УДК 004.89

Бурківський О. С., *здобувач 2 курсу спеціальності 122 Комп'ютерні науки,*
науковий керівник:
Фриз І. В., *канд. фіз.-мат. наук, старший викладач кафедри інформаційних технологій*

ЙМОВІРНІСНІ МЕТОДИ В МАШИННОМУ НАВЧАННІ

Донецький національний університет імені Василя Стуса, м. Вінниця

Вступ. Одним із ключових аспектів машинного навчання є застосування ймовірнісних методів, які дають змогу ефективно моделювати та управляти невизначеністю в даних.

Теорія ймовірностей, в основі якої лежить математична обробка ймовірностей подій та випадкових величин, відіграє одну з ключових ролей у розвитку та застосуванні різноманітних методів машинного навчання. Вона допомагає створювати моделі, які не лише адаптуються до наявних даних, але й враховують ступінь невизначеності та ймовірнісні закономірності.

Розглянемо один з аспектів застосування теорії ймовірності у машинному навчанні, а саме, як ймовірнісні методи використовуються для кластеризації і допомагають вирішувати складні завдання, пов'язані з аналізом даних та прийняттям рішень.

Виклад основного матеріалу. Кластеризація, або кластерний аналіз – це статистична процедура, завдання якої полягає в розбитті вибірки об'єктів на підмножини (кластери), які не перетинаються, так, що об'єкти всередині одного кластера мають високу подібність між собою, а об'єкти різних кластерів відрізняються. Задача кластеризації є однією з найпоширеніших задач без нагляду в машинному навчанні та аналізі даних. Кластеризація використовується для виявлення структури в наборах даних, групуванні подібних об'єктів і для виявлення схожих підгруп серед даних.

Як описано у [1], алгоритми кластеризації розподіляються на два типи ієрархічних методів: агломеративні та дивизимні. Перші ґрунтуються на тому, що всі об'єкти спочатку розташовуються у власні класи, а потім, використанням метрики подібності, ці об'єкти поступово об'єднуються, зменшуючи кількість кластерів до моменту отримання одного. Другі – навпаки, починають з віднесення об'єктів до одного кластера, а зі збільшенням відстані розподіляють їх за окремими кластерами.

Розглянемо типовий алгоритм кластерного аналізу, наведений у [1], на основі аналізу робіт [2, 3, 4], який застосовується у різних галузях, зокрема і в бізнесі. Нехай $A = \{a_1, a_2, \dots, a_n\}$ – множина об'єктів, а B – множина номерів кластерів. Обирається метрика $d_{ij}(x_{ik}, x_{jk})$, водночас є формулою відстані. Мета полягає у розбитті множини A на підмножини (кластери), які не перетинаються, і кожен кластер містить об'єкти, що близькі за метрикою $d_{ij}(x_{ik}, x_{jk})$, водночас об'єкти

різних класів значно відрізняються. Кожному об'єкту a_i призначається номер кластера B_i.

Множина A може містити об'єкти з різними одиницями вимірювання або різним діапазоном значень, тому необхідно здійснити нормалізацію вхідних даних. Для цього можна скористатися MinMax-нормалізацією або Z-нормалізацією.

Перетворення вхідних даних за допомогою MinMax-нормалізації виконується так:

$$x' = \frac{x - \min[X]}{\max[X] - \min[X]} \quad \text{або} \quad x' = a + \frac{x - \min[X]}{\max[X] - \min[X]}(b - a)$$

для даних, які знаходяться в діапазоні [0, 1] або ж для довільного діапазону $[a, b]$ відповідно.

У процесі Z-нормалізації кожен вхідний параметр перетворюється таким чином, щоб його середнє значення було рівним 0, а стандартне відхилення – 1. Перетворення вхідних даних за допомогою Z-нормалізації здійснюється за такою формулою

$$x' = \frac{x - M[X]}{\sigma[X]},$$

де $M[X]$ – математичне сподівання, $\sigma[X]$ – середнє квадратичне відхилення.

У кластерному аналізі можуть використовуватися різні міри подібності (зазвичай описуються невід'ємною функцією), як-от коефіцієнт кореляції, міри відстані, зокрема евклідова відстань між двома точками на площині: $d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$, коефіцієнти асоціативності та ймовірнісні коефіцієнти подібності.

У кластеризації широко використовується функція Гауса, особливо для методів, що базуються на моделі гаусової суміші (GMM). Ці методи моделюють дані як суміш декількох розподілів Гауса, що дає змогу ефективно виявляти кластери з різними формами та розмірами. Кожен кластер описується власним розподілом Гауса, параметри якого потрібно визначити, і з урахуванням цього розподілу знайти ймовірність, і з якою кожна точка належить до певного кластера.

Нагадаємо, що для d-вимірного випадку функція щільності розподілу Гауса має такий вигляд [5]:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

де x – d-вимірний вектор спостереження, μ – вектор середніх значень, Σ – коваріаційна матриця, $|\Sigma|$ – детермінант коваріаційної матриці, Σ^{-1} – обернена коваріаційна матриця.

На заключному кроці обирається та впроваджується відповідний алгоритм кластерного аналізу, а результати перевіряються на достовірність.

Висновки. Застосування ймовірнісних методів у машинному навчанні дає змогу ефективно моделювати та управляти невизначеністю в даних, що є важливим для створення адаптивних та точних моделей. Методи кластеризації на основі моделей, як-от GMM, дають змогу виявляти кластери з різними формами та розмірами, враховуючи ступінь невизначеності. Для покращення якості класте-

ризації важливо здійснювати нормалізацію даних за допомогою MinMax або Z-нормалізації, що забезпечує порівнянність значень об'єктів. Використання функції густини ймовірності Гауса та алгоритму GMM сприяє точному визначенню параметрів кластерів та їх ймовірностей, підвищуючи гнучкість і точність аналізу, що розширює можливості прийняття рішень у різних галузях.

Список використаних джерел

1. Шевченко С. М., Жданова Ю. Д., Шевцова Т. І. Застосування кластерного аналізу для просування бізнесу у соціальних мережах. *Вісник ХНТУ*. 2023. № 4(87). С. 271–281. DOI: 10.35546/kntu2078-4481.2023.4.32 (дата звернення 08.05.2024).
2. Koirala J. Understanding the Use of Cluster Analysis in Business (March 27, 2023). DOI: 10.2139/ssrn.4400674 (дата звернення 08.05.2024).
3. Безпарточний М. Г. Використання кластерного аналізу при оцінці ефективності діяльності торговельних підприємств. *Торгівля, комерція, підприємництво:* збірник наукових праць. Львів: Львівська комерційна академія. 2014. Вип. 17. С. 24–27.
4. Модель експертної системи для медичного скринінгу на основі методів кластерного аналізу / С. М. Шевченко, Ю. Д. Жданова, О. В. Негоденко, В. А. Куцук. *Moderní aspekty vědy:* XXVII: díl mezinárodní kolektivní monografie. Mezinárodní Ekonomický Institut s. r. o. Česká republika. Mezinárodní Ekonomický Institut s. r. o., 2023. С. 478–494. URL: http://perspectives.pp.ua/public/site/mono/mono-27.pdf (дата звернення 08.05.2024).
5. Reynolds D. Gaussian Mixture Models / S. Z. Li, A. Jain (eds). *Encyclopedia of Biometrics*. Springer, Boston, MA, 2009. DOI: 10.1007/978-0-387-73003-5_196 (дата звернення 08.05.2024).

УДК 004.8:519.2:658.6

Яценко В. В., *здобувач 2 курсу спеціальності 122 Комп'ютерні науки,*
Комаров В. Ф., *канд. техн. наук, старший викладач кафедри інформаційних технологій*

ПОРІВНЯННЯ АЛГОРИТМІВ КЛАСИФІКАЦІЇ НА ПРИКЛАДІ ЗАДАЧІ ВИЗНАЧЕННЯ КАТЕГОРІЇ ТОВАРУ МАГАЗИНУ

Донецький національний університет імені Василя Стуса, м. Вінниця

Вступ. Швидкий аналіз і обробка вхідних даних – основний виклик сучасних автоматизованих систем. Від швидкості прийняття рішень залежить ефективність вирішення багатьох прикладних задач кібернетики у різних галузях застосування. Обсяги даних, які необхідно аналізувати, зростають експоненційно, здатність оперативно обробляти великі потоки даних та виявляти в них критично важливі закономірності стає конкурентною перевагою під час розробки нових програмних продуктів, а в окремих задачах є вирішальною. Затримка з прийняттям рішень на основі отримуваної інформації може мати вкрай негативні наслідки – від збитків у бізнесі до небезпечних помилок у критично більш важливих системах. В умовах активного екстенсивного розвитку й проникнення смарт-технологій у всі сфери життєдіяльності людини ручне програмування рішень під кожен набір даних не є швидким і оптимальним вибором.