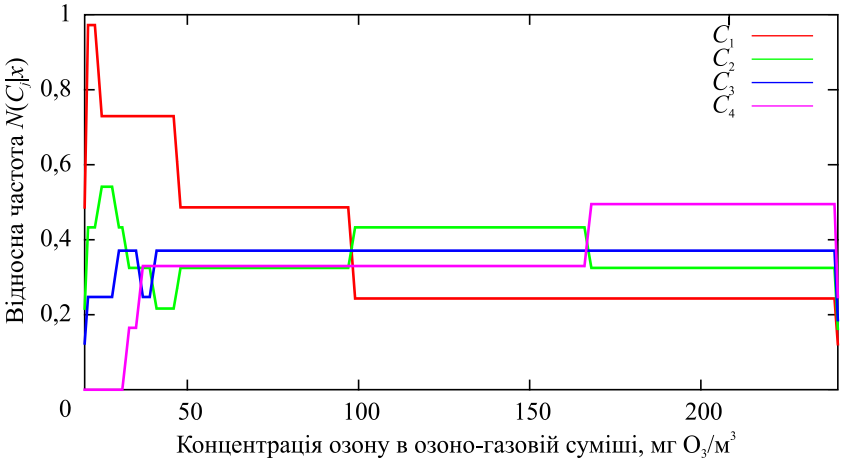

Рис. 2. Полігони відносних частот параметра x для кожного класу

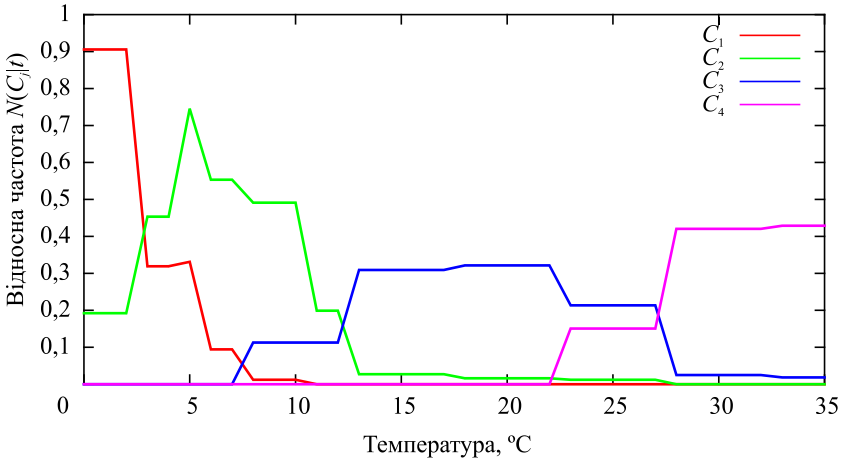

Рис. 3. Полігони відносних частот параметра *T* для кожного класу

Для вирішення проблеми нульових частот застосовано згладжування методом Безьє (рис. 4, 5) (із значенням параметра $n = 3$, використовувались поліноми Бернштейна), співвідношеннями:

$$P(t) = \sum_{i=0}^{n} P_i \cdot J_{n,i}(t), \quad 0 \le t \le 1; \tag{4}$$

$$J_{n,i}(t) = C_n^i \cdot t^i \cdot (1-t)^{n-i}. \tag{5}$$

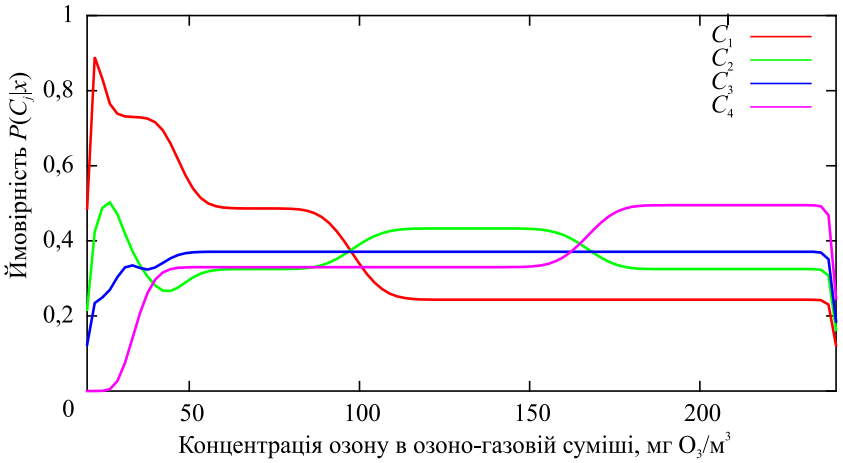

Рис. 4. Розподіл імовірностей належності до класів для параметра *x*

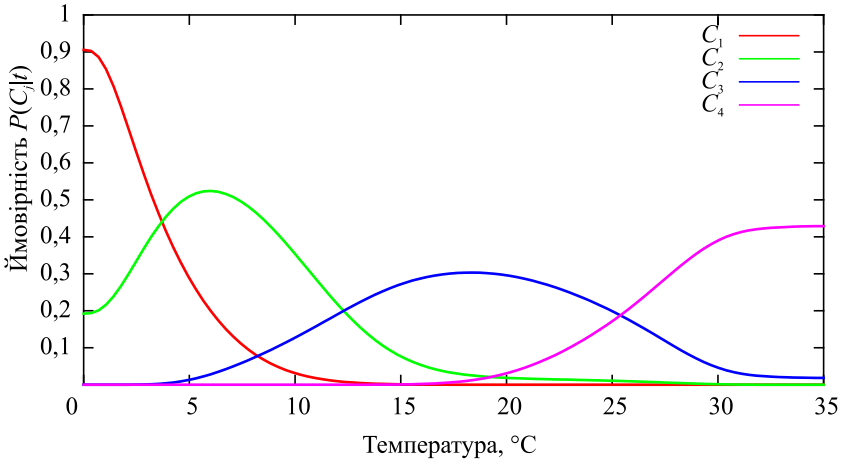

Рис. 5. Розподіл імовірностей належності до класів для параметра *T*

Аналіз мод частотних полігонів і математичного сподівання розподілів імовірностей M_0 показав (табл. 3), що температура має значний вплив на ефективність процесу озонування, а також дає змогу визначити належність до класу.

Таблиця 3. **Статистичний аналіз розподілів для температури**

Діапазон температур, °С	Математичне сподівання M_0	Належність до класу
0—2,5	0	C_1
3—10	6,5	C_2
15—22	17,5	C_3
30—35	33	C_4

Другий параметр *x* — концентрація озону в газовій суміші, показав низьку інформативність для визначення ефективності процесу озонування (визна-

чення класу). Розподіл для класу C_1 (рис. 2, 4) можна пояснити так, що при малій концентрації майже весь озон розчиняється, при збільшенні концентрації спостерігається поступове зменшення кількості розчиненого озону. Починаючи з концентрації x = (*110~240*) г(O^3)/м^3, імовірність належності процесу до класу C_1 становить ~27% і не змінюється. Розподіли на класи для x сильно накладаються один на одний і лише в діапазоні ($0 \le x \le 60$) г(O^3)/м^3 чітко виділяється клас C_1. Проте спостерігаються певні діапазони для різних класів і x може бути застосований, як додатковий параметр для класифікації.

Таблиця 4. **Статистичний аналіз розподілів для початкової концентрації озону**

Діапазон концентрацій, г(O^3)/м^3	Належність до класу
0—60	C_1
110—160	C_2
50—240	C_3
170—240	C_4

Принцип дії алгоритму наївного Баєсового класифікатора такий:

1. Користувач вводить запит, вказуючи температуру T_0 [°C] та концентрацію озону в газовій фазі x_0 [г/м^3].

2. Розраховується апріорна ймовірність для кожного класу:

$$P(C_j) = \frac{N(C_j)}{N}, \tag{6}$$

де $N(C_j)$ — кількість записів у навчальній вибірці, що відповідають даному класу C_j; N — об'єм навчальної вибірки, $C_j = \{C_1, C_2, C_3, C_4\}$ — сукупність класів.

3. Перевірка апріорних ймовірностей за класами:

$$\sum P(C_i) = 1, \tag{7}$$

4. Визначаємо апріорну ймовірність предикторів:

$$P(x = x_0) = \frac{N(x_0)}{N}; \tag{8}$$

$$P(T = T_0) = \frac{N(T_0)}{N}, \tag{9}$$

де $N(x_0)$, $N(T_0)$ — частоти значень параметрів $x=x_0$ та $T=_0$ відповідно.

5. Розрахунок «правдоподібності» для кожного параметра за всіма класами:

$$P(x = x_0 \mid C_j) = \frac{N(x_0 \mid C_j)}{N_j}; \tag{10}$$

$$P(T = T_0 \mid C_j) = \frac{N(T_0 \mid C_j)}{N_j}, \tag{11}$$

де $N(x_0|C_j)$, $N(T_0|C_j)$ — частоти значень параметрів, при певному класі, для x_0 і T_0 відповідно; N_j — об'єм вибірки для певного класу.

6. За формулою Байєса для кожного класу розраховуємо ймовірність того, що такий набір параметрів $x = x_0$ та $T = T_0$ відповідає певному класу C_j:

$$P(C_j \mid x = x_0, T = T_0) = \frac{P(x = x_0 \mid C_j) \cdot P(T = T_0 \mid C_j) \cdot P(C_j)}{P(x = x_0) \cdot P(T = T_0)}. \qquad (12)$$

7. Класифікатор визначає найбільшу ймовірність належності до класу і видає прогноз користувачу.

8. Якщо користувач підтверджує факт того, що класифікація пройшла вірно, то програма доповнює навчальну вибірку записом (значення x_0, T_0 відповідають певному класу C_j).

Функція самовдосконалення програми реалізована за технологією «машинного навчання» в режимі «Supervised Learning». Це навчання з учителем, тобто користувач підтверджує або спростовує вірність розрахованого результату.

Серед переваг представленої ймовірнісно-статистичної моделі можна виділити такі:

1. Алгоритм не потребує значних затрат комп'ютерних ресурсів.

2. Коли припущення про незалежність вхідних параметрів виконується, то алгоритм працює з високою точністю (~90%).

3. Прогнозування краще працює з категорійними ознаками, ніж з безперервними. Для безперервних ознак передбачається нормальний розподіл, що виконується не у всіх випадках.

Головним недоліком є те, що при зростанні кількості ознак (сотні ознак та більше) алгоритм перестає бути ефективним.

Висновки

У результаті проведеного дослідження було розроблено ймовірнісно-статистичну модель з технологією «машинного навчання» (в режимі Supervised Learning), що дає змогу робити прогнози для процесів озонування. Модель апробовано для визначення ефективності процесу озонування залежно від початкових параметрів (температура T, початкова концентрація озону x). Для діапазону температур 0..35°С та початкових концентрацій озону 20..240 (г/м^3) модель дає прогнози з точністю 91%.

У перспективі отриманий комплекс з допомогою «машинного навчання» може стати потужним інструментом в руках інженерних спеціалістів і науковців для дослідження технологій озонування, а також може бути адаптований для застосування у сферах баромембранних процесів. Програмна реалізація моделі проста у використанні, з вбудованою можливістю самовдосконалення за технологією «машинного навчання», що поступово буде збільшувати точність та ефективність запропонованої моделі і відповідного програмного забезпечення.

Статистичний аналіз даних показав, що для озонування температура процесу є одним із найвпливовіших параметрів, що впливає на розчинність озону в рідкій фазі.

Література

1. Thiemann N., Igel C., Wintenberger O., Seldin Y. A Strongly Quasiconvex PAC-Bayesian. *Algorithmic Learning Theory (ALT)*. 2017.

2. Krause O., Arbonès D. R., Igel C. CMA-ES with Optimal Covariance Update and Storage Complexity. *Advances in Neural Information Processing Systems (NIPS)*. 2016.

3. Dogan Ü., Glasmachers T., Igel C. A Unified View on Multi-class Support Vector Classification. *Journal of Machine Learning Research*. 2016. № 17(45).

4. Pandiselvam R., Sunoj S., Manikantan M. R., Kothakota A., Hebbar K. B. Application and Kinetics of Ozone in Food Preservation. *Ozone: Science & Engineering*. 2017. № 39(2). P. 115—126.

5. Cullen P. J., Tiwari B. K., O'Donnell C. P., Muthukumarappan K. K. Modelling approaches to ozone processing of liquid foods. *Trends in Food Science & Technology*. 2009. №20. P. 125—136.

6. Pereira M., Faroni L. R. D., Silva A. G., Sousa A. H., Paes J. L. Economical viability of ozone use as fumigant of stored corn grains. *Engenharia na Agricultura*. 2008. №16(2). P. 144—154.

7. Ferral-Pérez H., Torres Bustillos H., Méndez L. Sequential Treatment of Tequila Industry Vinasses by Biopolymer-based Coagulation/Flocculation and Catalytic Ozonation. *Ozone: Science & Engineering*. 2016. № 38. P. 279—290.

8. Karaca H. Use of Ozone in the Citrus Industry. *Ozone: Science & Engineering*. 2010. № 32. P. 122—129.

9. Shigezo N., Takahara H. Ozone Contribution in Food Industry in Japan. *Ozone: Science and Engineering*. 2006. № 28. P. 425—429.

10. Захаров В. В., Устінов О. А., Змієвський Ю. Г., Мирончук В. Г. Застосування алгоритму наївного баєсового класифікатора. *Наукові праці Національного університету харчових технологій*. К. 2018. Том 24. № 5. Ч. 1. P. 91—99.

11. Мова програмування Phyton. URL: https://www.python.org/downloads/release/python-2715/

12. Langseth H., Nielsen T. D. Classification using Hierarchical Naive Bayes models. *Mach Learn*. 2006. № 63. P. 135—159.