УДК 004.8, 004.93

А.О. Охріменко, Н.М. Куссуль

МЕТОД ВИЯВЛЕННЯ СКЛАДНИХ ДЛЯ РОЗПІЗНАВАННЯ ЗРАЗКІВ У НАБОРАХ ДАНИХ ДЛЯ ЗАДАЧ КЛАСИФІКАЦІЇ У МАШИННОМУ НАВЧАННІ*

Охріменко Антон Олександрович

Навчально-науковий Фізико-технічний інститут НТУУ «Київський політехнічний інститут імені Ігоря Сікорського»,
orcid: 0009-0004-8520-0278

antoh-ipt21@lll.kpi.ua

Куссуль Наталія Миколаївна

Навчально-науковий Фізико-технічний інститут НТУУ «Київський політехнічний інститут імені Ігоря Сікорського»,
orcid: 0000-0002-9704-9702

nataliia.kussul@gmail.com

Розглядається проблема неоднозначності в задачах класифікації в області машинного навчання. Задача класифікації полягає у навчанні моделі відрізняти екземпляри даних, що належать різним класам. Однак можливі ситуації, коли правильна класифікація певної множини екземплярів даних складна або навіть неможлива, незалежно від складності моделі машинного навчання. Запропоновано метод та алгоритм виявлення таких неоднозначних екземплярів даних, що базуються на використанні методу найближчого сусіда та аналізу класів екземплярів даних, розташованих поряд у просторі ознак, та дозволяють виділити підмножину неоднозначних екземплярів даних, що можуть негативно впливати на процес навчання моделі класифікації. З метою демонстрації практичного застосування алгоритму проведено експеримент на чотириканальному супутниковому композиті, що використовується для попіксельної класифікації сільськогосподарських культур. Визначено відсоток ненадійних даних загалом та окремо для кожної культури. Одним з основних результатів дослідження є можливість використання запропонованого алгоритму під час конструювання датасету (набору даних, dataset) для навчання моделі класифікації. Він допомагає виявити потенційно проблемні екземпляри даних та забезпечити якість вхідного набору даних. Крім того, розглянуто можливості застосування алгоритму після процесу навчання моделі при використанні в операційному режимі. Виявлення неоднозначних екземплярів може допомогти знайти потенційні помилки класифікації та покращити результати роботи моделі. Представлений алгоритм може стати важливим інструментом для дослідника впродовж повного циклу розробки моделі машинного навчання, починаючи від підготовки даних для навчання і закінчуючи її практичним впровадженням. Його застосування

* Робота виконана в межах проєкту Національного фонду досліджень України 2020.02/0292 «Методи і моделі глибинного навчання для прикладних задач супутникового моніторингу» (конкурс «Підтримка досліджень провідних та молодих учених»).

 ISSN 2786-6491

скорочуватиме час на отримання якісних навчальних даних, покращуватиме метрики класифікації та забезпечуватиме більш надійні результати у задачах машинного навчання.

Ключові слова: машинне навчання, класифікатор, метод найближчого сусіда, оцінка якості датасету (набору даних), незбалансовані набори даних, перекриття класів.

Вступ

Якість та продуктивність моделей машинного навчання значно залежать від розміру вибірки та якості навчальних даних. Чим більше даних, тим краща буде модель. Однак якість навчальних даних також важлива, не можна просто дублювати дані чи постійно відбирати їх з того самого джерела чи об'єкта. Вони мають бути різноманітними, щоб охопити якомога більший об'єм у просторі ознак. При розгляді задачі класифікації набір даних має бути роздільним, наприклад, повинні існувати граничні поверхні, які чітко відокремлюють точки даних, що належать різним класам. В ідеалі такі поверхні мають бути достатньо гладкими, щоб уникнути перенавчання моделі. В такому разі 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 логіки чи структури.

За наявності вказаної проблеми дослідник стикається з питаннями: чи підходить набір даних для даного завдання класифікації? чи треба удосконалити процес збору даних? як досягти найкращого можливого результату, використовуючи набір даних, якщо проблему з накладанням класів не вдалося виправити?

Для глибшого розуміння розглянемо причини виникнення проблеми «перекриття» класів у просторі ознак. Ця ситуація може бути спричинена похибками в процесі збору даних та/або їх розмітки або недостатньою інформативністю ознак (відсутністю складності даних). В останньому разі додавання додаткових ознак, що еквівалентно додаванню додаткових вимірів до простору ознак може значно покращити придатність даних. Водночас збільшення розмірності простору ознак може призводити до перенавчання моделі. Як наслідок, дослідникам потрібен алгоритм для визначення недоліків датасетів, тобто виявлення частки неоднозначних даних і підпросторів з такими даними у навчальному та тестовому наборах даних. Результати роботи алгоритму можна використати для прийняття рішення про модифікацію 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. Масштабні датасети з великим числом ознак неможливо візуалізувати без додаткових перетворень. У цьому разі для відображень даних з простору великої розмірності у простір меншої розмірності можна використовувати такі алгоритми, як PCA [2], tSNE [3], та нанести отриманий результат на двовимірний графік. До недоліків цих методів слід віднести залежність від людської суб'єктивності та нездатність ефективно візуалізувати й аналізувати дані великої розмірності через великі втрати інформації під час 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 разі йому також потрібен алгоритм, який оцінить потенційну точність класифікації на заданому наборі даних і визначить ненадійні екземпляри для корекції навчання моделі.

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

Для задачі класифікації зображень визначення неоднозначних екземплярів даних стає особливо важливим. Зазвичай згорткова нейронна мережа (CNN) [4]

посилає відображення з простору зображень у простір ознак, таким чином вхідне зображення перетворюється на вектор ознак. Остаточна класифікація виконується на базі одновимірного вектора. Виявлення неоднозначних екземплярів даних у просторі ознак є важливою науковою проблемою для покращення якості розпізнавання зображень за допомогою CNN. Варто зауважити, що вищевказане перетворення у простір ознак недетерміноване, а роздільність класів залежить не лише від якості даних, а й від якості роботи згорткових шарів нейронної мережі.

У цьому дослідженні запропоновано новий метод виявлення неоднозначних екземплярів даних на основі методу K найближчих сусідів (KNN — K Nearest Neighbors) [5], продемонстровано його роботу на супутникових даних (багатоспектральних оптичних знімках) у задачі класифікації сільськогосподарських культур та розглянуто можливі варіанти використання даного алгоритму під час розробки моделей машинного навчання. Вперше ідею цього методу представлено на конференції ICAIIT 2023, де його роботу проілюстровано на прикладі штучно згенерованого набору даних в двовимірному просторі [6]. У даній роботі метод узагальнено та продемонстровано на багатовимірних даних, а також описано конкретний алгоритм його реалізації і можливості застосування на всіх стадіях розробки моделі машинного навчання.

1. Огляд літератури

Проблемі перекриття класів у датасетах присвячені численні дослідження. Більшість з них розглядають етап навчання моделі та пропонують модифікації стратегії вибору навчальних даних для згладжування впливу недосконалості датасету [7]. Для удосконалення і створення збалансованого набору даних часто використовують методи аугментації (штучного доповнення датасету) [8]. Проте таке доповнення навчального набору даних прийнятне та ефективне не в усіх задачах і предметних областях. Проблему незбалансованості класів можна пом’якшити, або ігноруючи дані в неоднозначних регіонах, або ігноруючи найбільш численний клас в таких регіонах, або створюючи окремі правила для довірених і ненадійних даних [9]. Це демонструє важливість надійного методу виявлення підпросторів, де відбувається перекриття класів.

Одним з відомих підходів для аналізу наборів навчальних даних є використання методу KNN [10]. Досліджувалася продуктивність алгоритму KNN на наборах даних, де мали місце проблеми незбалансованості та перекриття класів [11]. Під час експериментів використовувалися штучно згенеровані датасети. Основними гіперпараметрами для його генерації були відсотки зон перекриття, а також співвідношення класів у них.

У роботі [12] описано метод вирішення обох проблем: перекриття класів і незбалансованість наборів даних. Для цього використовується модифікація навчального процесу, спрямована на балансування класів, які модель отримує на вхід при навчанні. Комбінація одразу двох проблем у датасеті вкрай негативно впливає на якість роботи моделі, яка була на ньому навчена. Така модель матиме тенденцію класифікувати всі екземпляри даних у суперечливих зонах як найбільш представлений у датасеті клас, ігноруючи класи з меншою кількістю екземплярів. Автори пропонують за допомогою підходу, заснованого на KNN, визначати, які екземпляри даних найпоширеніших класів знаходяться поблизу представників найменш поширених класів. Згодом ці точки видаляються з набору навчальних даних, зменшуючи частку найбільш представленого класу і таким чином балансуючи класи. Як наслідок, модель буде навчатися приділяти практично однакову увагу всім представленим класам.

Автори [13] також досліджують випадок одночасного перекривання класів і проблему їх незбалансованості. Вони пропонують видалити найбільш представлений клас у зоні перекриття класів. Наступним етапом є генерація нових датасетів для ансамблю моделей шляхом випадкового відбору збалансованого набору даних з базового незбалансованого датасету. Ансамбль робить прогноз на основі низки простих класифікаторів, кожен з яких навчається на власному унікальному наборі даних.

Інший спосіб вирішення проблеми дисбалансу класів — це окремі класифікації для різних підпросторів у просторі ознак. У [14] дослідники пропонують тренувати дві класифікаційні моделі: перша приймає бінарне рішення щодо приналежності до підпростору, де наявне перекриття класів; друга, на основі SVM, приймає рішення про класифікацію у разі приналежності екземпляру даних до надійного підпростору. Кілька досліджень також вводять дві моделі: першу — для підпросторів з проблемою перекриття класів, другу — для решти простору.

Подібний підхід також можна використовувати для динамічного вибору моделі з ансамблю на етапі визначення остаточного рішення ансамблю [15]. Крім підпросторів, де класи перекриваються, існують «локальні несправедливі» зони, де більшість моделей ансамблю не можуть зробити правильне передбачення, і як наслідок, їх ансамбль також зробить помилку. Однак менша частина моделей може зробити правильне передбачення. Автори [15] запропонували рішення для виявлення таких моделей і вибору підмножини моделей, які здатні робити коректне передбачення у даній конкретній зоні. Таким чином, остаточна модель матиме кращі метрики для екземплярів даних у всьому просторі ознак.

Розглянуті методи боротьби з перекриттям класів сильно залежать від правильності визначення проблемних зон та дещо спотворюють оригінальний датасет, видаляючи небажані дані, що не зовсім коректно. Це ще раз доводить важливість точного визначення ненадійних екземплярів даних для подальших маніпуляцій. Також важливо використовувати дану інформацію не лише під час навчання моделі, а й на інших етапах розробки.

2. Проблема визначення неоднозначних екземплярів даних у вибірках для задач машинного навчання

Під час вирішення завдання класифікації можуть виникнути дві проблеми — це перекриття класів і викиди (рис. 1: *а* — проблема викидів; *б* — проблема перекриття класів). Проблема викидів полягає у наявності серед множини екзамплярів класу поодиноких представників з нехарактерними ознаками. У просторі ознак такі поодинокі представники потрапляють у підпростір, де розташовані «типові» представники іншого класу (рис. 1, *а*). Проблема перекриття класів показана рис. 1, *б*. 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, зокрема супутникових даних. Пропонуємо алгоритм, який вирішує обидві проблеми одночасно, виявляючи неоднозначні та ненадійні дані в датасеті. Неоднозначними або ненадійними будемо називати 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.

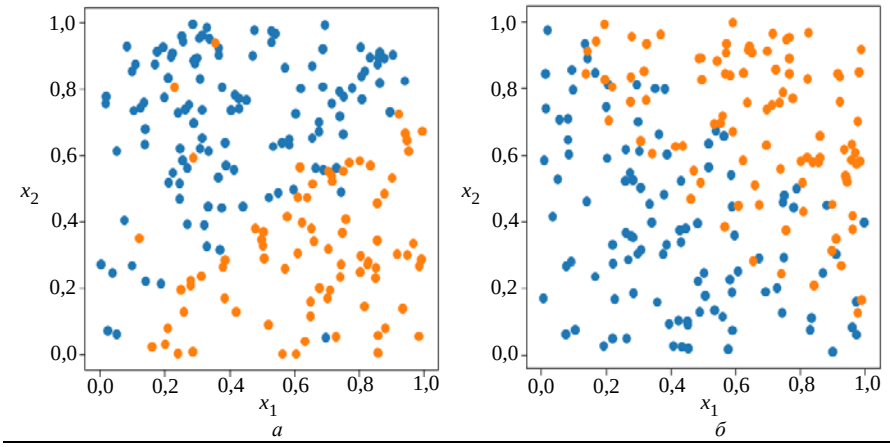
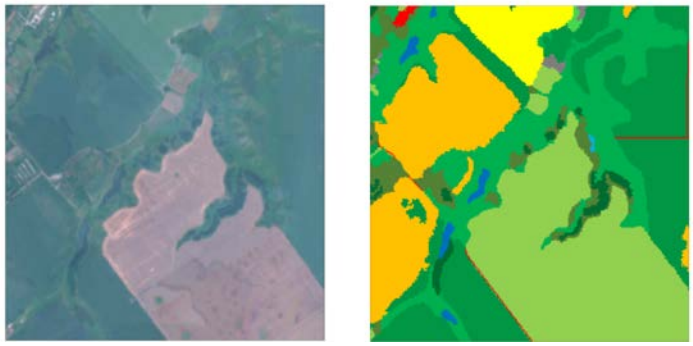


Рис. 1

3. Опис даних

У ролі досліджуваного набору даних розглянемо навчальну вибірку для задачі класифікації земного покрову, використану для побудови нейромережевої задачі класифікації в [16]. Вона включає дані оптичного супутника Sentinel-2 з просторовим розрізненням 10 м, які покривають Київську область. Для уникнення проблеми хмарності в даному дослідженні використовується композит, отриманий з супутникових знімків Київської області, знятих супутником Sentinel-2 впродовж липня 2021 року. Даний композит має чотири канали: синій (490 нм), зелений (560 нм), червоний (665 нм) та інфрачервоний (842 нм). Фрагмент триканального зображення композиту для Київської області в палітрі true color, яка включає канали 490, 560 та 665 нм, показаний на рис. 2, *а*. Цей композит залучали до побудови карти класифікації земного покриву з використанням згорткової нейромережевої моделі, розробленої Інститутом космічних досліджень НАНУ та ДКАУ та описаної в [17]. Приклад фрагменту карти класифікації показаний на рис. 2, *б*.



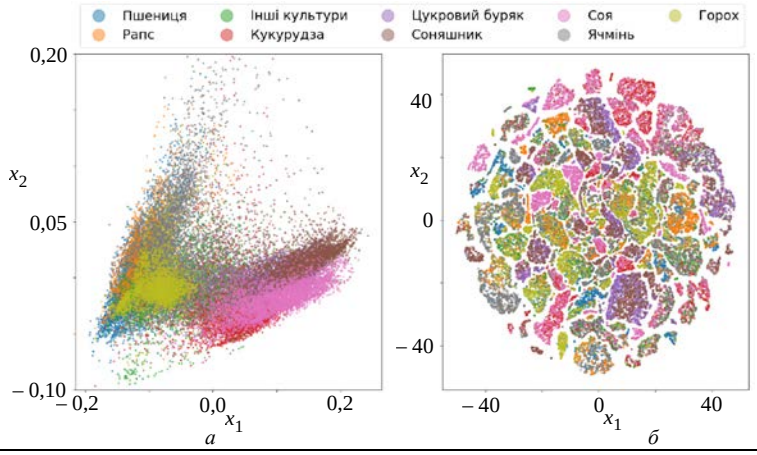
а *б*

Рис. 2

Розглянемо задачу попіксельної класифікації, що зводить дану проблему до задачі класифікації на декілька класів за чотирма ознаками. Для формування навчального датасету виберемо вісім класів сільськогосподарських культур, інші культури віднесемо до окремого класу — «Інші культури». Класи, які не відносяться до сільськогосподарських культур, відкинуті, наприклад забудівлі чи водні ресурси.

Для зменшення обсягу досліджуваного набору даних з восьми класів довільно виберемо 25 тисяч відповідних йому пікселів на супутниковому композиті. Якщо певний клас включає меншу кількість пікселів, для експерименту візьмемо всі.

На рис. 3 представлені візуалізації отриманого датасету: *а* — візуалізація датасету за допомогою PCA; *б* — за допомогою tSNE. Через неможливість візуалізації чотиривимірних даних на площині розмірність даних була зменшена до двовимірної за допомогою алгоритмів PCA та tSNE.



а *б*

Рис. 3

ISSN 2786-6491

Отриманий датасет містить дев’ять класів та має проблему перекриття класів. Для візуалізації використовуємо редуковані двовимірні дані. Зазначимо, що всі обчислення запропонованого алгоритму виконуються на повних чотиривимірних даних, а сам алгоритм здатний працювати з даними будь-якої розмірності.

4. Метод визначення неоднозначних екземплярів даних алгоритму

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 $\vec{x}_i \in \hat{X}$ отримаємо вектор $\vec{m}_i$, де елемент m_i^j — це результат класифікації $\vec{x}_i$ за допомогою класифікатора KNN з параметром числа сусідів рівним j, який було навчено за допомогою набору даних $\hat{X} \setminus \vec{x}_i$:

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

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

$$M : M_i^j = m_i^j.$$

Тепер можна порівняти кожен вектор $\vec{m}_i$ з істинним класом y_i. Розглянемо декілька можливих варіантів:

- більшість елементів $\vec{m}_i$ відповідає справжньому класу y_i;
- перші елементи $\vec{m}_i$ відповідають справжньому класу y_i, решта — ні;
- більшість елементів $\vec{m}_i$ не відповідає справжньому класу y_i;
- передбачений клас m_i^j постійно змінюється в залежності від j, відбуваються «стрибки між класами».

Для того щоб екземпляр даних сприймався як надійний та однозначний, вважаємо перші дві умови обов’язковими. Однак перша умова завжди істинна, якщо 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 зони 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, невеликі цілі числа. Загалом 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, третя умова C_3: якщо присутня часта зміна класів, екземпляр даних ненадійний. Математично 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. Сума результату згортки не повинна виконувати деяке достатньо низьке ціле число q.

Таким чином, отримаємо остаточне правило для класифікації екземпляру даних:

$$C_1 \wedge \bar{C}_2 \wedge \bar{C}_3,$$

де k, r, q — гіперпараметри. Змінюючи їх значення, зробимо одну умову важливішу за іншу. Як правило, $q < r, q < k$.

Алгоритм визначення ненадійних представників даних в датасеті мовою псевдокоду можна представити таким 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
```

Результати роботи алгоритму. За результатами роботи алгоритму на супутниковому композиті визначено ненадійні точки для задачі класифікації за чотирма ознаками. Візуалізацію надійних та ненадійних точок представлено на рис. 4.

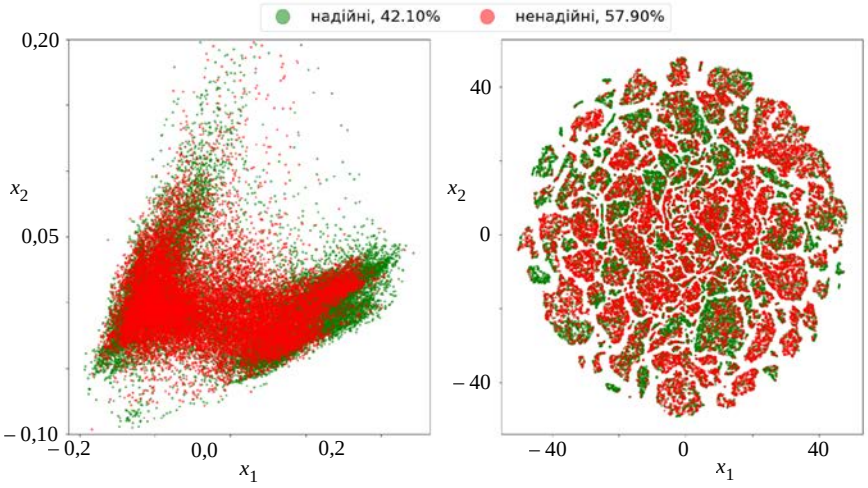


Рис. 4

ISSN 2786-6491

Також обраховано відсоток ненадійних точок кожного класу та загалом, дані наведено у таблиці.

Таблиця

Назва культури	Кількість пікселів, тис.	Відсоток ненадійних точок, %
Пшениця	25	51,36
Рапс	25	63,17
Кукурудза	25	68
Цукровий буряк	25	78,66
Соняшник	25	61,99
Соя	25	54,88
Ячмінь	25	47,44
Горох	18	60,55
Інші культури	25	35,8
Разом	218	57,90

Як видно з таблиці, більшість точок з датасету потрапила до ненадійних даних, що свідчить про неможливість точної попіксельної класифікації обраних восьми видів сільськогосподарських культур на основі єдиного знімку. Це узгоджується з емпіричними результатами сучасних досліджень у сфері дистанційного зондування Землі, які свідчать про те, що різні типи сільськогосподарських культур неможливо відрізнити на основі одного знімку. Тому в сучасних роботах для класифікації сільськогосподарських культур використовуються часові ряди супутникових даних, отриманих протягом всього вегетаційного сезону [18, 19]. Для подолання цієї проблеми необхідно розширити датасет різночасовими знімками з додатковими каналами та/або використати просторові властивості знімків, наприклад, за допомогою згорткових нейронних мереж.

5. Можливості практичного застосування

Алгоритм визначення ненадійних екземплярів даних здатний стати ще одним інструментом дослідника і може використовуватися протягом усього циклу розробки моделі машинного навчання; від збору даних до розгортання моделі для використання в реальних умовах (рис. 5).

У наступних підрозділах детально розглядаються можливі варіанти використання запропонованого алгоритму.

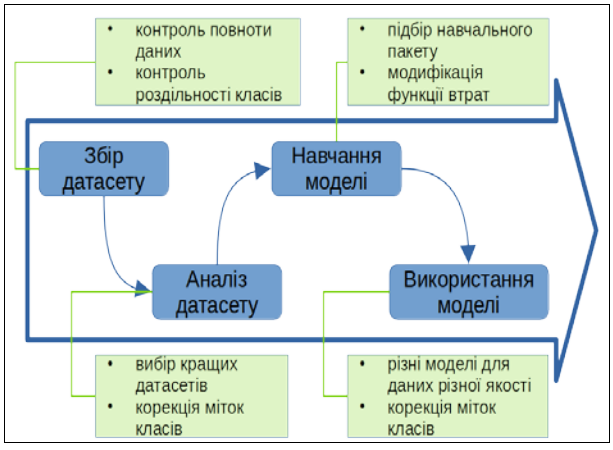


Рис. 5

Оцінка якості датасету. Представлений алгоритм може бути корисним як під час збору даних, так і при аналізі наявних датасетів. Більшість сучасних задач машинного навчання вимагає великих датасетів з великою розмірністю простору ознак та численними екземплярами даних. Досить часто відсутній набір даних, який відповідає поточній задачі. Тому багато дослідників змушені конструювати свій власний датасет, наприклад, через аутсорсинг або краудсорсинг [20, 21]. Подібна ситуація виникає, наприклад, коли в ролі навчальних даних в державних системах агромоніторингу використовується інформація про посіви, надана фермерами. Пілотний проєкт, проведений в Україні із залученням даних від респондентів Державного департаменту статистики, показав, що недостовірна інформація в даних респондентів може складати до 30 %. У цій ситуації надзвичайно важливо мати інструмент для контролю якості отриманих даних. Таким чином, можна виявити проблему на ранніх стадіях, і дослідники матимуть змогу скорегувати процес збору даних.

Можлива і протилежна ситуація, коли для даної проблеми існує декілька датасетів. Зазвичай вибираються найкращі з них або їх комбінації. Маючи можливість визначити відсоток неоднозначних даних, дослідник може вибрати дані найкращої якості. Також можливо побудувати новий об'єднаний датасет з кількох найкращих, з найменшою часткою ненадійних даних у них.

Корекція формування навчального пакета. Здатність визначити ненадійні та неоднозначні екземпляри даних дає можливість покращити процес навчання моделі. Один з можливих методів — це змінити стратегію формування пакета даних при навчанні.

Під час навчання моделі можна використовувати надійні дані для навчання частіше, ніж ненадійні. Змінюючи частоту потрапляння ненадійних даних у навчальний пакет, можна надати моделі бажані властивості, аж до повного видалення неоднозначних екземплярів даних. Отримана модель прийматиме рішення здебільшого на основі надійних зразків даних. Це дозволяє робити 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 чітко розрізняти різні типи екземплярів даних за їх якістю дозволяє навчити кілька моделей, кожна з яких працює лише у певних зонах простору ознак. Таким чином, правила класифікації для підпростору з надійними екземплярами даних будуть відрізнятися від ненадійних. Остаточний ансамбль моделей складатиметься з двох підмножин: 1) з моделей, призначених для надійних даних; 2) з моделей, призначених для даних, які знаходяться біля неоднозначних екземплярів даних у просторі ознак. Подібний підхід до побудови динамічних ансамблів моделей у разі незбалансованих вибірок запропоновано в [15].

Варто зауважити, що ненадійний екземпляр даних може або лежати в зонах, що перекриваються, або бути викидом. Зрозуміло, що в останньому випадку не можемо використовувати другу підмножину класифікаторів і слід фільтрувати такі випадки, тому потрібне подальше вдосконалення даного підходу.

Модифікація датасету. Багато методів роботи з незбалансованими датасетами і проблемами перекриття класів пропонують видалити з датасету екземпляри даних, які відносяться до найбільш представленого класу та знаходяться в зонах перекриття класів [22]. Натомість можна внести зміни до їх істинного класу без зайвого видалення даних. Потрібно замінити мітки класів у зонах, де класи перек-

риті. 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. Щоб боротися з проблемою дисбалансу класів, усі мітки класів у сумнівній зоні можна змінити на найменш представлений клас.

Як і в попередньому підрозділі, необхідно відфільтрувати викиди, для них мітки класів можна змінити на мітки екземпляра даних, які оточують їх у просторі ознак.

Висновок

Дана робота присвячена дослідженню проблеми перекриття класів разом з іншими можливими важкими випадками, такими як викиди та незбалансованість класів. Існує безліч причин виникнення описаних проблем, найважливіші — недостатня точність збору даних при формуванні датасету та замала кількість ознак, з чого випливає надто мала розмірність простору ознак. Незалежно від причин, деякі екземпляри даних просто неможливо класифікувати правильно, оскільки вони мають схожі ознаки з іншими екземплярам з іншою міткою класу. Як приклад можна розглянути задачу сегментації супутникових знімків для визначення землекористування [23]. Через схожість спектральних характеристик таких культур, як пшениця та ячмінь, їх не можна розділити у просторі ознак. Такі екземпляри даних заважають процесу навчання моделі та можуть стати причиною неправильного передбачення на реальних даних.

У цьому дослідженні представлено новітній алгоритм, який дає змогу будь-якому досліднику мати більш чітке уявлення про якість датасетів та не залежить від розмірності простору ознак. На відміну від методів візуальної оцінки, заснованих на декомпозиції, описаний алгоритм видає детерміновані числові показники якості даних, наприклад відсоток надійних даних. Крім того, він дозволяє чітко відрізнити достовірні екземпляри даних від недостовірних, відкриваючи можливість коригування.

Розглянуто задачу попіксельної класифікації чотириканального супутникового композиту для визначення сільськогосподарських культур. Визначено ненадійні точки, які важко класифікувати правильно, та обчислено відсоток ненадійних даних загалом та по кожному класу окремо. Більшість точок з датасету, частка яких становить 59,7 %, не може бути чітко відділена 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 вказує на необхідність використовувати більшу кількість знімків та оптичних каналів для задач класифікації землекористування або ж математичну модель, яка враховує просторову структуру сигналів на знімку.

Запропонований алгоритм може використовуватися протягом усього циклу розробки моделі машинного навчання — це контроль та корекція процесу збору навчальних даних, а також вибір найкращих датасетів та їх змішування. Виявивши неоднозначні екземпляри даних, можна застосувати різні правила для них і для надійних даних як під час навчання, так і під час роботи на реальних даних. Алгоритм також відкриває простір для модифікації набору даних, мітки класів ненадійних зразків даних змінюються згідно з поточними задачами, такими як високі метрики або балансування класів.

DATA MINING OF MACHINE LEARNING DATASETS FOR HARD CASE IDENTIFICATION

Anton Okhrimenko

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,

ant.ohrimenko@ukr.net

Nataliia Kussul

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,

nataliia.kussul@gmail.com

This article addresses the issue of ambiguity in classification tasks in the field of machine learning. Classification involves training a model that is able to distinguish between data samples belonging to different classes. However, there are situations where correctly classifying certain data samples becomes a difficult or even impossible task, regardless of the complexity of the machine learning model. In this study, a method and algorithm for detecting such ambiguous data samples are proposed. The method is based on the nearest neighbor approach and analyzes the class labels of data samples that are closely located in the feature space, it makes possible the identification of a subset of ambiguous data samples that may negatively impact the classification model's training process. To demonstrate the practical application of the algorithm, an experiment was conducted using a four-channel satellite composite for pixel-to-pixel classification of agricultural crops. The percentage of unreliable data was determined both total and separately for each crop. One of the main findings of the research is the potential use of the proposed algorithm in constructing the dataset for classification model training. It helps identify potentially problematic data samples and ensures the quality of the input data set. Additionally, there were considered the possibilities of applying the algorithm after the model training process while using it in operational mode. Detecting ambiguous data samples can help identify potential classification errors and improve the model's performance. The presented algorithm can be a valuable tool for researchers through the entire cycle of machine learning model development, starting from data preparation for training and ending with its deploying for a practical use. An algorithm can contribute to reducing the time required for obtaining high quality training data, improving the classification metrics, and providing more reliable results in machine learning tasks.

Keywords: machine learning, classifier, the nearest neighbor method, dataset quality assessment, imbalanced datasets, hard cases.

ПОСИЛАННЯ

1. Chawla N.V. Data mining for imbalanced datasets: an overview. *Data Mining and Knowledge Discovery Handbook*. 2006. P. 875–888. DOI:10.1007/0-387-25465-x_40
2. Abdi H., Williams L.J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. *Wiley Interdisplinary Rev. Comput. Stat.*, 2010. P. 433–459.
3. Van Der Maaten L., Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res*. 2008. Vol. 9. P. 2579–2605.
4. O'Shea K., Nash R. An introduction to convolutional neural networks. arXiv:1511.08458. 2015. P. 1–11.
5. Cover T.M., Hart P.E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*. 1967. Vol. 13, N 1. P. 21–27. DOI: 10.1109/TIT.1967.1053964
6. Okhrimenko A., Kussul N. KNN-based algorithm of hard case detection in datasets for classification. *Proceedings of the 11th International Conference on Applied Innovations in IT*, Anhalt University of Applied Sciences, Mar. 2023. [Online]. P. 113–118. http://dx.doi.org/10.25673/101926

7. Mikołajczyk A., Grochowski M. Data augmentation for improving deep learning in image classification problem. *International Interdisciplinary PhD Workshop, IIPhDW 2018*. P. 117–122. DOI: 10.1109/IIPHDW.2018.8388338
8. Shorten C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *J. Big Data*. 2019. Vol. 6, N 1. P. 1–48. DOI: 10.1186/s40537-019-0197-0
9. Almutairi W.A., Janicki R. On relationships between imbalance and overlapping of datasets. *EPiC Series in Computing*. 2020. P. 141–150. DOI: 10.29007/h71z
10. Kramer O. Dimensionality reduction with unsupervised nearest neighbors. *Intell. Syst. Ref. Libr*. 2013. Vol. 51. P. 13–23. DOI: 10.1007/978-3-642-38652-7
11. García V., Mollineda R.A., Sánchez J.S. On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Anal. Appl*. 2008. Vol. 11, N 3–4. P. 269–280. DOI: 10.1007/s10044-007-0087-5
12. Nwe M.M., Lynn K.T. KNN-based overlapping samples filter approach for classification of imbalanced data. *Studies in Computational Intelligence*. 2020. P. 55–73. DOI: 10.1007/978-3-030-24344-9_4
13. Chen L., Fang B., Shang Z., Tang Y. Tackling class overlap and imbalance problems in software defect prediction. *Softw. Qual. J*. 2018. Vol. 26, N 1. P. 97–126. DOI: 10.1007/s11219-016-9342-6
14. Tang Y., Gao J. Improved classification for problem involving overlapping patterns. *IEICE Trans. Inf. Syst.* 2007. Vol. E90-D, N 11. P. 1787–1795. DOI: 10.1093/ietisy/e90-d.11.1787
15. Lässig N., Oppold S., Herschel M. Metrics and algorithms for locally fair and accurate classifications using ensembles. *Datenbank-Spektrum*. 2022. Vol. 22, N 1. P. 23–43. DOI: 10.1007/s13222-021-00401-y
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 Sens.* 2022. Vol. 14, N 1. 35 p. 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. *International Geoscience and Remote Sensing Symposium (IGARSS)*. 2018. P. 4812–4815. DOI: 10.1109/IGARSS.2018.8518263
18. Shelestov A., Lavreniuk M., Kussul N., Novikov A., Skakun S. Large scale crop classification using Google earth engine platform. *International Geoscience and Remote Sensing Symposium (IGARSS)*. 2017. P. 3696–3699. DOI: 10.1109/IGARSS.2017.8127801
19. 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. *IEEE*. 2019. P. 698–740. DOI: 10.1109/igarss.2019.8900517
20. Zheng F. *et al.* Crowdsourcing methods for data collection in geophysics: state of the art, issues and future directions. *Reviews of Geophysics*. 2018. Vol. 56, N 4. P. 698–740. DOI: 10.1029/2018RG000616
21. Bayas J.C.L. *et al.* Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. *Remote Sens.* 2016. Vol. 8, N 11. P. 1–36. DOI: 10.3390/rs8110905
22. 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, N 4. P. 198–201. DOI: 10.1145/3343440
23. 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. *International Geoscience and Remote Sensing Symposium (IGARSS)*. 2016. DOI: 10.1109/IGARSS.2016.7729043

Отримано 26.05.2023