ЧАСТИНА 2

МЕТОДИ КОМП'ЮТЕРНОГО ЗОРУ І ГЛИБИННИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ СУПУТНИКОВОГО ІНТЕЛЕКТУ

1. МЕТОД ВИЯВЛЕННЯ АНОМАЛІЙ В ДАНИХ НАВЧАННЯ МОДЕЛЕЙ КОМП'ЮТЕРНОГО ЗОРУ НА ОСНОВІ МЕТОДІВ КЛАСТЕРИЗАЦІЇ

Антон Охріменко, аспірант
Кафедра математичного моделювання і аналізу даних
Навчально-науковий Фізико-технічний інститут
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

antoh-ipt21@lll.kpi.ua

ВСТУП

Якість та продуктивність моделей машинного навчання значною мірою залежать від розміру вибірки та якості навчальних даних. У загальному випадку, чим більше даних, тим кращою буде модель. Але якість навчальних даних також важлива, оскільки не можна просто дублювати дані чи постійно відбирати їх із того самого джерела. Дані мають бути різноманітними, щоб охопити якомога більший об'єм у просторі ознак. З огляду на задачу класифікації набір даних має бути роздільним, наприклад, повинні існувати граничні поверхні, які чітко відокремлюють точки даних, які належать до різних класів. В ідеальному випадку такі поверхні мають бути достатньо гладкими для уникнення перенавчання моделі. В цьому випадку дані утворюють певні кластери, кожен з яких містить екземпляри даних лише одного класу.

Проблема перекриття класів унеможливлює побудову чітких та однозначних граничних поверхонь [1]. Як наслідок, дані не можуть бути розділені на окремі кластери, а значить, частину даних не можна чітко розрізнити між собою у просторі ознак. Крім того, в просторі ознак існують деякі підпростори, які містять суміш точок різних класів без будь-якої структури.

При наявності вказаної проблеми важливими є наступні питання. Чи підходить даний набір даних для розв'язання даної задачі класифікації? Чи потрібно удосконалити процес збору даних? Як досягти найкращого можливого результату, використовуючи даний набір даних, якщо проблему з накладанням класів не вдалося виправити?

Для глибшого розуміння проблеми розглянемо причини виникнення проблеми «перекриття» класів в просторі ознак. Ця ситуація може бути спричинена похибками в процесі збору даних та/або їх розмітки, або недостатньою інформативністю ознак. В останньому випадку додавання нових ознак, що еквівалентно додаванню додаткових вимірів до простору ознак, може значно покращити придатність даних для розв'язання поставленої задачі. В той же час, збільшення розмірності простору ознак, в свою чергу, може призводити до перенавчання моделі. Як наслідок, дослідникам потрібен алгоритм для визначення недоліків датасетів, тобто виявлення частки неоднозначних даних і підпросторів із такими даними у навчальному та тестовому наборі даних. Результати роботи алгоритму можна використати для прийняти рішення щодо модифікації процесу збору даних, додавання нових ознак або збільшення кількості і точності екземплярів даних у «сумнівних» підпросторах і навколо них.

Найбільш простим методом дослідження датасету є його візуалізація у двовимірному просторі. Датасети з великим числом ознак неможливо візуалізувати без додаткових перетворень. У цьому випадку для відображення даних з простору з великою розмірністю у простір з меншою розмірністю можна використовувати такі алгоритми, як PCA [2] та tSNE [3], та нанести отриманий результат на двовимірний графік. До недоліків цих методів відноситься залежність від суб'єктивності дослідника та нездатність ефективної візуалізації та аналізу даних великої розмірності через великі втрати інформації в процесі перетворення.

Однак у більшості випадків дослідник не може вплинути на процес збору даних і змушений працювати з даними, в яких наявна проблема перекриття класів. У цьому випадку йому також

потрібен алгоритм, який дозволить оцінити потенційну точність класифікації на заданому наборі даних і визначити ненадійні екземпляри даних для корекції процесу навчання моделі.

Визначені на попередніх кроках «сумнівні» підпростори в просторі ознак можуть бути використані для корекції результатів (передбачень) моделі. Для прикладу, для екземплярів даних, що потрапляють у ненадійну частину простору ознак, можна використовувати інші правила та навіть інші моделі.

Для задачі класифікації зображень визначення неоднозначних екземплярів даних стає особливо важливим. Як правило, згорткова нейронна мережа (CNN) [4] простір зображень відображає у простір ознак. Таким чином вхідне зображення перетворюється на вектор ознак. Остаточна класифікація виконується саме на базі цього одновимірного вектору. Виявлення неоднозначних екземплярів даних у просторі ознак є важливою науковою задачею, розв'язання якої дозволяє покращити якість розпізнавання зображень з використанням CNN. Варто зауважити, що вищевказане перетворення у простір ознак не детерміноване, а роздільність класів залежить не лише від якості даних, а також від якості роботи згорткових шарів нейронної мережі.

На тему проблеми перекриття класів у датасетах проведено численні дослідження [5-13]. У даному розділі запропоновано новий метод виявлення неоднозначних екземплярів даних на основі метода K найближчих сусідів (KNN – K Nearest Neighbors) [14], продемонстровано його роботу на супутникових даних (багатоспектральних оптичних знімках) в задачі класифікації сільськогосподарських культур та розглянуто можливі варіанти використання даного алгоритму під час розробки моделей машинного навчання [15].

1.1. ПОСТАНОВКА ЗАДАЧІ

Дві можливі проблеми, які можуть виникнути під час розв'язання задачі класифікації, — це перекриття класів і викиди (рис. 1). Проблема викидів полягає у наявності серед множини екземплярів класу поодиноких представників з нехарактерними ознаками. В просторі ознак такі поодинокі екземпляри потрапляють у підпростір, де розташовані «типові» представники іншого класу (рис. 1-а). Вона характеризується наявністю в просторі ознак областей, де рівномірно представлені декілька класів. З цими проблемами часто стикаються фахівці при розробці

моделей машинного навчання на геопросторових і, зокрема, супутникових даних. В даному розділі запропоновано алгоритм, який дозволяє розв'язати обидві задачі одночасно, виявляючи неоднозначні та ненадійні дані в датасеті. Неоднозначними або ненадійними будемо вважати складні для розпізнавання екземпляри даних, які будь-яка класифікаційна модель з високою ймовірністю не зможе правильно розпізнати. Пошук таких ненадійних екземплярів даних у датасеті може стати важливим допоміжним інструментом для дослідника протягом усього циклу розробки моделі машинного навчання.

В якості набору даних, який досліджувався, розглянуто навчальну вибірку для задачі класифікації земного покрову, використану для побудови нейромережевої задачі класифікації в [16]. До її складу входять дані оптичного супутника Sentinel-2 з просторовим розрізненням 10 м для Київської області. Для уникнення проблеми хмарності, в даному дослідженні використовується композит, отриманий з супутникових знімків Київської області Sentinel-2 впродовж липня 2021 року. Даний композит має 4 канали: синій (490 нм), зелений (560 нм), червоний (665 нм) та інфрачервоний (842 нм).

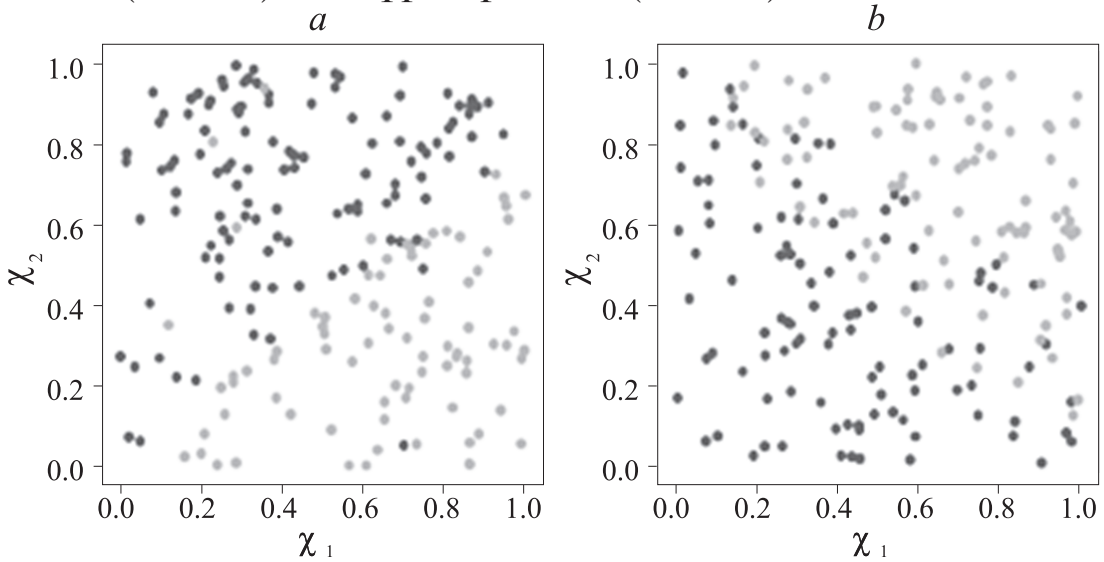
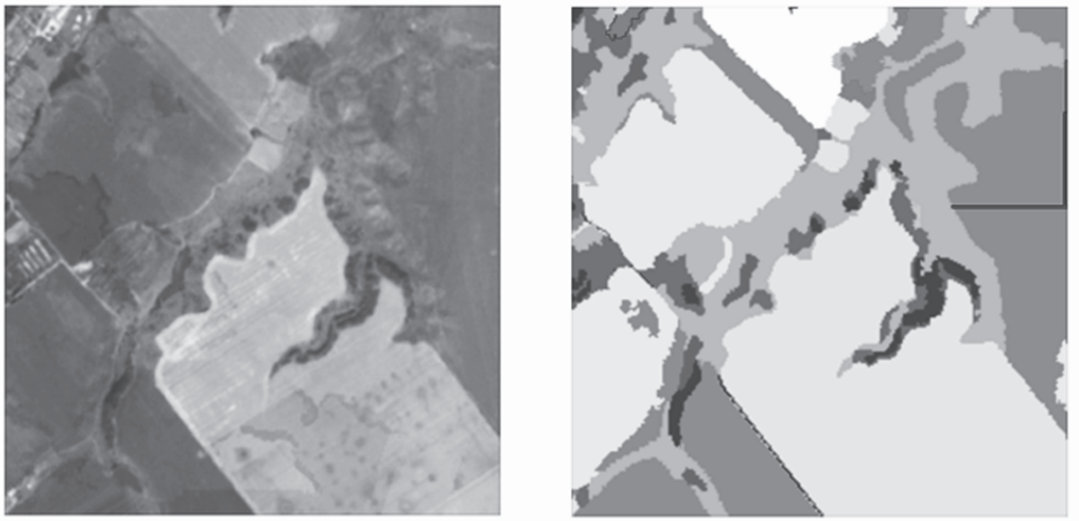


Рис. 1. Проблема викидів (a) та проблема перекриття класів (b).

Фрагмент 3-канального зображення композиту для Київської області в палітрі true color, яка включає канали 490, 560 та 665 нм, показано на рис. 2-a. Цей композит використовувався для побудови карти класифікації земного покриву з використанням згорткової нейромережевої моделі, розробленої Інститутом космічних досліджень НАНУ-ДКАУ [17], [18]. Приклад фрагменту карти класифікації показаний на рис. 2-b.



a *b*

Рис. 2. Фрагмент 3-канального супутникового зображення (a) та відповідної карти класифікації земного покриву (b)

У даному розділі розглядається задача попіксельної класифікації [19, 20], що зводить дану проблему до задачі класифікації на декілька класів за чотирма ознаками. Для формування навчального датасету виберемо 8 класів сільськогосподарських культур, при цьому інші культури віднесено до окремого класу "Інші культури". Класи, які не відносяться до сільськогосподарських культур, були відкинуті, наприклад забудівлі чи водні об'єкти.

Для зменшення об'єму досліджуваного набору даних з обраних 8 класів випадковим чином виберемо 25 000 відповідних йому пікселів на супутниковому композиті. Якщо певний клас включає меншу кількість пікселів, для проведення експерименту будуть використані всі наявні пікселі.

На рис. 3 наведено візуальне представлення отриманого датасету. Для візуалізації чотиривимірних даних на площині розмірність даних була понижена до двовимірної за допомогою алгоритмів PCA та tSNE.

Отриманий датасет містить 9 класів та має проблему перекриття класів. Для візуалізації ми використовуємо редуковані двовимірні дані. Слід зазначити, що всі обчислення запропонованого алгоритму виконано на повних чотиривимірних даних, а сам алгоритм здатний працювати з даними будь-якої розмірності. Запропонований метод визначення неоднозначних даних описано в наступному підрозділі.

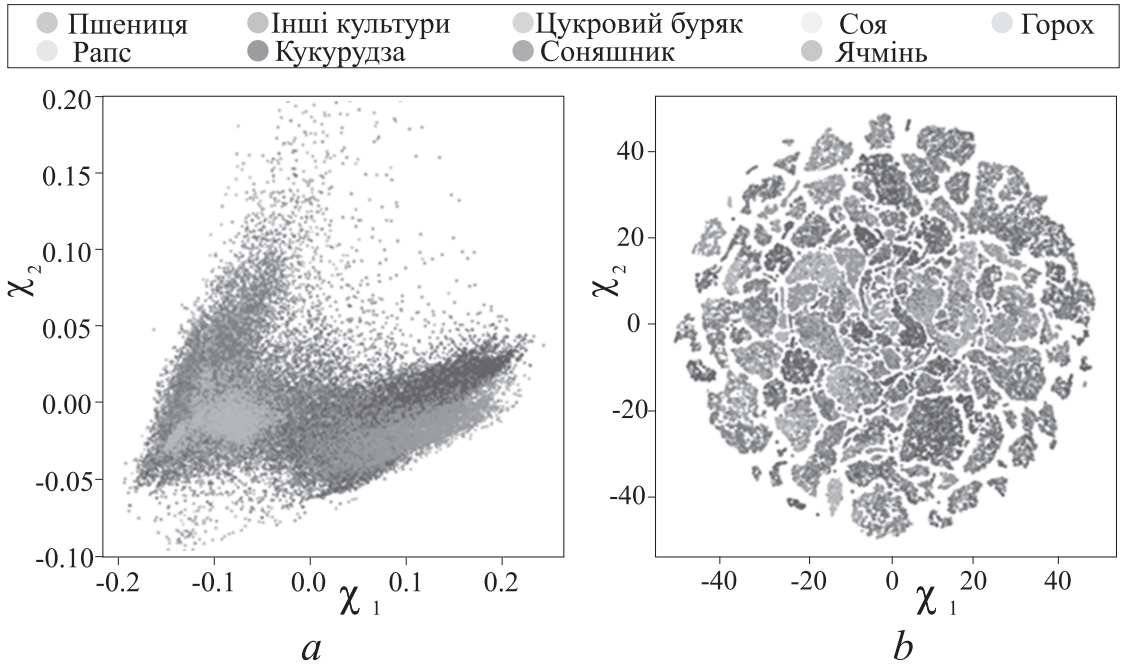


Рис. 3. Візуалізація датасету за допомогою PCA (a) та tSNE (b)

1.2. РОЗВ'ЯЗАННЯ ЗАДАЧІ

Нехай $\hat{X}$ — множина екземплярів даних, а $\vec{x}_i \in \hat{X}$ — i-й екземпляр даних з цієї множини. Аналогічно $\hat{Y}$ — множина міток класів, до яких можуть належати екземпляри даних $\vec{x}_i$, і $y_i = \hat{Y}$ — істинний клас для екземпляра даних $\vec{x}_i$.

Потрібно відповісти на питання: чи можливо деякий екземпляр даних $\vec{x}_i$ правильно класифікувати як клас y_i. Для цього буде використано ансамбль класифікаторів KNN з різними номерами сусідів $n = [0,1,\dots N], n \in \mathbb{N}$. Для кожного $\vec{x}_i \in \hat{X}$ буде отримано вектор $\vec{m}_i$, де елемент $\vec{m}_i$ — це результат класифікації $\vec{x}_i$ за допомогою класифікатора KNN з параметром числа сусідів рівним j, який було навчено за допомогою набору даних $\hat{X} \setminus \vec{x}_i$.

$$\vec{m}_i : m_i^j = KNN(\vec{x}_i, j, \hat{X} \setminus \vec{x}_i). \quad (1)$$

Таким чином, кожен екземпляр даних $\vec{x}_i$ матиме відповідний йому вектор $\vec{m}_i$ та з цих векторів можна побудувати матрицю M:

$$M : M_i^j = m_i^j. \quad (2)$$

Тепер кожний вектор $\vec{m}_i$ можемо порівняти з істинним класом y_i. Розглянемо декілька можливих випадків: більшість елементів $\vec{m}_i$ відповідають справжньому класу y_i; перші елементи $\vec{m}_i$ відповідають справжньому класу y_i, решта — ні; більшість елементів $\vec{m}_i$ не відповідають справжньому класу y_i; передбачений клас m_i^j постійно змінюється в залежності від j, відбуваються "стрибки між класами".

Будемо вважати, що для того щоб вважати екземпляр даних надійним та однозначним, перші дві умови є обов'язковими. Але перша умова завжди істинна, коли і друга істинна, тому залишається лише одна умова.

Екземпляр даних не може бути надійним, якщо справджується третя або четверта умова. Третя умова означає, що цей екземпляр з високою долею ймовірності є викидом, а четверта — що екземпляр даних у просторі ознак оточений іншими екземплярами з іншими мітками класу та, ймовірно, належить до зоні перекриття класів.

Таким чином, перша умова C_1: з перших k елементів вектора $\vec{m}_i$ хоча б r має дорівнювати істинному класу y_i. k, r — гіперпараметри, що представляють собою невеликі цілі числа. У загальному випадку найкращі значення цих параметрів залежать від щільності набору даних у просторі ознак.

Друга умова C_2: екземпляр даних є ненадійним, якщо y_i не є найчастішим класом серед перших k елементів вектора $\vec{m}_i$.

Остання умова C_3: якщо присутня часта зміна класів, екземпляр даних є ненадійним. З математичної точки зору це можна розглядати як одновимірну згортку вздовж вектора $\vec{m}_i$ з ядром K=[1, 1]. Якщо два сусідні елементи однакові, результат згортки буде нульовим. Для кожного випадку зміни класів результат згортки буде ненульовим. Результат згортки не повинен дорівнювати деякому цілому числу q, яке повинно бути достатньо невеликим.

Таким чином, остаточне правило для класифікації екземпляру даних можна представити наступним чином:

$$C_1 \wedge \overline{C}_2 \wedge \overline{C}_3, \tag{3}$$

де k, r, q — гіперпараметри. Змінюючи їх значення, ми можемо зробити одну умову важливішою за іншу. Як правило, $q < r, q < k$.

Алгоритм визначення ненадійних представників даних в датасеті мовою псевдокоду можна представити наступним чином.

```
FOR data_sample, label IN dataset:
        vector knn_results
        FOR i IN 1, 2, ..., n:
                knn_results[i] = // KNN classification of 'data_sample'
                  // with neibour number 'i'
        Condition_1 = //  BOOL: the most of first elements of 'knn_results'
             //        is equal to 'label'
        Condition_2 = // BOOL: 'label' is not most frequent class of
'knn_results'
        Condition_3 = // BOOL: convolution of 'knn_results' with [-1, 1]
gives
                        //       non-zero result more than the threshold
        IF (Condition_1 AND (NOT Condition_2) AND (NOT Condition_3)
):
                 // data_sample is reliable
        ELSE:
                //data_sample is not reliable
```

1.3. АНАЛІЗ РЕЗУЛЬТАТІВ

За результатами роботи алгоритму на супутниковому композиті визначено ненадійні точки в задачі класифікації за 4-ма ознаками. Надійні та ненадійні точки представлено на рис. 4. Також обраховано відсоток ненадійних точок по кожному класу та в цілому. Отримані результати наведено у табл. 1.

Як видно з табл. 1, більшість точок з датасету потрапила до ненадійних даних, що свідчить про неможливість точної попіксельної класифікації обраних 8 видів сільськогосподарських культур на основі одного знімку. Це узгоджується з емпіричними результатами сучасних досліджень в сфері дистанційного зондування Землі, які свідчать про те, що різні типи сільськогосподарських культур неможливо відрізнити по одному знімку. Тому в сучасних роботах для класифікації сільськогосподарських культур використовуються часові ряди супутникових даних, отриманих протягом всього вегетаційного

сезону [21], [22]. Для подолання цієї проблеми необхідно розширити датасет різночасовими знімками з додатковими каналами та/або використовувати просторові властивості знімків, наприклад за допомогою згорткових нейронних мереж.

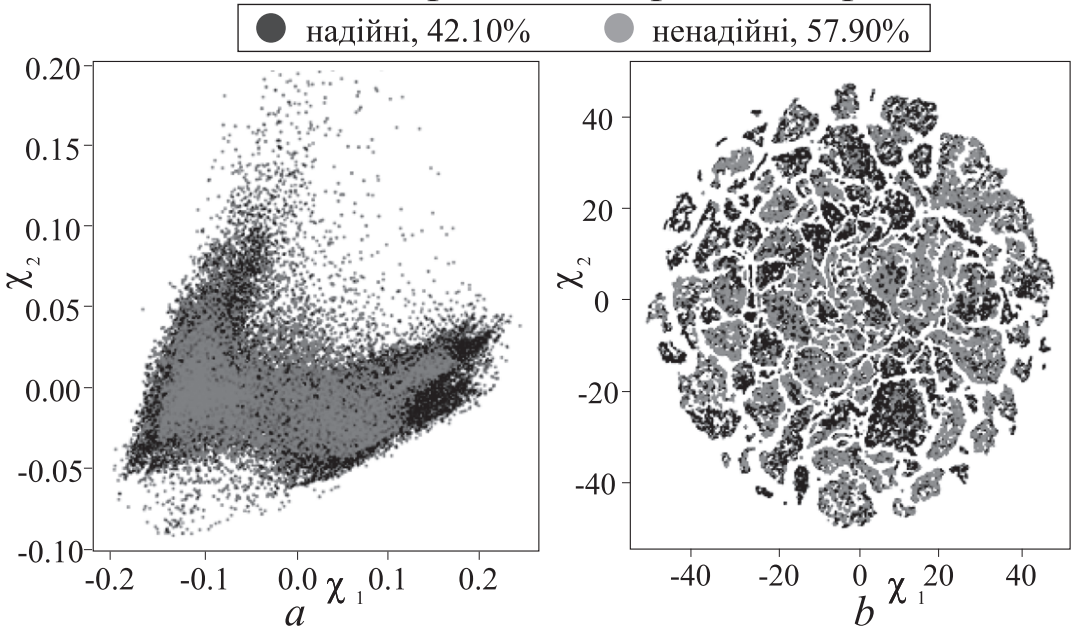


Рис. 4. Візуалізація визначених ненадійних точок датасету за допомогою PCA (a) та tSNE (b). Зеленим кольором позначено надійні точки червоним – ненадійні

Таблиця 1 Загальна кількість пікселів та частка ненадійних даних по кожному класу

Назва культури	**К-сть пікселів, тис.**	**Відсоток ненадійних точок, %**
Пшениця	25	51.36
Ріпак	25	63.17
Кукурудза	25	68.00
Цукровий буряк	25	78.66
Соняшник	25	61.99
Соя	25	54.88
Ячмінь	25	47.44
Горох	18	60.55

Інші культури	25	35.80
Разом	218	57.90

1.4. МОЖЛИВОСТІ ПРАКТИЧНОГО ЗАСТОСУВАННЯ

Алгоритм визначення ненадійних екземплярів даних можна вважати ще одним інструментом, який може бути використаний протягом усього циклу розробки моделі машинного навчання, від збору даних до розгортання моделі для використання у реальних умовах (рис. 5).

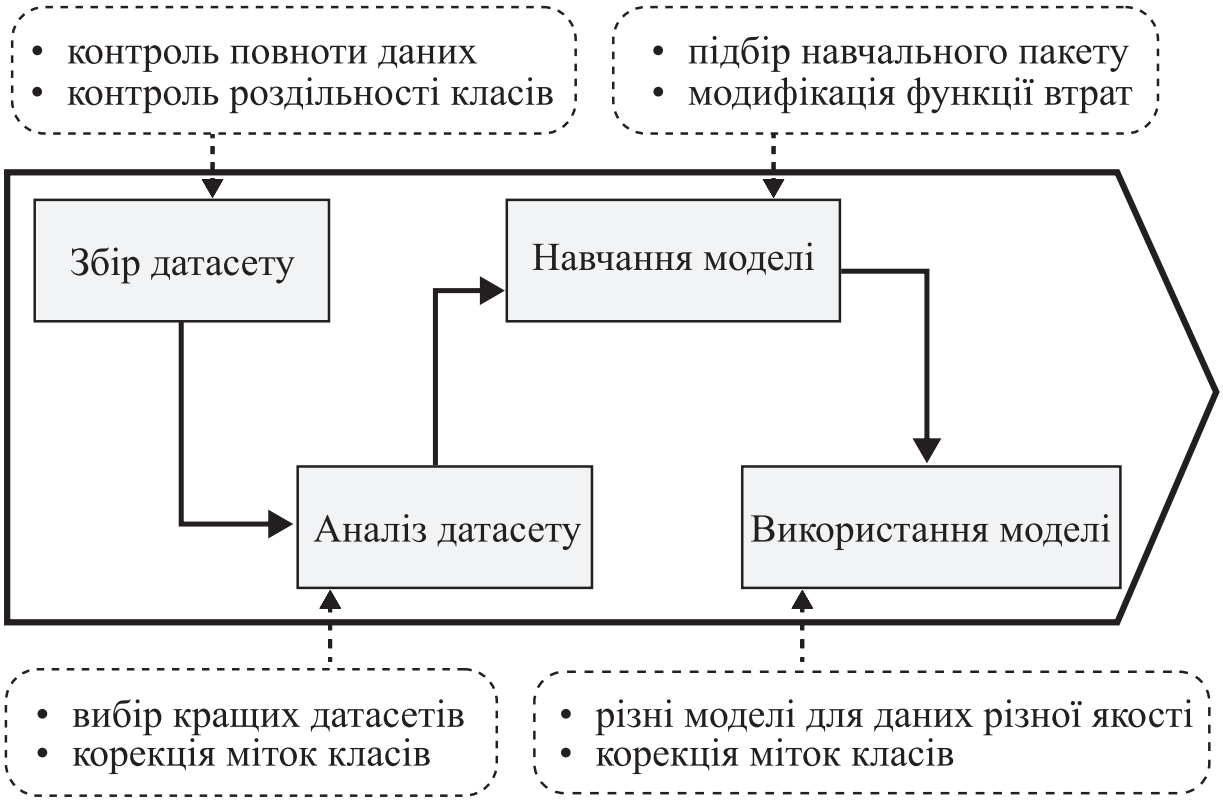


Рис. 1. Можливості використання алгоритму в повному циклі розробки моделі машинного навчання

Нижче наведено опис застосувань запропонованого методу.

Оцінка якості датасету. Представлений метод може бути корисним як в процесі збору даних, так і для аналізу наявних датасетів. Більшість сучасних задач машинного навчання вимагають великих датасетів, з великою розмірністю простору ознак та кількістю екземплярів даних. Досить часто наявний набір даних не в повній мірі відповідає поточній задачі. Тому в багатьох

випадках потрібно конструювати окремий датасет, наприклад, застосовувавши аутсорсинг або краудсорсинг [23], [24].

Подібна ситуація виникає, наприклад, в тому випадку, коли в якості навчальних даних в державних системах агромоніторингу використовується інформація про посіви, надана фермерами. Пілотний проєкт, проведений в попередні роки виконавцями проєкту з залученням даних від респондентів Державної служби статистики України, показав, що недостовірна інформація в таких даних може складати до 30%. У цьому випадку надзвичайно важливо мати інструмент для контролю якості отриманих даних, що дозволить виявити проблему ще на ранніх стадіях виконання робіт і скорегувати процес збору даних.

Можлива і протилежна ситуація, коли для розв'язуваної задачі є в наявності декілька датесетів. Зазвичай вибираються найкращі з них або їх комбінації. За допомогою розробленого методу можна визначити відсоток неоднозначних даних і, таким чином, вибрати дані найкращої якості або сформувати новий об'єднаний датасет з найменшою часткою ненадійних даних.

Корекція формування пакету даних для навчання. Розроблений метод дозволяє також покращити процес навчання моделі. Один з можливих методів полягає у зміні стратегії формування пакету даних для навчання. Під час навчання моделі можна використовувати надійні дані для навчання частіше, ніж ненадійні. Змінюючи частоту потрапляння ненадійних даних у навчальний пакет до повного видалення неоднозначних екземплярів даних, можна надати моделі нові властивості.

Отримана модель прийматиме рішення більшою мірою (суттєво) на основі надійних зразків даних, а не на ненадійних. Це дозволить зробити правильний прогноз у підпросторах, де немає перекриття класів, і приділяти менше уваги підпросторам, де класи перекриваються, оскільки у таких областях неможливо правильно визначити належність до класу.

Особлива стратегія ансамблювання. Можливість оцінювання якості окремих екземплярів даних дозволяє навчити кілька моделей, кожна з яких буде використовуватись лише у певних зонах простору ознак. Таким чином, правила класифікації для підпростору з надійними екземплярами даних будуть відрізнятися від ненадійних. Тоді остаточний ансамбль моделей буде складатися з двох підмножин: перша — з моделей, призначених

для надійних даних, а друга — для даних, які знаходяться біля неоднозначних екземплярів даних у просторі ознак. Подібний підхід до побудови динамічних ансамблів моделей у випадку незбалансованих вибірок запропоновано в [13].

Варто зауважити, що ненадійний екземпляр даних може або лежати в зонах, що перекриваються, або бути викидом. Зрозуміло, що в останньому випадку друга підмножина класифікаторів використовуватися не може, тому потрібно фільтрувати такі випадки.

Модифікація датасету. В багатьох методах роботи з незбалансованими дататетами та класами, які перекриваються, пропонується видалити з датасету екземпляри даних, які відносяться до найбільш представленого класу та знаходяться у зонах перекриття класів [25]. Натомість, можна внести зміни до їх істинного класу без зайвого видалення даних. В цьому випадку потрібно замінити мітки класів у зонах, де присутнє перекриття класів. Цільова мітка істинного класу, на яку необхідно поміняти мітки екземплярів даних у таких зонах, сильно залежить від поточних цілей. Для найкращих метрик має сенс змінити всі мітки класів на мітку найбільш представленого класу. Для боротьби з проблемою дисбалансу класів, усі мітки класів у сумнівній зоні можна змінити на найменш представлений клас.

Як і в попередньому підрозділі, при цьому потрібно відфільтрувати викиди, для яких мітки класів можна змінити на мітки екземплярів даних, які оточують їх у просторі ознак.

ОБГОВОРЕННЯ І ВИСНОВКИ

В розділі 1 наведено результати розробки методу для дослідження аномалій в даних і, зокрема, вирішення проблеми перекриття класів разом з іншими можливими випадками, такими як викиди та незбалансованість класів. Існує багато причин виникнення описаних проблеми, найважливішими з яких є недостатня точність процесу збору даних при формування датасету та недостатня кількість ознак і, відповідно, мала розмірність простору ознак. Незалежно від причин, деякі екземпляри даних просто неможливо класифікувати правильно, оскільки вони мають схожі ознаки з іншими екземплярами, які мають іншу мітку класу. Одним з прикладів таких задач є задача сегментації супутникових знімків для визначення типів землекористування. Через схожість спектральних характеристик таких культур, як пшениця та ячмінь,

їх неможливо розділити у просторі ознак. Такі екземпляри даних заважають процесу навчання моделі та можуть стати причиною неправильного прогнозування на реальних даних.

В даному розділі представлено новий метод, який не залежить від розмірності простору ознак та дозволяє сформувати більш чітке уявлення про якість датасетів. На відміну від методів візуальної оцінки, заснованих на декомпозиції, представлений метод дозволяє отримати детерміновані числові показники якості даних, такі як відсоток надійних даних. Крім того, він дозволяє чітко відрізнити достовірні екземпляри даних від недостовірних, відкриваючи можливість їх подальшої модифікації.

В якості апробації запропонованого підходу розглянуто практичну задачу попіксельної класифікації чотириканального супутникового композиту для визначення сільськогосподарських культур. Було визначено ненадійні точки, які важко класифікувати правильно, та обчислено відсоток ненадійних даних в цілому та окремо по кожному класу. Більшість точок з датасету, частка яких становить 59.7%, не можуть бути чітко відділені від точок іншого класу у просторі ознак. Особливо важкими для класифікації виявилися цукровий буряк та ріпак — частка ненадійних точок 78.66% і 63.17% відповідно. Це показує необхідність використання більшої кількості знімків та оптичних каналів для задач класифікації типів землекористування або використовувати математичну модель, яка б враховувала просторовий розподіл даних на знімку.

Запропонований метод може бути використаний протягом усього циклу розробки моделі машинного навчання, в тому числі етапів контролю та корекції процесу збору навчальних даних, а також вибору найкращих датасетів та їх компонування. Після виявлення неоднозначних екземплярів даних, можна застосувати різні правила для них і для надійних даних, як під час навчання, так і під час роботи на реальних даних. Метод також надає можливості по модифікації набору даних, зміні міток класів ненадійних зразків даних відповідно до поточних задач, таких як балансування класів тощо.

ПЕРЕЛІК ПОСИЛАНЬ

1. Chawla N.V. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook. 2005. pp. 853-867. DOI: 10.1007/0-387-25465-x_40.

2. Abdi H., Williams L.J. Principal component analysis. WIREs Computational Statistics. 2010. Vol. 2, no. 4. pp. 433-459. DOI: 10.1002/wics.101.

3. Van Der Maaten L., Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008. Vol. 9, no. 86. pp. 2579-2605.

4. O'Shea K., Nash R. An Introduction to Convolutional Neural Networks. arXiv preprint arXiv. 2015. DOI: 10.48550/ARXIV.1511.08458.

5. Mikołajczyk A., Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW). 2018. pp. 117-122. DOI: 10.1109/IIPHDW.2018.8388338.

6. Shorten C., Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019. Vol. 6, no. 1. DOI: 10.1186/s40537-019-0197-0.

7. Almutairi W.A., Janicki R. On relationships between imbalance and overlapping of datasets. EPiC Series in Computing. 2020. Vol. 69. pp. 141-150. DOI: 10.29007/h71z.

8. Kramer O. Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library. 2013. Vol. 51. pp. 13-23. DOI: 10.1007/978-3-642-38652-7.

9. Garcна V., Mollineda R.A., Sбnchez J.S. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications. 2008. Vol. 11, no. 3-4. pp. 269-280. DOI: 10.1007/s10044-007-0087-5.

10. Nwe M.M., Lynn K.T. KNN-Based Overlapping Samples Filter Approach for Classification of Imbalanced Data. Studies in Computational Intelligence. 2020. pp. 55-73. DOI: 10.1007/978-3-030-24344-9_4.

11. Chen L., Fang B., Shang Z., Tang Y. Tackling class overlap and imbalance problems in software defect prediction. Software Quality Journal. 2018. Vol. 26, no. 1. pp. 97-125. DOI: 10.1007/s11219-016-9342-6.

12. Tang Y., Gao J. Improved classification for problem involving overlapping patterns. IEICE Transactions on Information and Systems.

2007. Vol. E90-D, no. 11. pp. 1787-1795. DOI: 10.1093/ietisy/e90-d.11.1787.

13. Lдssig N., Oppold S., Herschel M. Metrics and Algorithms for Locally Fair and Accurate Classifications using Ensembles. Datenbank-Spektrum. 2022. Vol. 22, no. 1. pp. 23-43. DOI: 10.1007/s13222-021-00401-y.

14. Cover T.M., Hart P.E. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 1967. Vol. 13, no. 1. pp. 21-27. DOI: 10.1109/TIT.1967.1053964.

15. Okhrimenko A., Kussul N. KNN-Based Algorithm of Hard Case Detection in Datasets for Classification. In Proceedings of the 11th International Conference on Applied Innovations in IT. 2023. DOI: 10.25673/101926.

16. Makarichev V., Vasilyeva I., Lukin V., Vozel B., Shelestov A., Kussul N. Discrete Atomic Transform-Based Lossy Compression of Three-Channel Remote Sensing Images with Quality Control. Remote Sensing. 2022. Vol. 14, no. 1. pp. 125. DOI: 10.3390/rs14010125.

17. Lavreniuk M., Kussul N., Novikov A. Deep learning crop classification approach based on sparse coding of time series of satellite data. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. 2018. pp. 4812-4815. DOI: 10.1109/IGARSS.2018.8518263.

18. Kussul N., Shelestov A., Lavreniuk M., Butko I., Skakun S. Deep learning approach for large scale land cover mapping based on remote sensing data fusion. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2016. pp. 198-201. DOI: 10.1109/IGARSS.2016.7729043.

19. Okhrimenko A., Kussul N. KNN-Based Algorithm of Hard Case Detection in Datasets for Classification. Proceedings of International Conference on Applied Innovation in IT. 2023. Vol. 11, no. 1, pp. 113-118. DOI:10.25673/101926.

20. Охріменко А.О., Куссуль Н.М. Метод виявлення складних для розпізнавання зразків у наборах даних для задач класифікації у машинному навчанні. Проблеми керування та інформатики. 2023. Vol. 68, no. 4, с. 84–95. doi: 10.34229/1028-0979-2023-4-7.

21. Shelestov A., Lavreniuk M., Kussul N., Novikov A., Skakun S. Large scale crop classification using Google earth engine platform. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017. pp. 3696-3699. DOI: 10.1109/IGARSS.2017.8127801.

22. Garnot V.S.F., Landrieu L., Giordano S., Chehata N. Time-Space Tradeoff in Deep Learning Models for Crop Classification on Satellite Multi-Spectral Image Time Series. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. 2019. pp. 6247-6250. DOI: 10.1109/igarss.2019.8900517.

23. Zheng F., Tao R., Maier H.R., See L., Savic D., Zhang T., et al. Crowdsourcing Methods for Data Collection in Geophysics: State of the Art, Issues, and Future Directions. Reviews of Geophysics. 2018. Vol. 56, no. 4. pp. 698-740. DOI: 10.1029/2018RG000616.

24. Laso Bayas J., See L., Fritz S., Sturn T., Perger C., Dьrauer M., et al. Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. Remote Sensing. 2016. Vol. 8, no. 11. pp. 905. DOI: 10.3390/rs8110905.

25. Kaur H., Pannu H.S., Malhi A.K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys. 2019. Vol. 52, no. 4. pp. 1-36. DOI: 10.1145/3343440.