

Експериментальне Порівняння Лінійного та Нелінійного SVM-класифікаторів

Катерина Горішня^{1*†} та Володимир Кобзєв^{1†}

¹ Харківський національний університет радіоелектроніки, пр. Науки, 14, м. Харків, 61166, Україна

Анотація

У цій роботі проведено експериментальне порівняння лінійного та нелінійного методів опорних векторів (SVM) для класифікації даних із нелінійними межами. Дослідження виконано на наборі даних `make_circles` із використанням різних типів ядер. Результати демонструють значну перевагу нелінійних ядер над лінійними у задачах із складними розділювальними межами. Отримані результати підкреслюють важливість правильного вибору ядра для забезпечення максимальної ефективності SVM-класифікації.

Ключові слова

класифікація, метод опорних векторів, лінійне ядро, нелінійне ядро, `make_circles`, медичні дані

1. Вступ

Обробка наявних результатів медичних досліджень за допомогою методів Big Data дозволяє виявляти закономірності, які пов'язані з перебігом захворювань [1]. Ці методи дають можливість аналізувати величезні обсяги інформації, що утворюються внаслідок сучасних томографічних досліджень. Результатом таких досліджень є набори зображень, які можуть містити ознаки захворювань різного ступеня тяжкості.

Томографічні зображення є основою для аналізу кількості, розмірів, розташування клітин і їх інших характеристик. Класифікація таких зображень за формальними правилами дає медичному персоналу базу для діагностики важких захворювань, визначення стадії розвитку хвороби та підготовки до вибору методів лікування. Актуальність статистичного аналізу такої інформації пояснюється наступними моментами.

1. Онкологічні захворювання є однією з головних причин смертності людей у світі.
2. Дані, пов'язані з онкологічними захворюваннями, мають великий обсяг інформації, використання методів виявлення закономірностей в яких може допомогти у визначенні характерних ознак, що вказують на розвиток раку.
3. Виявлені закономірності або аномалії допомагають окреслити характеристики пацієнта, які визначають ефективний метод лікування.

2. Метод опорних векторів

Для аналізу таких даних важливо вибрати метод, який не лише ефективно класифікує клітини на здорові та хворі, але й забезпечує точну локалізацію ракових клітин на

Information Systems and Technologies (IST-2024), November 26-28, 2024, Kharkiv, Ukraine

** Corresponding author.*

† These authors contributed equally.

✉ kateryna.horishnia@nure.ua (К. Горішня); volodymyr.kobziev@nure.ua (В. Кобзєв)

ORCID [0009-0001-9032-4249](https://orcid.org/0009-0001-9032-4249) (К. Горішня); [0000-0002-8303-1595](https://orcid.org/0000-0002-8303-1595) (В. Кобзєв)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

зображеннях. Метод опорних векторів (SVM) є одним із найефективніших для таких завдань, оскільки дозволяє визначити чіткі межі між класами клітин і виявити приховані закономірності в розташуванні хворих клітин [2, 3].

В основі методу лежить уявлення про гіперплощину, яка розділяє точки (клітини) в просторі відповідно до їх класу. У найпростішому випадку гіперплощина відповідає оболонці органу, навколо якої розташовані клітини. Однак в реальних системах класи рідко бувають лінійно роздільними. У таких випадках застосовуються нелінійні ядра, такі як поліноміальні та радіальні базисні функції, які забезпечують адаптацію до складних форм меж [4].

Локалізація та класифікація клітин на томографічних зображеннях ґрунтується на кількох ключових підходах, які забезпечують точне визначення координат і стану клітин. Першим етапом є визначення центру (середини) групи точок, що представляють клітини на зображенні. Це досягається шляхом обчислення середнього положення точок на основі їх координат. Такий аналіз дозволяє ідентифікувати центральне розташування патології, що є важливим для подальшої діагностики.

Наступним кроком є обчислення радіусу зони класифікації. Радіус визначається як відстань від центру до найбільш віддалених точок хмари, які відповідають клітинам. Це дозволяє встановити межі аналізованої зони, в межах якої відбувається класифікація. Використовуючи SVM, ці межі можуть адаптивно розширюватися залежно від характеристик ядра.

Класифікація точок відбувається у багатовимірному просторі, що дозволяє враховувати їхні просторові координати та інші властивості. Метод SVM створює оптимальну межу між класами, розділяючи точки на здорові та хворі. У випадках, коли ракові клітини знаходяться на межі оболонки органу або виходять за її межі, SVM забезпечує точність класифікації шляхом налаштування ширини розділювальної смуги. Це дозволяє мінімізувати помилки та враховувати специфіку розташування клітин, що є критично важливим для аналізу ранніх стадій онкологічних захворювань.

3. Проведення експерименту

Для проведення експерименту було обрано набір даних `make_circles` із бібліотеки `Scikit-learn` [5]. Цей набір є особливо цінним для тестування методів класифікації, оскільки утворює два концентричні кола, які моделюють дані зі складними нелінійними межами. Використання таких даних дозволяє оцінити ефективність різних типів ядер у SVM для класифікації точок у просторових структурах із нетривіальними розділювальними межами.

У рамках експерименту було протестовано лінійне ядро (Linear), поліноміальні ядра другого та третього порядків (Poly-2, Poly-3) та радіальне базисне ядро (RBF). Основними метриками для оцінки ефективності класифікації стали точність (Precision), повнота (Recall) та F1-міра (F1-score). Результати експерименту наведені в таблиці 1.

Таблиця 1

Результати експерименту

Ядро	Precision	Recall	F1-score
Linear	0.5372	0.55	0.5330
Poly-2	0.7138	0.68	0.6790
Poly-3	0.8297	0.78	0.7784
RBF	0.9634	0.96	0.9602

Лінійне ядро (Linear) продемонструвало найгірші результати серед усіх протестованих варіантів. Це пояснюється тим, що лінійне ядро ефективно лише для задач із лінійно

розділними межами між класами. У випадку набору даних `make_circles`, де класи утворюють концентричні кола, лінійне ядро не може коректно описати розділювальні межі. Як наслідок, точність, повнота та F1-міра для цього ядра залишаються низькими.

Поліноміальне ядро другого порядку (Poly-2) показало значно кращі результати. Воно здатне враховувати дугоподібні межі між класами, що робить його придатним для задач, де межі мають просту криволінійну форму. Однак Poly-2 має обмежену здатність до адаптації, через що його ефективність знижується при обробці більш складних структур даних.

Поліноміальне ядро третього порядку (Poly-3) забезпечило ще кращу адаптацію до складних меж між класами. Це свідчить про те, що збільшення ступеня полінома дозволяє краще враховувати нерівномірний розподіл даних і складніші форми меж. Poly-3 показало помітне зростання точності, повноти та F1-міри, що підтверджує його переваги над ядром Poly-2.

Найвищі результати продемонструвало радіальне базисне ядро (RBF). Завдяки своїй здатності враховувати складні структури даних, RBF забезпечує максимальну точність, повноту та F1-міру серед усіх протестованих ядер. Його ефективність пояснюється тим, що RBF ядро враховує не лише просторове розташування точок, але й їх відстань від центру класифікаційної зони, дозволяючи створити найбільш оптимальну межу між класами. Це робить RBF ідеальним вибором для задач, де класи утворюють складні нелінійні структури, як-от концентричні кола.

4. Висновки

Результати експерименту підтвердили, що метод SVM із лінійним ядром підходить лише для задач, де класи розділені прямою лінією. У задачах з більш складними геометричними формами об'єктів і меж поміж ними, що представлено у наборі даних `make_circles`, ефективність класифікації значно зростає при використанні нелінійних ядер, таких як поліноміальні або радіальне базисне ядро. Ядро RBF продемонструвало свою особливу перевагу завдяки здатності адаптуватися до складних форм меж та мінімізувати помилки класифікації.

Таким чином, результати дослідження підкреслюють необхідність використання нелінійних ядер для задач із нелінійними межами, особливо у випадках аналізу медичних даних, де точна класифікація є критично важливою для ранньої діагностики та розробки персоналізованих підходів до лікування.

Список використаних джерел

- [1] Choong, H.-J.Y., & Lee, C.H. (2017). Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, 36(1), 3–11. <https://doi.org/10.23876/j.krcp.2017.36.1.3>.
- [2] Кобзєв В.Г., Яковлєв С. В., Назаров О.С., Назарова Н.В., Горішня К.О. Модифікації методу опорних векторів для класифікації та виявлення аномалій у задачах обробки зображень // Застосування інформаційних технологій у підготовці та діяльності сил охорони правопорядку: Збірник тез доповідей Міжнародної науково-практичної конференції. - Харків: НАНГ, 2024. - с. 149-151.
- [3] Kobziev V., Yakovlev S., Gorishnia K. Modifications of the support vector method for classification and detection of anomalies in image processing problems // Інформаційні технології та комп'ютерне моделювання; матеріали статей Міжнародної науково-практичної конференції, м. Івано-Франківськ, 21-24 травня 2024 року. – Електронне видання. Івано-Франківськ: Прикарп. нац. ун-т ім. В. Стефаника, 2024. – с. 221-222.
- [4] Anwar, M. (2018). Support vector machines (SVM): Linear and non-linear SVMs. Medium.
- [5] Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.