

АНОТАЦІЯ

Бердник М. І. Метод L_1 -регуляризації для опису фізико-хімічних властивостей молекул. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття ступеня доктора філософії за спеціальністю 102 - Хімія (Галузь знань 10 – Природничі науки). – Харківський національний університет імені В. Н. Каразіна Міністерства освіти і науки України, Харків, 2021.

Роботу присвячено дослідженню можливостей використання L_1 -регуляризації в побудові хеометричних моделей «структура-активність» і квантовохімічних розрахунках. Для виконання завдань дисертації розроблено оригінальний комплекс програм, що реалізують різні статистичні (хеометричні) підходи до побудови регресійних моделей й аналізу їх прогностичної здатності. Також створено комплекс квантовохімічних програм, у яких L_1 -регуляризація використовується для побудови хвильових функцій методів, що ураховують електронну кореляцію.

Зокрема, у дисертаційній роботі розглядалося використання L_1 -регуляризації для побудови лінійних емпіричних моделей опису різних фізико-хімічних параметрів молекул. Серед таких параметрів розглянуто pK_a та температури кипіння органічних сполук різної природи, які включають карбонові кислоти, феноли, сульфіді, флуороалкани. Розглядалися також кореляції в'язкості рідин та тиску насиченого пару різних органічних сполук. Спираючись на досліджені вибірки молекул, було показано, що з використанням L_1 -регуляризації завжди можна сформулювати послідовний (упорядкований) набір дескрипторів. Систематично додаючи дескриптори з цього набору до моделей лінійної регресії або штучних нейронних мереж, можна отримати рівняння (або відповідно нейронні мережі) з послідовно зростаючими величинами критеріїв валідації. Оскільки після ранжування дескрипторного набору обрані предиктори можуть використовуватися в різних підходах до побудови лінійної регресії, нами було проведено відповідне дослідження якості цих альтернативних моделей. При цьому розглядалися:

метод найменших квадратів (*Ordinary Least squares*, OLS), метод найменших модулів (*Least Absolute Deviation*, LAD), метод ортогональних відстаней (*Orthogonal Distances Regression*, ODR), а також запропонований нещодавно метод найменших абсолютних відхилень ортогональних відстаней (*Least Absolute Deviation of Orthogonal Distances*, LADOD). Було показано, що той чи інший метод може мати кращі прогностичні властивості відповідно до критеріїв зовнішньої або внутрішньої валідації. Показано, що методом штучних нейронних мереж з використанням впорядкованого дескрипторного набору, який був отриманий методом L_1 -регуляризації, також може бути зроблено якісні прогнози властивостей речовини. Також було проведено співставлення отриманих рівнянь лінійної регресії з альтернативними підходами, що працюють із нескороченими (неоптимізованими) дескрипторними наборами. А саме: з методом PCR (*Principal Component Regression*), а також методом PLS (*Partial Least Squares* або *Projection on Latent Structure*). Слід зазначити, що хоча з використанням цих методів для деяких задач і були отримані досить надійні прогностичні моделі, але такі моделі не надають ясної інформації стосовно природи отриманих рівнянь і не відповідають на питання: які структурно-хімічні особливості або молекулярні дескриптори, призводять до змін у відгуку (активності). У вивчених прикладах L_1 -регуляризація дозволила сформулювати компактні одно-, двух- або трьох- параметричні моделі, які здатні задовільно описати набір даних. Відповідно до вивчених прикладів, моделі отримані з попереднім відбором із використанням LARS-LASSO виявились кращими, ніж результати розрахунків PLS та PCR.

Певну увагу в дисертації приділено методам валідації й оцінкам якості регресійних рівнянь. З цією метою було використано модельну задачу, у яку вносилися похибки як в залежну, так і в незалежну змінні. Для полегшення аналізу, а також, щоб вивчити валідаційні характеристики рівнянь в усіх досліджених методах лінійної регресії, розглядався найпростіший, але далеко нетривіальний випадок – регресія з однією незалежною змінною. Така постановка задачі дала можливість оцінювати рівняння відповідно до

близькості коефіцієнтів регресійних рівнянь до «ідеальних» теоретичних значень. З використанням модельної задачі було досліджено вплив раціонального розбиття вибірки на тренувальну та тестову на якість отриманих регресійних рівнянь. Було продемонстровано, що випадкове одиничне розбиття вибірки не є інформативним, оскільки в залежності від систем, що опинилися в тестовій вибірці, валідаційні характеристики для початкової (повної) вибірки можуть бути як дуже погані, що веде до недооцінки, так і дуже добрі, що веде до переоцінки якості рівняння. Отже, показано, що для адекватної оцінки регресійного рівняння, а також дослідження якості вхідних даних у цілому, необхідно створювати та вивчати якомога більше розбивань на тренувальну й тестову вибірку. Також було досліджено вплив урахування границь застосовності моделі (*Applicability Domain*, AD) на валідаційні характеристики регресійних рівнянь. Встановлено, що при випадкових розбиваннях вибірки на тренувальну та тестову, максимумами розподілу внутрішніх та зовнішніх характеристик, зазвичай, співпадають. На відміну від цього, коли розбиття на тренувальну й тестову вибірки здійснюються таким чином, що тестова вибірка знаходиться в AD моделі, отриманої виходячи з тренувальної вибірки, відповідний розподіл зовнішніх критеріїв валідації зміщується відносно внутрішніх у сторону збільшення. При цьому найбільш інформативними є розбиття, що попадають близько до максимуму густини точок, оскільки саме з аналізу цих областей можна отримати найбільш повне адекватне розуміння якості моделі. Також було досліджено відомі, запропоновані на сьогодні, критерії валідації. Виходячи з модельної задачі, було зроблено висновок, що деякі з критеріїв валідації надто сильно корельовані один з одним, що робить їх одночасне використання малоінформативним. Серед таких параметрів пари: (R_{test}^2 – CCC та Q_{F3}^2 – RMSEP_{test}). Встановлено, що для даних із вираженим розкидом типовою картиною є зворотна (суттєво нелінійна) залежність R_{train}^2 – R_{test}^2 . При цьому покращення (збільшення) коефіцієнтів внутрішньої валідації (R_{train}^2 , Q_{LOO}^2), взагалі кажучи, не є свідомством покращення прогностичної

властивості моделі, оскільки для лінійної регресії за достатньо великої кількості точок залежність між цими двома критеріями була близька до лінійної. Проте, критерій Q_{Loo}^2 може бути успішно використано для малих вибірок. Спираючись на розрахункові дані, показано, що для більшості випадків метод OLS давав найкращі результати. Однак, для великих вибірок із похибкою як в залежній, так і в незалежних змінних, у методі ODR (та LADOD) можна отримати найкращі рівняння.

Інша тісно пов'язана із побудовою статистичних моделей проблема - це побудова класифікаційної функції. З цією метою в роботі використано L_1 -регуляризований розрахунок логістичної регресії. Розглянуто дві задачі. У першій проведено класифікацію молекул на сильні та слабкі основи відносно іону літію. У другій задачі органічні системи були класифіковані на активні або неактивні відповідно до спорідненості зв'язування молекул до рецепторів естрогену. Показано, що з використанням L_1 -регуляризованої логістичної регресії можна досягнути таких результатів класифікації, які є конкурентно-спроможними до результатів, отриманих з використанням інших, більш складних у розрахунковому сенсі, методів. Використання спеціального L_1 -регуляризованого алгоритму (його позначено як LR-LARS-LASSO) дало можливість отримати досить прості класифікаційні рівняння, які є інтерпретуємими (на відміну від результатів, отриманих в інших популярних методах класифікації, таких як: метод опорних векторів, метод випадкових лісів, метод штучних нейронних мереж). Також отримані рівняння логістичної регресії є однозначними й відтворюваними.

Показано, що метод L_1 -регуляризації може бути використаний і в квантовій хімії. За допомогою процедури L_1 -регуляризації можливо створення впорядкованого (ранжованого) набору електронно-збуджених відносно Гартрі-Фоківського стану конфігурацій. Включаючи різну кількість конфігурацій з створеного набору, можливо отримати прогресивний набір наближених розв'язків до точних даних методу. Метод реалізовано в рамках теорії збурень Меллера-Плессета другого порядку (MP2) та різних рівнів теорії зв'язаних

кластерів. Продemonстровано, що такі наближені розв'язки дають доволі точні значення енергетичних характеристик молекул, при цьому кількість конфігурацій у розрахунках може бути значно нижчою, ніж у розрахунках з використанням повного конфігураційного набору точного методу. Для ефективного розв'язку відповідних рівнянь теорії зв'язаних кластерів, реалізовано низку розрахункових алгоритмів з використанням багатокрокових методів першого порядку.

Ключові слова: L_1 -регуляризація, QSAR/QSPR, pK_a органічних сполук, температури кипіння органічних сполук, в'язкість рідини, тиск насиченого пару, ліганди рецептору естрогену, основність до катіону літію, лінійна регресія, метод найменших квадратів, метод найменших модулів, метод ортогональних відстаней, штучні нейронні мережі, валідація, логістична регресія, теорія збурень Меллера-Плессета, теорія зв'язаних кластерів.