КВАНТОВО-ПІДСИЛЕНА СТІЙКІСТЬ ДО ЗМАГАЛЬНИХ АТАК У МАШИННОМУ НАВЧАННІ

Левандовський В.С., Замуруєва О. В., Бондарчук М. В., Захарчук С. І.

Волинський національний університет імені Лесі Українки, Levandovskyi.Viktor@vnu.edu.ua

Машинне навчання (ML) є основою більшості сучасних автономних, інформаційних та інтелектуальних систем, зокрема у сферах безпеки, комп'ютерного зору, розпізнавання образів і прийняття рішень. Водночас однією з ключових проблем ML залишається його вразливість до змагальних (adversarial) атак – спеціально сформованих вхідних даних, які майже не відрізняються від коректних зразків, але призводять до хибної класифікації. Наявність таких атак становить серйозну загрозу для систем, що працюють у критично важливих або безпекових застосуваннях.

Останніми роками значну увагу привертає інтеграція машинного навчання з квантовими обчисленнями, що дало початок напряму квантового машинного навчання (Quantum Machine Learning, QML). Очікується, що QML може забезпечити квантову перевагу не лише у швидкодії або точності, але й у підвищеній стійкості до змагальних атак. Це призвело до формування окремого дослідницького напряму – квантового змагального машинного навчання (Quantum Adversarial Machine Learning, QAML).

Квантові класифікатори зазвичай реалізуються у вигляді варіаційних квантових схем, у яких класичні дані кодуються у квантові стани, обробляються параметризованими квантовими операціями, а результат визначається шляхом вимірювань. Проте ранні теоретичні дослідження показали, що такі класифікатори можуть бути надзвичайно вразливими до змагальних атак через фундаментальну геометричну властивість багатовимірних гільбертових просторів — явище концентрації міри. Згідно з цим явищем, більшість квантових станів у просторі високої розмірності розташовані поблизу меж класифікації, що робить їх чутливими навіть до надзвичайно малих збурень.

Емпіричні дослідження з використанням стандартних наборів даних, зокрема MNIST, підтвердили, що квантові варіаційні класифікатори можуть бути успішно атаковані класичними методами змагальних атак, такими як FGSM або PGD. Ці атаки призводять до різкого падіння точності навіть тоді, коли змагальні приклади візуально майже не відрізняються від оригінальних даних. Ба більше, перші експериментальні реалізації QAML на реальному квантовому апаратному забезпеченні продемонстрували, що вразливість до змагальних атак зберігається і за наявності квантового шуму.

Разом з тим, квантовий шум може відігравати і конструктивну роль у підвищенні змагальної стійкості квантових моделей. Для ілюстрації цього ефекту на рис.1. наведено залежність точності класифікації від параметра шуму p, який моделює, зокрема, деполяризаційний шум у квантовій схемі (1).

$$accuracy(p) = b + (1 - b)(1 - p)^k \quad (1)$$

де $b=0.10$ – «базова» точність (нижня межа), $k=2$ – параметр чутливості до шуму.

Видно, що зі зростанням рівня шуму точність класифікації поступово зменшується, однак цей спад є плавним, а не різким. Така поведінка вказує на те, що випадкові квантові флуктуації можуть частково «згладжувати» змагальні збурення, зменшуючи їхній вплив на результат класифікації.

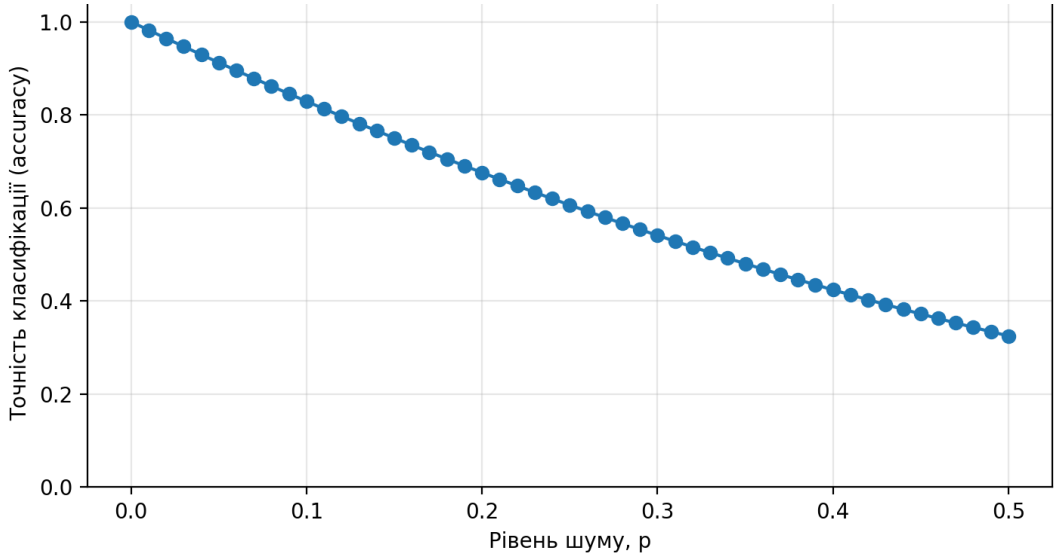

Рис. 1. Залежність точності класифікації від параметра квантового шуму p. Зі збільшенням рівня шуму точність поступово зменшується, що відображає компроміс між точністю та змагальною стійкістю квантового класифікатора.

З фізичної точки зору, шум виконує роль стохастичного усереднення, подібного до термодинамічних флуктуацій у статистичній фізиці. Внаслідок цього малі, спеціально сконструйовані змагальні збурення втрачають ефективність, оскільки їхній вплив маскується випадковими квантовими процесами. Саме цей механізм лежить в основі ідей сертифікованої змагальної стійкості, де гарантії безпеки випливають із фундаментальних меж розрізнення квантових станів.

Окрему увагу приділено методу змагального навчання, коли змагальні приклади включаються безпосередньо у процес тренування. Попри відсутність строгих теоретичних гарантій, цей підхід показав значне підвищення практичної стійкості квантових класифікаторів. Однак залишається відкритим питання узагальнення такої стійкості на нові типи атак і різні схеми кодування даних.

У підсумку, квантове змагальне машинне навчання є перспективним і швидко зростаючим напрямом досліджень. Хоча квантові моделі не є апріорно захищеними від змагальних атак, вони відкривають нові можливості для побудови більш надійних систем машинного навчання. Подальший розвиток квантового апаратного забезпечення, методів кодування даних та архітектур квантових моделей може зробити QAML важливим інструментом для створення безпечних і стійких інтелектуальних систем майбутнього.

Список літератури

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep Learning. Nature 521, 436–444 (2015).

2. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G. & Roli, F. Evasion attacks against machine learning at test time. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 387–402 (Springer, 2013)

3. Du, Y., Hsieh, M.-H., Liu, T., Tao, D. & Liu, N. Quantum noise protects quantum classifiers against adversaries. Physical Review Research 3, 023153 (2021).