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.

Одним зі способів вирішення проблеми є методи машинного навчання [1]. Машинне навчання – це галузь штучного інтелекту, що спеціалізується на розробці і дослідженні моделей та алгоритмів, що здатні навчатися з даних, аналізувати їх і на основі аналізу приймати рішення без явних інструкцій.

Метою дослідження є аналіз та порівняння різних методів класифікації для визначення доречного для поставленої задачі. У межах роботи необхідно пояснити, як відбувається машинне навчання, представити кожен алгоритм, визначити переваги і недоліки та підсумувати відповідну інформацію.

Виклад основного матеріалу. Машинне навчання моделі складається з декількох етапів, а саме: збір та підготовка даних, вибір моделі машинного навчання, безпосередньо навчання моделі, оцінка та налаштування моделі.

На першому етапі відбувається початковий збір інформації. Відбір інформації може відбуватися з вебсайтів, сенсорів обладнання, баз даних, даних з платформ соціальних мереж. Так, наприклад, для задачі визначення категорії товару в магазині необхідно відібрати найменування і ознаки товарів. Далі дані піддаються обробці. Відбувається очищення початкових даних, видалення пропусків або їх заповнення методами інтерполяції чи спеціальними значеннями. До того ж дані перевіряються на предмет помилок. Вибірка перевіряється на наявність аномальних значень – даних, що виходять за задані діапазони, або нереалістичної інформації. Обов'язково видаляються дублікати, оскільки наявність повторюваних даних може призвести до погіршення здатності розпізнавати невідомі дані і навіть до перенавчання моделі. Згодом початкові дані нормалізуються і стандартизуються [2]. Ідея нормалізації і стандартизації полягає в тому, щоб подати дані для моделі в однаковому масштабі в однакових мірах вимірювання, тобто так, щоб модель правильно порівнювала дані з певними ознаками і не надавала перевагу ознаці через неправильний акцент у даних.

На другому етапі відбувається вибір моделі машинного навчання, яка оптимально підходить для вирішення задачі з урахуванням складності, характеристик даних і витрат ресурсів. До основних типів моделей машинного навчання належать: регресійні моделі – моделі, які займаються прогнозуванням певної інформації; класифікаційні – такі, що відносять певне явище до класу згідно з його ознаками; кластеризаційні – моделі, що групують подібні явища. У випадку категоризації товарів класифікаційна модель є оптимальною для розв'язування задачі.

Основні методи класифікації інформації – це логістична регресія, дерево рішень, випадковий ліс і метод опорних векторів. У роботах [3, 4] подаються нові методи прискореного і більш ефективного машинного навчання моделей. За допомогою паралелізованої обробки інформації і методів ансамблювання, тобто поєднання прогнозувань декількох моделей, можна значно пришвидшити процес машинного навчання, зробити його більш ефективним, знизити потреби у ресурсах і спростити подальше масштабування моделі.

Логістична регресія застосовує модель, яка описує зв'язок між незалежними змінними і залежними змінними. Зазвичай це лінійне рівняння, де кожна змінна має свій коефіцієнт. Після формування моделі використовується логістична функція, яка перетворює вихідні значення рівняння у ймовірність, у діапазоні значень від 0 до 1 (подія не відбудеться і подія відбудеться). Далі класифікація

здійснюється за допомогою обчислення ймовірності для точки даних. Логістична регресія дуже проста у використанні і доволі ефективна. Недоліки методу – регресія не підходить для задач з більше ніж двома категоріями і припускає, що зв'язок між незалежними змінними і залежними лінійний, що може призвести до неточних прогнозів. Цей метод не є ефективним у використанні через відсутність можливості роботи з множинною класифікацією [5].

Дерево рішень працює так: починаючи з кореневого вузла вибудовується дерево, за допомогою якого згодом будуть прийматись рішення. У кожному вузлі вибирається ознака, яка найкраще описує дані за змінною, яку користувач хоче прогнозувати. Ця ознака використовується для розмежування даних на групи, представлені дочірніми вузлами. Процес повторюється, поки не будуть досягнуті кінцеві вузли. Під час аналізу змінної відбувається проходження дерева від кореневого вузла до листа. Дерево рішень – доволі простий у використанні метод, але водночас дуже ефективний, особливо для великих наборів даних. До того ж дерево рішень можна візуалізувати, що полегшує аналіз рішень моделі. Однак цей метод потребує особливо ретельної обробки даних, оскільки дерево чутливе до якості даних. Дерево рішень схильне до перенавчання, що ускладнює навчання з різними випадками класифікації.

Випадковий ліс – ансамбль дерев рішень. Цей метод є доволі оптимальним для вирішення даної задачі. Фактично алгоритм лісу аналогічний побудові дерева, з відмінністю у тому, що вибудовується множина дерев рішень, і під час прийняття рішень відбувається голосування кожного дерева. Прогнози дерев узагальнюються, і на основі остаточного підрахунку голосів відбувається прийняття рішення. Випадковий ліс поєднує в собі високу точність і надійність. Цей алгоритм дуже стійкий до шуму. Основним недоліком є великі затрати в ресурсах для побудови цієї моделі і низька швидкість прийняття рішень. До того ж рішення таких моделей важко інтерпретувати через поєднання прогнозів багатьох дерев. Випадковий ліс буде доречним у задачі з категоріями товарів.

Метод опорних векторів шукає гіперплощину, тобто багатовимірну пряму, яка розділяє простір між двома точками даних різних класів, причому відстань від найближчої точки площини має бути максимальна. Точки даних, які розташовані до гіперплощини найближче, називаються опорними векторами. На основі того, в яку площину потрапить нова точка, буде визначено, до якого класу належить ця точка. Алгоритм дуже точний і стійкий до шуму, доволі легко інтерпретується, але побудова такої моделі може бути ресурсозатратним процесом, а процес прийняття рішень повільним.

Далі відбувається ***навчання моделі і оцінка її продуктивності.*** Залежно від результатів тестування можуть бути модифіковані параметри моделі. Якщо результати будуть невдалі, навчання може початися спочатку.

Висновки. У підсумку машинне навчання значно спрощує процес обробки і прийняття рішень. Кожен з алгоритмів має свої переваги та недоліки і може бути корисним у певній ситуації. Логістична регресія – дуже швидкий метод, але має незадовільну точність у комплексних задачах і неактуальний у задачах із множинною класифікацією. Дерево рішень надає оптимальний за часом і точністю результат, але цей метод нестійкий до шуму і його успішність сильно залежить

від початкових даних. Випадковий ліс стійкий до шуму, але ресурсозатратний у генерації і роботі. Метод опорних векторів дає максимально точний результат, але порівняно повільний у роботі. Вибір алгоритму повинен відповідати потребам – обсягу вхідних даних та вимогам користувача.

Список використаних джерел

1. Machine learning explained. 2024. URL: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (дата звернення: 16.05.2024).

2. Optimization vs standartization. 2024. URL: https://www.geeksforgeeks.org/normalization-vs-standardization/ (дата звернення: 16.05.2024).

3. Harmon M., Klabjan D. Activation Ensembles for Deep Neural Networks. arXiv, 2017. DOI: 10.48550/arXiv.1702.07790 (дата звернення: 16.05.2024).

4. Luo H., Haffner P., Paiement J.-F. Accelerated Parallel Optimization Methods for Large Scale Machine Learning. arXiv, 2014. DOI: 10.48550/arXiv.1411.6725 (дата звернення: 16.05.2024).

5. What is logistic regression. 2024. URL: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/ (дата звернення: 16.05.2024).

УДК 519.2

Мороз Д. В., *здобувачка 1 курсу спеціальності 111 Математика,*
Луценко А. В., *д-р філос. з математики, старший викладач кафедри прикладної математики та кібербезпеки*

ЗАСТОСУВАННЯ ЛІНІЙНОЇ АЛГЕБРИ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ДАНИХ

Донецький національний університет імені Василя Стуса, м. Вінниця

Актуальність. У світі, що переповнений даними, важливість їх аналізу та інтерпретації зростає з кожним днем. Інтелектуальний аналіз даних стає ключовим інструментом для виявлення закономірностей, отримання цінних інсайтів та прийняття стратегічних рішень у різних сферах, від бізнесу до медицини. Одним із фундаментальних математичних інструментів, який забезпечує основу для інтелектуального аналізу даних, є лінійна алгебра. Розуміння та ефективне використання принципів лінійної алгебри в цьому контексті стає важливим завданням для дослідників, аналітиків та фахівців з обробки даних.

Метою роботи є розгляд застосування лінійної алгебри в інтелектуальному аналізі даних. У роботі будуть розглянуті основні поняття лінійної алгебри, як-от вектори, матриці, лінійні простори та операції над ними, а також конкретні сфери застосування цих принципів у сучасному інтелектуальному аналізі даних, як-от обробка та аналіз даних, методи машинного навчання, обробка зображень та сигналів, а також рекомендаційні системи. Результатом роботи буде розуміння важливості лінійної алгебри для ефективного використання інтелектуальних методів аналізу даних та виявлення перспектив подальшого використання цих знань у практиці.