ОБЧИСЛЮВАЛЬНІ ПРОЦЕСИ ТА СИСТЕМИ

Doi: https://doi.org/10.15407/emodel.42.05.051
УДК 004.85

Н.І. Недашківська, д-р техн. наук, **С.О. Лупаненко**
Інститут прикладного системного аналізу
Національного технічного університету України
«Київський політехнічний інститут ім. Ігоря Сікорського»
Україна, 03056, Київ, пр-т Перемоги, 37,
тел. +38 067 899 1473; e-mail: n.nedashkivska@gmail.com;
тел. +38 098 671 0950; e-mail: lupanenko.sophia@gmail.com

Порівняльний аналіз моделей машинного навчання для прогнозування поширення коронавірусу COVID-19 в різних країнах

Побудовано математичні моделі опису поширення коронавірусу COVID-19 в різних країнах. Проведено порівняльний аналіз цих моделей для США, Мексики, Росії, Бельгії та України. Початкові дані щодо кількості випадків отримано зі щоденних звітів Всесвітньої організації охорони здоров'я та Центру системних наук та інженерії при Університеті Джона Хопкінса. Для моделювання поширення коронавірусу обрано два потужних методи машинного навчання, що прогнозують нелінійні часові ряди: опорних векторів та багатошарових нейронних мереж прямого розповсюдження. Виявлено переваги і недоліки цих методів, розглянуто питання регуляризації. Побудову і навчання моделей часових рядів для опису поширення коронавірусу COVID-19 в різних країнах світу, вибір найкращої моделі, побудову прогнозу поширення та візуалізацію результатів виконано у реалізованому програмному модулі в середовищі python з використанням сучасних бібліотек scikit-learn, pandas та matplotlib. За допомогою методу решітчастого пошуку з крос-валідацією підібрано найкращі параметри нейронних мереж та опорних векторів в моделях опису поширення COVID-19 в США, Мексиці, РФ, Бельгії та Україні. На основі побудованих моделей виконано прогнозування кількості приросту захворювань на COVID-19 в цих країнах.

К л ю ч о в і с л о в а: метод опорних векторів, багатошарові нейронні мережі прямого розповсюдження, регуляризація, COVID-19, прогнозування поширення епідемії.

Всесвітньою організацією охорони здоров'я 11 березня 2020 року епідемію коронавірусу COVID-19 названо пандемією [1]. Україна оголосила загальнонаціональний карантин 12 березня 2020 року [2, 3]. Новий тип вірусу, швидко розповсюджуючись, завдає шкоди здоров'ю населення всіх країн світу і суттєво впливає на національні економіки. Тому наразі дуже важливо провести детальне дослідження цієї пандемії.

Перший випадок захворювання людини вірусом COVID-19 було зафіксовано в кінці 2019 року в Китаї в місті Ухань. У людей коронавіруси зазвичай викликають респіраторні хвороби — від звичайної застуди до пневмонії. Після різких спалахів COVID-19 в Італії експерти висловлювалися щодо необхідності масового тестування населення для виявлення позитивних результатів, відслідковування їх контактів, що могло сприяти наступному поширенню захворювання [1]. Тепер тестування виконують майже у всьому світі. До найбільш вразливої частини населення входять групи ризику. Це люди з хронічними захворюваннями та особи старше 60 років. Станом на кінець липня 2020 року на COVID-19 у світі захворіли майже 18 мільйонів людей, померло у світі близько 680 тисяч осіб. Найбільша кількість хворих наразі залишається в США, Бразилії та Індії [4—6].

Науковою спільнотою проводяться дослідження щодо побудови моделей для опису розгортання епідемії коронавірусу COVID-19 [7—16]. В роботі [7] подано аналіз і прогнозування розповсюдження COVID-19 в Китаї, Італії і Франції на основі SIR моделі, яка часто використовується для моделювання епідемій. У [8] на основі SEIR моделі зроблено спробу спрогнозувати потреби в приміщеннях і лікарняних ліжках в Чілі за використання різних стратегій втручання з боку держави, а саме блокування та поєднання стратегій часткової ізоляції, закриття шкіл та університетів, домашнього карантину та соціальної відстані.

Модель SARIIqSq, яка надає пояснення динаміки передачі COVID-19 [9], також базована на епідеміологічній SEIR моделі. На основі моделі, описаної у роботі [9], прогнозують терміни закінчення COVID-19 в Індії. Аналіз та прогноз пандемії COVID-19 в Індії виконано за допомогою генетичного програмування [10].

Відношення кількості щоденних нових заражень в Італії в момент часу t апроксимується розподілом Пуассона і розраховується імовірне число нових заражень на наступний місяць [11]. В [12, 13] описано розроблені гібридні моделі вейвлет-розкладу та ARIMA, за допомогою яких виконують прогнозування кількості смертей [12] та щоденних випадків захворювання [13] на COVID-19 в різних країнах світу.

Прогнозування кількості підтверджених випадків захворювання, смертей і одужань від COVID-19 в Пакистані виконано на основі моделі ARIMA [14]. Узагальнену логістичну модель зростання пандемії COVID-19 в Азії запропоновано в роботі [15]. В [16] описано методи порівняння розвитку епідемії між країнами з різними стратегіями стримування. Результати досліджень підтвердили, що раннє стримування є ключовим фактором у згладжуванні кривої розвитку епідемії [16].

Задача полягає у побудові моделей опису поширення коронавірусу COVID-19 в різних країнах світу та отриманні прогнозів поширення COVID-19 на основі цих моделей. Серед методів машинного навчання для аналізу обрано методи опорних векторів та багатошарових нейронних мереж прямого розповсюдження, а вхідними даними — нелінійні часові ряди про поширення COVID-19.

Методи машинного навчання для прогнозування часових рядів. Одна з основних проблем машинного навчання — створити алгоритм, ефективний не лише для наявних навчальних даних, але й для отриманих нових. В [17] будь-яка модифікація методу, що дозволяє зменшити помилки узагальнення, не зменшуючи помилки навчання, називається регуляризацією. Проаналізуємо два потужних методи – опорних векторів та нейронних мереж — та розглянемо їх регуляризацію.

Метод опорних векторів. За допомогою методу опорних векторів для регресії використовується інвертована ціль на відміну від класифікації. В задачі класифікації відшукується найбільш широка роздільна смуга між двома класами і одночасно обмежується порушення зазору. Регресійна модель опорних векторів, навпаки, дозволяє помістити якомога більше навчальних прикладів на смузі з одночасним обмеженням кількості прикладів за межами цієї смуги, ширина якої визначається параметром ε.

У лінійно нероздільному випадку для «розмиття» відступу використовується гіперпараметр C, який означає, що деяким навчальним прикладам дозволяється порушувати роздільну смугу, коли це покращує узагальнюючу здатність алгоритму. Такий захід вважається регуляризацією: при меншому значенні C модель більш регуляризована, відступ стає більш розмитим і вміщує в себе більшу кількість точок. Оптимальне значення параметру C залежить від конкретного набору даних, який налаштовують за допомогою перехресної перевірки.

Для нелінійної регресії використовують процедуру kernel trick та функцію ядра. Суть цієї процедури полягає в тому, що відшукується гіперплощина у зміненому просторі функцій більшої розмірності, при цьому кожний скалярний добуток у результуючому алгоритмі замінюється нелінійною функцією ядра. Використовуються такі ядра: поліноміальне та RBF—гаусівська радіальна базисна функція.

Метод має декілька переваг:

зведення процесу навчання до задачі квадратичного програмування;

етап прогнозування після навчання моделі займає дуже мало часу;

дійсне узагальнення на нелінійний випадок.

Оскільки на результат впливають точки, які лежать біля відступу, метод придатний для використання багатовимірних даних, зокрема коли кількість ознак є більшою за розмірність навчальної множини. Інші алгоритми не завжди працюють в таких випадках.

Недоліком методу опорних векторів вважають великі обчислювальні витрати, порядку $O(N^3)$, де N — розмірність навчальної вибірки [18, 19]. Параметри методу потрібно ретельно підбирати за допомогою перехресної перевірки, зокрема пошуку за сіткою, що може призвести до значних обчислювальних витрат при зростанні значення N. Ще один недолік — відсутність імовірнісної інтерпретації результатів, яку можна отримати за допомогою внутрішньої перехресної перевірки, проте це також може потребувати великих обчислювальних витрат.

Метод багатошарових нейронних мереж прямого розповсюдження. Проблема вибору структури мережі тісно пов'язана з проблемою перенавчання: занадто складні мережі мають надмірну кількість вільних параметрів, які в процесі навчання налаштовуються не тільки на відновлення цільової залежності, але і на відтворення шуму. Для опису нелінійних даних на практиці буває достатньо використання двошарових мереж. Тришарові нейронні мережі використовуються для опису складних багатозв'язних областей [17—20]. Вважається, що чим більше шарів має мережа, тим більше функцій вона реалізує, і тим важче її навчити, оскільки гірше сходяться градієнтні методи.

Для боротьби з проблемою перенавчання використовується регуляризація, яка в нейронних мережах заснована на зменшенні дисперсії оцінок параметрів при невеликому збільшенні зміщення цих оцінок. У багатьох випадках при регуляризації використовують додавання штрафу за нормою параметра $\Omega(w)$ до цільової функції $J(w; X, y)$ [17]:

$$\tilde{J}(w; X, y) = J(w; X, y) + \alpha\Omega(w),$$

де X, y — вхідний і вихідний вектори; w — вектор ваг мережі; $\alpha \in [0, \infty)$ — параметр регуляризації, який задає вагу доданку Ω відносно J. Наприклад, в задачі лінійної регресії цільовою функцією обирають середньоквадратичну помилку моделі на тестовому наборі даних:

$$J(w; X, y) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2,$$

де вектор $\hat{y}$ — значення прогнозу на основі моделі; y — вектор реальних значень вихідної змінної.

Часто для цільової функції $\tilde{J}(w; X, y)$ використовують штраф за нормою L^2, що називається методом зменшення ваг, гребневою регресією або регуляризацією Тихонова і визначається за формулою $\Omega(w) = \frac{1}{2}\| w \|^2$. Оскільки градієнт регуляризованої цільової функції за параметрами має вигляд $\nabla_w \tilde{J}(w; X, y) = \nabla_w J(w; X, y) + \alpha w$, то на кожному кроці алгоритму

градієнтного спуску перед стандартним оновленням градієнта вектор ваг маємо помножити на постійний коефіцієнт, менший одиниці:

$$w = (1 - \eta\alpha)\, w - \eta \nabla_w J\,(w; X, y).$$

Якщо припустити квадратичну апроксимацію цільової функції в колі того значення ваг w^*, при якому досягається мінімум цільової функції J без регуляризації, то мінімальне значення регуляризованої функції $\hat{J}$ досягається в точці $\tilde{w} = (H + \alpha I)^{-1} H w^*$, де H — матриця Гессе J відносно w в точці w^*. Тому результат зниження ваг — це масштабування w^* вздовж напрямків власних векторів матриці H, а саме компоненти w^* практично дорівнюють нулю для напрямків $\lambda_i << \alpha$, які не впливають на зменшення цільової функції.

Для лінійної регресії, цільова функція якої квадратична і не потребує апроксимації, вектор ваг у регуляризованому випадку отримуємо за допомогою рівняння

$$w^p = (X^T X + \alpha I)^{-1} X^T y.$$

Дисперсія вхідних ознак описується діагональними елементами матриці $(X^T X + \alpha I)^{-1}$, збільшеними на α порівняно з нерегуляризованим варіантом. Тому в процесі регуляризації матриця входів X сприймається такою, яка має більшу дисперсію, і відповідно зменшуються ваги тих ознак, для яких коваріація з y незначна порівняно з доданою дисперсією [17].

Регуляризовану цільову функцію можна також визначати по-іншому у випадку, коли використовується штраф за нормою L^1, який визначається за формулою

$$\Omega\,(\theta) = \| w \|_1 = \sum_i | w_i |.$$

Тоді градієнт L^1 регуляризованої цільової функції за параметром w визначається за допомогою рівняння

$$\nabla_w \tilde{J}\,(w; X, y) = \alpha\, \text{sign}\,(w) + \nabla_w J\,(w; X, y)$$

і відрізняється від градієнта для норми L^2.

Розглянемо задачу лінійної регресії і наступні припущення, які полегшують подальший аналіз:

квадратична цільова функція апроксимується першими членами свого ряду Тейлора, так що $\nabla_w \hat{J}\,(w) = H\,(w - w^*)$;

гесіан — діагональна матриця $H = \text{diag}\,([H_{1,1}, ..., H_{n,n}])$, $H_{i,i} > 0,\ \forall i.$

Останнє припущення справедливо, якщо дані були попередньо оброблені для усунення кореляції між вхідними ознаками, наприклад з використанням методу головних компонент. Мінімальне значення цієї апроксимації L^1 регуляризованої цільової функції $\hat{J}$ досягається в точці [17]

$$w_i = \text{sign}\,(w_i^*)\,\max\left\{|\,w_i^*\,| - \frac{\alpha}{H_{i,i}}, 0\right\},\ \ \forall i = 1, ..., n,$$

де L^1-регуляризація дає більш розріджене рішення порівняно з L^2-регуляризацією (оптимальне значення деяких елементів вектора параметрів w дорівнює нулю). Це дійсне у наступних випадках: $w_i^* > 0$ і $w_i^* \le \alpha / H_{i,i}$ або $w_i^* < 0$ і $w_i^* \ge -\alpha / H_{i,i}$. Розрідженість рішення, властиву L^1-регуляризації, використано в задачах відбору ознак.

Побудова і аналіз моделей поширення COVID-19 в різних країнах світу. Побудову і навчання моделей часових рядів для опису поширення коронавірусу COVID-19, вибір найкращої моделі, побудову прогнозу поширення та візуалізацію результатів виконано у реалізованому програмному модулі в середовищі python з використанням сучасних бібліотек scikit-learn [21], pandas [22] та matplotlib [23]. Як вхідні дані розглянуто часові ряди з кількістю нових випадків заражень коронавірусом щоденно та сумарною кількістю випадків заражень для п’яти країн світу: США, Мексики, Бельгії, РФ та України (рис. 1). Ряди побудовано за офіційними даними [4—6].

Для опису вхідних рядів були побудовані альтернативні моделі поліноміальної регресії, опорних векторів та багатошарових нейронних мереж прямого розповсюдження. Оцінювання якості моделей здійснювалося на перевірочному наборі даних за наступними показниками:

помилки Root Mean Square Error (RMSE), Mean Absolute Error (MAE) та Mean Absolute Percentage Error (MAPE) —

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2},\ \ \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\,\hat{y}_i - y_i\,|,$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,\hat{y}_i - y_i\,|}{y_i}100\%;$$

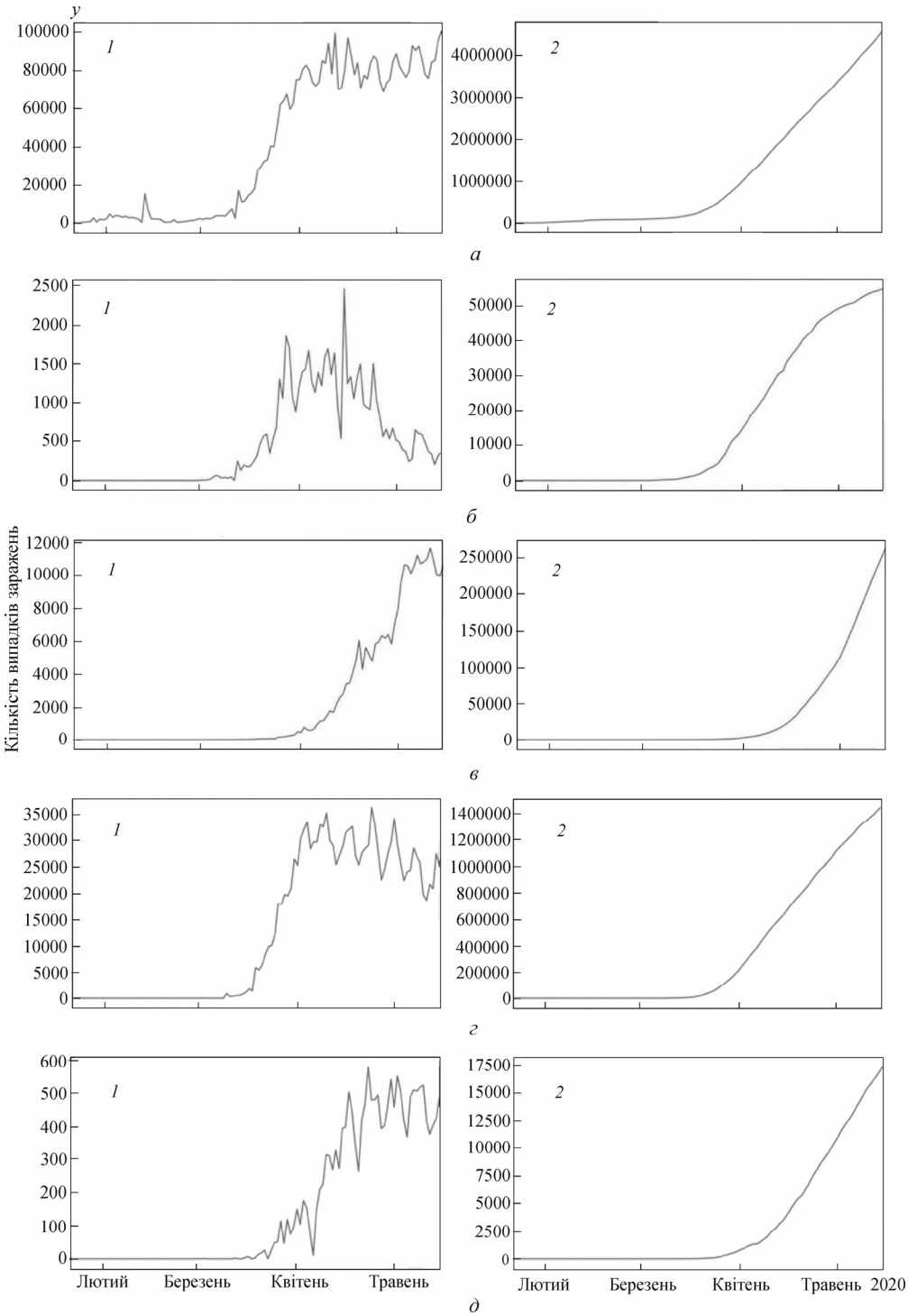


Рис. 1. Кількість нових випадків заражень щоденно (*1*) та сумарна кількість випадків заражень (*2*): *а* — у світі; *б* — в Бельгії; *в* — у РФ; *г* — в США; *д* — в Україні

коефіцієнт детермінації

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2},$$

який показує відсоток дисперсії, що пояснено моделлю. Тут вектор $\hat{y}$ містить значення прогнозу на основі моделі; y — вектор реальних значень вихідної змінної; N — розмірність перевірочного набору даних. Це основні показники, за якими порівнюються різні моделі прогнозування часових рядів [24].

Розглянемо результати навчання альтернативних моделей різних класів.

Модель багатошарової нейронної мережі прямого розповсюдження. Побудову і навчання цієї моделі для задачі регресії, що розглядається, виконано алгоритмом MLPRegressor бібліотеки scikit-learn: model = MLPRegressor (hidden_layer_sizes=[32, 26, 10,], max_iter=100000, alpha= 0.0005, random_state=26, solver='lbfgs', learning_rate='constant', validation_ fraction=0.1).

Варіювалися наступні параметри алгоритму MLPRegressor:

hidden_layer_sizes — розмір скритих шарів мережі;

solver — алгоритм оптимізації розрахунку ваг;

alpha — параметр регуляризації;

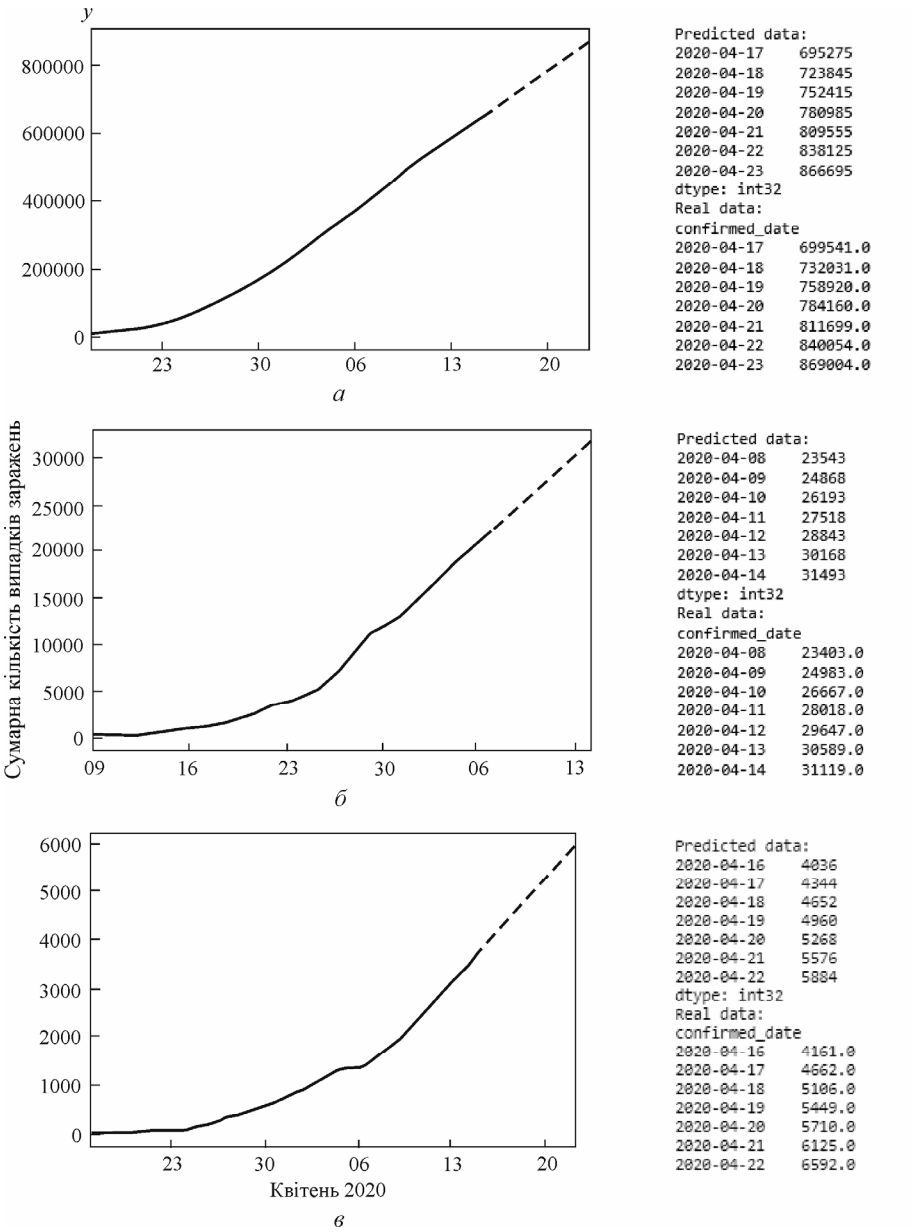
f — функція активації для скритого шару;

learning_rate — швидкість навчання;

max_iter — максимальна кількість ітерацій.

У таблиці наведено коефіцієнти детермінації R^2 для різних моделей нейронних мереж на навчальних даних при різних значеннях параметрів α і solver. Порівнюючи наведені у таблиці дані, можна зробити висновок, що найадекватнійшою моделлю багатошарового персептрона, яка описує

Країна	Значення R^2 для				
	α			solver	
	0,00001	0,001	0,1	«adam»	«lbfgs»
США	0,972	0,973	0,973	0,997	0,999
РФ	0,985	0,999	0,991	0,836	0,999
Бельгія	0,971	0,979	0,974	0,971	0,979
Мексика	0,673	0,896	0,789	0,536	0,824
Україна	0,705	0,941	0,873	0,813	0,941



а

б

в

Рис. 2. Прогноз приросту захворюваності на основі моделі багатошарового персептрона: *а* — для США (MAE = 4073,4, RMSE = 4648,4, MAPE=0,536%, R^2=0,999); *б* — для Бельгії (MAE =384,0, RMSE =1212,8, MAPE=1,420%, R^2=0,979); *в* — для України (MAE= =440,7, RMSE =434,4, MAPE=7,876%, R^2=0,941); суцільна крива — поточні значення сумарної кількості захворювань; шрихова крива — спрогнозовані значення сумарної кількості захворювань

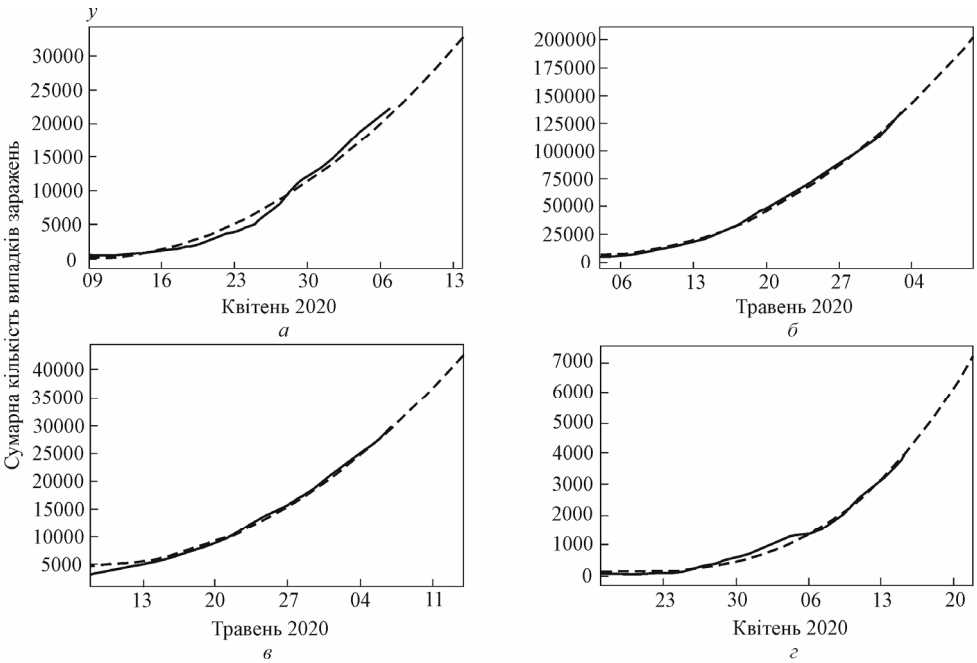


Рис. 3. Прогноз приросту кількості хворих на основі обраної моделі опорних векторів для Бельгії (*а*), Мексики (*б*), РФ (*в*), України (*г*): суцільна крива — поточні значення сумарної кількості захворювань; шрихова крива — спрогнозовані значення сумарної кількості захворювань

навчальні дані з коронавірусу COVID-19 для п'яти країн, є модель з параметром регуляризації α = 0,001 за умови використання алгоритму оптимізації Бройдена — Флетчера — Гольдфарба — Шанно (solver= «lbfgs») для розрахунку ваг мережі. За допомогою методу решітчастого пошуку з крос-валідацією [20] знайдено параметри моделі нейронних мереж та визначено архітектуру hidden_layer_sizes =[32, 26, 10].

Побудовану модель нейронної мережі застосовано для прогнозування приросту захворювань. Для оцінювання якості прогнози розраховувалися на перевірочних множинах даних з відомими значеннями реальних приростів захворювань. Прогнози на сім днів та оцінки їх якості побудовано у програмному середовищі python (рис. 2). Коефіцієнти детермінації мали значення, близькі до одиниці, помилки RMSE, MAE та MAPE для прогнозів на основі багатошарового персептрона мали менші значення порівняно з прогнозами на основі інших досліджених моделей.

Модель опорних векторів. Для побудови і навчання моделі опорних векторів використано алгоритм SVR бібліотеки scikit-learn: model = SVR (kernel='poly', degree=3, C=40, epsilon=0,1, shrinking=True, cache_size=200, verbose=False, tol=0,001, max_iter= - 1).

Зважаючи на особливості початкових даних, були побудовані моделі опорних векторів з поліноміальними ядрами. Варіювалися наступні параметри алгоритму SVR: degree — ступінь поліному; epsilon, *C* — параметри моделі, які задають ширину роздільної смуги. За допомогою методу решітчастого пошуку з крос-валідацією [20] знайдено найкращі комбінації параметрів для розглянутих навчальних даних:

для РФ — поліном третього ступеня, epsilon=0,1, C=10;

для Мексики і Бельгії — поліном другого ступеня, epsilon=0,1, C=35;

для США та України — поліном третього ступеня, epsilon=0,1, C=40.

На основі моделей з цими параметрами побудовано прогнози приросту захворюваності. На рис. 3 наведено побудовані у програмному середовищі python прогнози на сім днів для Бельгії, Мексики, РФ та України. Результати виявилися дещо гіршими порівняно з результатами на основі моделі нейронних мереж.

Порівняльний аналіз досліджених моделей показав, що приріст кількості хворих на коронавірус COVID-19 у Мексиці найкраще описується моделлю опорних векторів: model = SVR (kernel='poly', degree=3, C=30, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, tol=0.001, max_iter= - 1). Для США, РФ, Бельгії та України найкращою серед досліджених виявилася модель багатошарового персептрона: model = MLPRegressor (hidden_ layer_sizes=[32, 26, 10,], max_iter=100000, alpha=0.0005, random_state=26, solver='lbfgs', learning_rate='constant', validation_fraction=0.1)

Висновки

Продовжується зростання кількості захворювань на COVID-19 в Україні та інших країнах світу. Тому побудова математичної моделі поширення COVID-19 для прогнозування та прийняття рішень з метою зменшення впливу наслідків COVID-19 на здоров’я населення є наразі дуже актуальною. Модель багатошарового персептрона показала найвищу точність прогнозування на сім днів для США Бельгії, України та РФ. Оцінки якості пргнозів за цією моделлю наступні: MAPE=0,536%, R^2=0,999 (США), MAPE = 1,420%, R^2 = 0,979 (Бельгія), MAPE = 7,876%, R^2 = 0,941 (Україна). Для Мексики найкращою виявилася модель опорних векторів (MAPE = 0,623%, R^2 = 0,990). Отримані оцінки якості прогнозів свідчать про ефективність побудованої моделі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. World Health Organization. Coronavirus disease (COVID-19) outbreak situation. 2020. Режим доступу: https://www.who.int/emergencies/diseases/ novel-coronavirus-2019. (Accessed: 20.07.2020)
2. Міністерство охорони здоров'я України. Актуально про COVID-19. Режим доступу: https://moz.gov.ua/koronavirus-2019-ncov. (Accessed: 20.07.2020)
3. Коронавірус в Україні. Офіційний інформаційний портал Кабінету Міністрів України. Режим доступу: https://covid19.gov.ua/. (Accessed: 20.07.2020)
4. Всесвітня організація охорони здоров'я. Coronavirus disease (COVID-2019) situation reports. Режим доступу: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/ situation-reports (Accessed: 20.07.2020)
5. Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering at Johns Hopkins University. Режим доступу: https://coronavirus.jhu.edu/map.html. (Accessed: 20.07.2020)
6. Карта розповсюдження і смертності коронавірусу COVID-19 в світі. Режим доступу: https://www.currenttime.tv/a/covid-19-interactive-map/30484955.html (Accessed: 20.07. 2020)
7. *Fanelli D., Piazza F.* Analysis and forecast of COVID-19 spreading in China, Italy and France // Chaos, Solitons & Fractals, 2020, Vol. 134. Doi: 10.1016/j.chaos.2020.109761.
8. *Rainisch G., Undurraga E.A., Chowell G.* A dynamic modeling tool for estimating healthcare demand from the COVID19 epidemic and evaluating population-wide interventions// International Journal of Infectious Diseases, 2020, Vol. 96, pp. 376—383. Doi: 10.1016/j.ijid.2020.05.043.
9. *Sarkar K., Khajanchi S., Nieto J.J.* Modeling and forecasting the COVID-19 pandemic in India// Chaos, Solitons & Fractals, 2020, Vol. 139. Doi: 10.1016/j.chaos.2020.110049.
10. *Salgotra R., Gandomi M., Gandomi A.H.* Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming // Ibid, 2020, Vol. 138. Doi: 10.1016/j.chaos.2020.109945.
11. *Chintalapudi N., Battineni G., Sagaro G.G., Amenta F.* COVID-19 outbreak reproduction number estimations and forecasting in Marche, Italy// International Journal of Infectious Diseases, 2020, Vol. 96, pp. 327—333. Doi: 10.1016/j.ijid.2020.05.029.
12. *Singh S., Parmar K.S., Kumar J., Makkhan S.J.S.* Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19// Chaos, Solitons & Fractals, 2020, Vol. 135. Doi: 10.1016/j.chaos.2020.109866.
13. *Chakraborty T., Ghosh I.* Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis// Ibid, 2020, Vol. 135. Doi: 10.1016/j.chaos. 2020.109850.
14. *Yousaf M., Zahir S., Riaz M. et al*. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan// Ibid, 2020, Vol. 138. Doi: 10.1016/j.chaos.2020.109926.
15. *Aviv-Sharon E., Aharoni A.* Generalized logistic growth modeling of the COVID-19 pandemic in Asia// Infectious Disease Modelling. Available online 24 July 2020. Doi: 10.1016/j.idm.2020.07.003.
16. *Middelburg R.A., Rosendaal F.R.* COVID-19: How to make between-country comparison // International Journal of Infectious Diseases, 2020, Vol. 96, pp. 477—481. Doi: 10.1016/j.ijid.2020.05.066.
17. *Гудфеллоу Я., Бенджио И., Курвилль А*. Глубокое обучение / Пер. с англ. А.А. Слинкина. 2-е изд., испр. М.: ДМК Пресс, 2018, 652 с.

18. *Бринк Х., Ричардс Дж., Феверолф М.* Машинное обучение. СПб.: Питер, 2017, 336 с.
19. *Aurelien Geron*. Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media Inc., Sebastopol, CA, 2017, 760 с.
20. *Плас Дж.В*. Python для сложных задач. Наука о данных и машинное обучение. СПб.: Питер, 2018, 576 с.
21. Scikit-Learn Documentation. Режим доступу: https://scikit-learn.org. 2020. (Accessed: 20.07.2020)
22. Pandas Documentation. Режим доступу: https://pandas.pydata.org/docs/.(Accessed: 20.07.2020)
23. Matplotlib Documentation. Режим доступу: https://matplotlib.org/contents.html. (Accessed: 20.07.2020)
24. *Бідюк П.І., Романенко В.Д., Тимощук О.Л.* Аналіз часових рядів. Київ: Політехніка, 2012, 360 с.

Отримано 05.08.2020;
після доопрацювання 01.09.2020

REFERENCES

1. “World Health Organization, Coronavirus disease (COVID-19) outbreak situation”, available at: https://www.who.int/emergencies/diseases/ novel-coronavirus-2019 (Accessed July 20, 2020).
2. “The Ministry of Health of Ukraine. Actually about COVID-19”, available at: https://moz.gov.ua/koronavirus-2019-ncov (Accessed July 20, 2020).
3. “Coronavirus in Ukraine”, Official information portal of the Cabinet of Ministers of Ukraine, available at: https://covid19.gov.ua/ (Accessed July 20, 2020).
4. “World Health Organization. Coronavirus disease (COVID-2019) situation reports”, available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (Accessed July 20, 2020).
5. “Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering at Johns Hopkins University”, available at: https://coronavirus.jhu.edu/map.html (Accessed July 20, 2020).
6. “The map of spread and death from the coronavirus COVID-19 in the world”, available at: https://www.currenttime.tv/a/covid-19-interactive-map/30484955.html (Accessed July 20, 2020).
7. Fanelli, D. and Piazza, F. (2020), “Analysis and forecast of COVID-19 spreading in China, Italy and France”, *Chaos, Solitons & Fractals*, Vol. 134. DOI: 10.1016/j.chaos.2020.109761.
8. Rainisch, G., Undurraga, E.A. and Chowell, G. (2020), “A dynamic modeling tool for estimating healthcare demand from the COVID19 epidemic and evaluating population-wide interventions”, *International Journal of Infectious Diseases*, Vol. 96, pp. 376-383. DOI: 10.1016/j.ijid.2020.05.043.
9. Sarkar, K., Khajanchi, S. and Nieto, J.J. (2020), “Modeling and forecasting the COVID-19 pandemic in India”, *Chaos, Solitons & Fractals*, Vol. 139. DOI: 10.1016/j.chaos.2020.110049.
10. Salgotra, R., Gandomi, M. and Gandomi, A.H. (2020), “Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming”, *Chaos, Solitons & Fractals*, Vol. 138. DOI: 10.1016/j.chaos.2020.109945.

11. Chintalapudi, N., Battineni, G., Sagaro, G.G. and Amenta, F. (2020), "COVID-19 outbreak reproduction number estimations and forecasting in Marche, Italy", *International Journal of Infectious Diseases*, Vol. 96, pp. 327-333. DOI: 10.1016/j.ijid.2020.05.029.
12. Singh, S., Parmar, K.S., Kumar, J., Makkhan, S.J. (2020), "Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19", *Chaos, Solitons & Fractals*, Vol. 135. DOI: 10.1016/j.chaos.2020.109866.
13. Chakraborty, T., Ghosh, I. (2020), "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis", *Chaos, Solitons & Fractals*, Vol. 135. DOI: 10.1016/j.chaos.2020.109850.
14. Yousaf, M., Zahir, S. and Riaz, M. (2020), "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan", *Chaos, Solitons & Fractals*, Vol. 138. DOI: 10.1016/j.chaos.2020.109926.
15. Aviv-Sharon, E. and Aharoni, A. (2020), "Generalized logistic growth modeling of the COVID-19 pandemic in Asia", *Infectious Disease Modelling*. DOI: 10.1016/j.idm.2020.07.003.
16. Middelburg, R.A. and Rosendaal, F.R. (2020), "COVID-19: How to make between-country comparison", *International Journal of Infectious Diseases*, Vol. 96, pp. 477-481. DOI: 10.1016/j.ijid.2020.05.066.
17. Gudfellou, YA., Bendzhio, I. and Kurvill', A. (2017), Deep Learning, The MIT Press, Massachusetts, England.
18. Brink, KH., Richards, Dzh., Feverolf, M. (2016), Real-World Machine Learning, Manning Publications.
19. Aurelien, G. (2017), Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media Inc.
20. Plas, J.V. (2017), Python Data Science Handbook. Essential Tools for Working with Data, O'Reilly Media Inc.
21. "Scikit-Learn Documentation", available at: https://scikit-learn.org (Accessed July 20, 2020).
22. "Pandas Documentation", available at: https://pandas.pydata.org/docs/ (Accessed July 20, 2020).
23. "Matplotlib Documentation", available at: https://matplotlib.org/contents.html (Accessed July 20, 2020).
24. Bidyuk, P.I., Romanenko, V.D. and Tymoshchuk, O.L. (2012), *Analiz chasovykh ryadiv* [Time Series Analysis], Politechnika, Kiev, Ukraine.

Received 05.08.2020;
After revision 01.09.2020

N.I. Nedashkovskaya, S.O. Lupanenko

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS
FOR FORECASTING COVID-19 SPREADING IN DIFFERENT COUNTRIES

In this work, mathematical models of the spread of the coronavirus COVID-19 in various countries are built, and a comparative analysis of these models for the United States, Mexico, Russia, Belgium and Ukraine was performed. Baseline data on the number of infections obtained from the daily reports of the World Health Organization and the the Center for Systems Science and Engineering at Johns Hopkins University. To simulate the spread of coronavirus, two powerful classes of machine learning methods have been selected that allow predicting nonlinear time series: support vector machines and feedforward multilayer neural networks. The advantages and disadvantages of these methods are revealed, and the issues of regulariza-

tion are considered. The construction and training of time series models to describe the spread of COVID-19 in different countries, the choice of the best model, the construction of forecast and the visualization of results were performed in an implemented software module in the python environment using modern scikit-learn, pandas and matplotlib libraries. Using the grid search method with cross-validation, the best parameters of neural network and support vector models which describe the spread of COVID-19 in the USA, Mexico, Russia, Belgium and Ukraine were selected. Based on the constructed models, the growth of COVID-19 diseases in these countries was predicted.

K e y w o r d s: support vector machines, multilayer feedforward neural networks, regularization, COVID-19, forecasting of epidemic spreading.

НЕДАШКІВСЬКА Надія Іванівна, д-р техн. наук, доцент кафедри ММСА Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського», який закінчила у 2004 р. Область наукових досліджень — системи і методи підтримки прийняття рішень, багатокритеріальний аналіз, аналіз ризиків, системний аналіз, машинне навчання, інтелектуальний аналіз даних, моделювання.

ЛУПАНЕНКО Софія Олександрівна, студентка Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського». Область наукових досліджень — машинне навчання, інтелектуальний аналіз даних, моделювання.