УДК 004.94

ПОРІВНЯННЯ ВЛАСТИВОСТЕЙ АЛГОРИТМІВ LASSO ТА ПЕРЕБІРНОГО КОРЕЛЯЦІЙНОГО АЛГОРИТМУ МГУА З РОЗРАХУНКОМ РЕЙТИНГУ РЕГРЕСОРІВ CRA

Г.А. Піднебесна

Міжнародний науково-навчальний центр
інформаційних технологій та систем НАН та МОН України, м. Київ

pidnebesna@irtc.org.ua

У роботі наводяться результати порівняння властивостей алгоритмів *LASSO* та перебірного кореляційного алгоритму МГУА з розрахунком рейтингів регресорів (кореляційно-рейтинговий алгоритм) *CRA*. Порівняння базується на аналізі результатів числових експериментів та застосуванні таких статистичних характеристик, як індекс Жаккара, чутливість та специфічність. Для дослідження бралась різна кількість регресорів (100, 500, 1000) та моделі різної складності.

Ключові слова: кореляційно-рейтинговий алгоритм CRA, МГУА, індуктивне моделювання, перебірні алгоритми, кореляційний аналіз.

The paper presents the results of comparing the properties of LASSO algorithms and the GMDH sorting-out correlation algorithm with the calculation of regressor's ratings (correlation- rating algorithm) CRA . The comparison is based on the analysis of the results of numerical experiments and the application of such statistical characteristics as Jacquard index, sensitivity and specificity. Different numbers of regressors (100, 500, 1000) and models of different complexity were used for the study.

Keywords: correlation- rating algorithm CRA, GMDH, searching-out algorithms, inductive modeling, correlation analysis.

В работе показаны результаты сравнения свойств алгоритмов *LASSO* и корреляционного алгоритма направленного перебора МГУА с учетом рейтинга регрессоров (корреляционно-рейтинговый алгоритм) *CRA*. Сравнение основано на анализе результатов числовых экспериментов с применением статистических характеристик, таких как индекс Жаккара, чувствительность и специфичности.

Ключевые слова: корреляционно-рейтинговый алгоритм CRA, МГУА, индуктивное моделирование, переборные алгоритмы, корреляционный анализ.

Вступ

У сучасному світі, де аналіз даних є однією з найважливіших задач, використання лінійної регресії залишається одним з найпотужніших інструментів для роботи з даними. Множинна лінійна регресія широко

використовується при моделюванні та прогнозуванні різних процесів. Для підвищення її ефективності природним вважається визначення інформативних аргументів та виключення з розгляду незначимих (неінформативних). Можна визначити декілька причин для проведення такої селекції.

По-перше, згідно принципу леза Оккама вважається, що за умови існування декількох ймовірних описів процесу (об'єкту, явища) кращим є найпростіше. Тобто, модель без надлишкових аргументів краще пояснює дані. До того ж, непотрібні фактори додають шум до оцінки впливу на процес інших аргументів.

По-друге, якщо в даних є колінеарність (взаємозалежність), то це означає, що кілька змінних повторно визначають вплив тих самих факторів на залежну змінну. Ще одним важливим аргументом на користь відсіювання з моделі неінформативних факторів є економічність: немає потреби вимірювати та/або розраховувати зайві фактори, що може істотно економити час розрахунків (це може бути критичним в задачах моніторингу процесів в реальному часі) та кошти для отримання даних спостережень.

Існує багато методів обробки даних з використання регресії. Метод групового урахування аргументів (МГУА) відомий як ефективний засіб роботи із статистичними даними [1–3]. Основним його принципом є автоматична процедура побудови множини моделей-кандидатів та вибору кращих за зовнішніми критеріями селекції. Одним з найвідоміших алгоритмів МГУА є метод повного перебору *COMBI*. Його ефективність було підтверджено шляхом багатьох досліджень та впроваджень на практиці. Але метод має певні вади. Зокрема, обмежений кількістю аргументів, що беруться до розгляду. Крім того, оскільки метод використовує зовнішній критерій, вагомим є розбиття вибірки даних спостережень на навчальну (тренувальну) та перевірну (тестувальну). Від такого розбиття може значно залежати результат моделювання.

В [4] запропоновано метод відбору інформативних аргументів за допомогою аналізу рейтингу регресорів, отриманого моделюванням при багаторазовому поділі вибірки – кореляційно-рейтинговий алгоритм (*correlation- rating algorithm*) *CRA*. Метод відноситься до класу перебірних алгоритмів МГУА з неповним (направленим) перебором. Крім традиційної для МГУА процедури побудови моделей-кандидатів остаточний вибір інформативних аргументів відбувається теж автоматично, за допомогою процедури кластеризації, що надає більше об'єктивності результату.

В роботі досліджуються властивості запропонованого алгоритму. Було проведено низку числових експериментів: змінювалась кількість істинних аргументів, їхня частка щодо загальної кількості; розглядались результати моделювання при різній кількості точок вимірювань.

1. Основна ідея перебірного кореляційного алгоритму МГУА з розрахунком рейтингу регресорів *CRA*

Для оцінки інформативності аргументів аналізується їхній рейтинг, який розраховується певним чином. На початку моделювання рейтинг всіх аргументів нульовий. Для отримання рейтингових балів аргументів багаторазово проводиться випадковий поділ вибірки на тренувальну та перевірну. Для кожного такого поділу будується множина часткових моделей: додається один *перспективний* аргумент до попередньої моделі. До *перспективних* вибираються аргументи згідно величини їхньої кореляції з вихідною змінною (для першої моделі) або з поточними залишками Y_r (різницею табличного значення змінної Y та модельного значення Y_{mod}: $Y_r = Y - Y_{mod}$). Одночасно обчислюється ймовірність того, що ця величина кореляції є статистично значущою (*p-Val*). Додається до моделі найбільш корельований регресор (з найменшим значенням *p-Val*).

З множини отриманих часткових моделей традиційно для МГУА вибирається краща згідно з мінімальним значенням критерія регулярності. Ті регресори, що увійшли до кращої моделі, отримують по одному рейтинговому балу.

Описана процедура поділу вибірки даних, згідно з яким проводиться моделювання, повторюється задану кількість разів (в даному дослідженні – 30). В результаті ми отримуємо рейтинг регресорів. Він показує, скільки разів кожен з аргументів потрапляв до складу кращих моделей для різних поділів набору даних. Вибір аргументів для остаточної моделі проводиться шляхом аналізу отриманого рейтингу. Такий аналіз може бути проведено по-різному: визначено експертом чи запроваджено автоматичну процедуру з застосуванням методів кластеризації або частотних критеріїв. В розглянутому дослідженні використовувалась автоматична процедура кластеризації за допомогою стандартної функції *k-means* в *MatLab R2018b*. До остаточної моделі вибираються регресори з кластера, для якого середнє значення є більшим. Коефіцієнти остаточної моделі обраховуюся за МНК на всій вибірці даних.

2. Порівняння результатів числових експериментів

Для дослідження властивостей алгоритму розглянемо результати моделювання пропонованим алгоритмом *CRA* та алгоритмом *LASSO* [5]. Для порівняння проведено низку експериментів з різною кількістю регресорів *m*, точок вимірювань у вибірці *n* та числом істинних аргументів *mt* (таких, що входять до істинної моделі).

Для того, щоб оцінити, наскільки ефективно запропонований алгоритм знаходить істині аргументи при побудові моделі, було застосовано індекс

Жаккара *IJ* [7], а також такі статистичні характеристики бінарної класифікації, як чутливість (*sensitivity – sns*)) і специфічність (specificity – *spc*) [8].

3.1 Порівняння індексу Жаккара числових експериментів

Індекс Жаккара (*the Jaccard index*) - бінарна міра подібності скінченних множин, яка визначається як міра спільної частини, поділена на міру об'єднання множин:

$$J(O,E)=\frac{|O\cap E|}{|O\cup E|}=\frac{|O\cap E|}{|O|+|E|-|O\cap E|},\ 0\le J\le 1$$

Подібність множин тим краща, чим ближче показник індексу до одиниці.

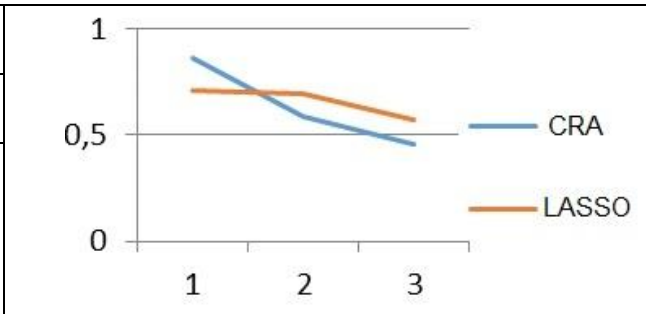
В описаному випадку порівнюються множини істинних табличних регресорів (*O*) та тих, що увійшли в модель (*E*), вибраних алгоритмом.

Залежність індексу Жаккара від частки істинних аргументів щодо загальної кількості регресорів для різної кількості точок вимірювань

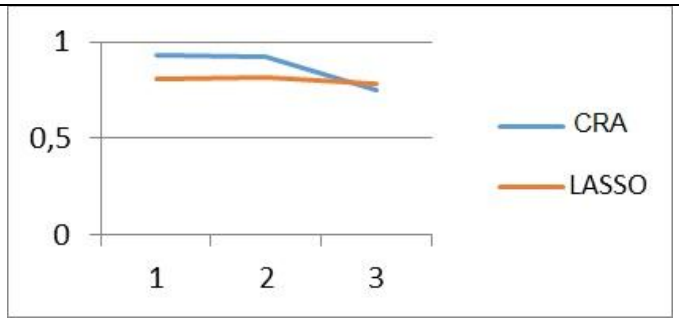
Було проведено низку експериментів, коли при однаковій загальній кількості аргументів *m* змінювались кількість істинних регресорів *mt*, а також кількість точок вимірювань *n* (Таблиця 1).

Таблиця 1. Середнє значення індексу Жаккара при збільшенні частки істиних аргументів *mt*

m =50	*n* =100	
	CRA	*LASSO*
mt =5	0,86	0,71
mt =15	0.59	0.69
mt =25	0.46	0.57

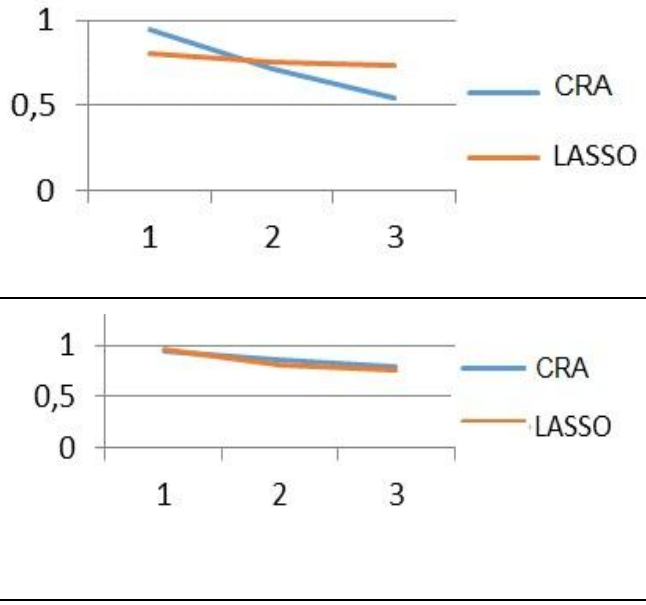


	n =500	
	CRA	*LASSO*
mt =5	0.93	0.81
mt =15	0.92	0.82
mt =25	0.75	0.78



Продовження таблиці 1

m =100	*n* =500	
	CRA	*LASSO*
mt =5	0.95	0.81
mt =15	0.72	0.76
mt =25	0.54	0.74
	n =1000	
	CRA	*LASSO*
mt =5	0.94	0.96
mt =15	0.86	0.81
mt =25	0.8	0.76



За результатами проведених експериментів можна зробити припущення, що алгоритм *LASSO* більш стійкий до збільшення частки істинних аргументів щодо їх загальної кількості.

Залежність індексу Жаккара від кількості точок вимірювань

Досліджувалась залежність результату моделювання від кількості точок вимірювань шляхом проведення серії експериментів при різних пропорціях істиних аргументів щодо загальної кількості регресорів. Значення індексу Жаккара, розраховані в ході експериментів, подано в таблиці 1 та візулізовано відповідними графіками.

Проведені експерименти показують, що індекс Жаккара запропонованого алгоритму *CRA* істотно покращується із збільшенням кількості точок вимірювань. Значення індексу Жаккара для алгоритму *LASSO* збільшується повільніше. Тобто при збільшення спостережень за допомогою *CRA* селекція інформативних та неінформативних агрументів відбувається точніше.

Таблиця 2. Середнє значення індексу Жаккара при збільшенні точок вимірювань *n*

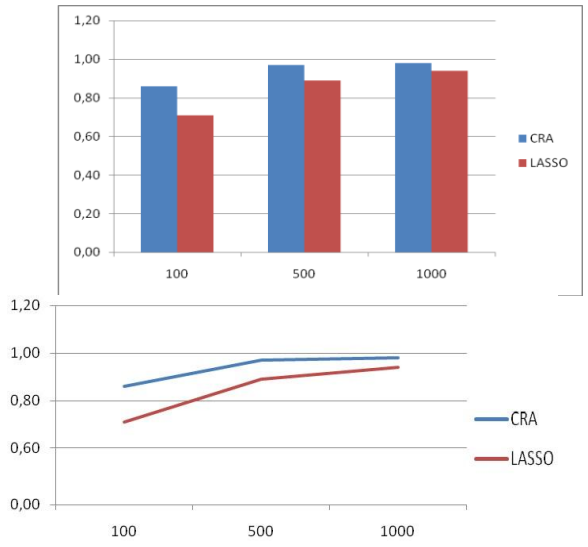
m =10, *mt* =5

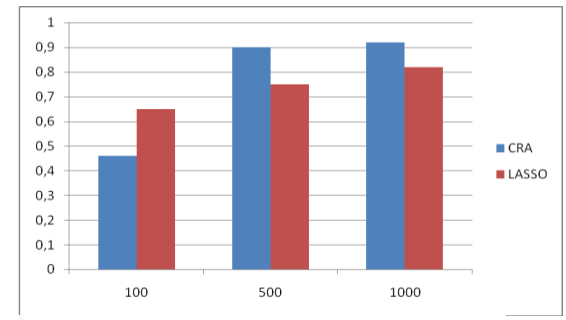
№	*n*	*CRA*	*LASSO*
1	100	0.86	0.71
2	500	0.97	0.89
3	1000	0.98	0.94

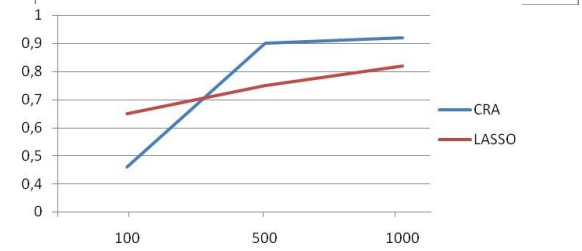
m =50, *mt* =15

№	*n*	*CRA*	*LASSO*
1	100	0.46	0.65
2	500	0.9	0.75
3	1000	0,92	0,82

Продовження таблиці 2

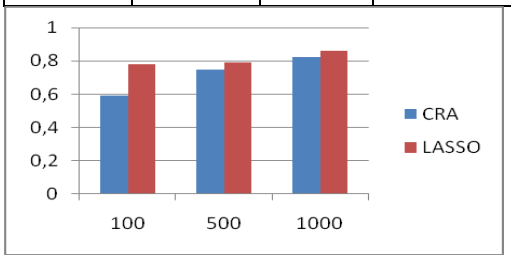


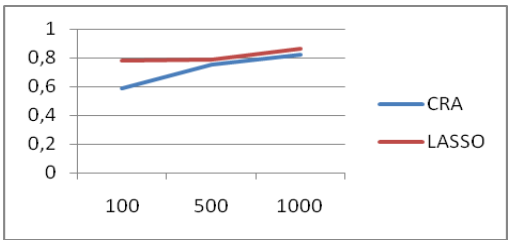




m =50, *mt* =25

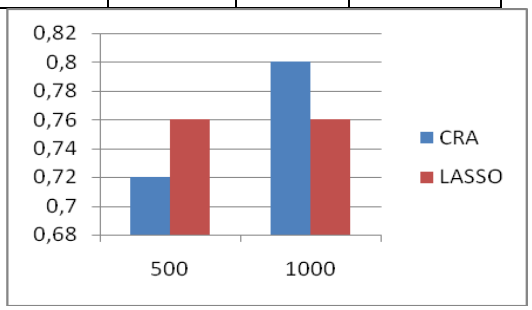
№	*n*	*CRA*	*LASSO*
1	100	0,59	0,78
2	500	0,75	0,79
3	1000	0,82	0,86

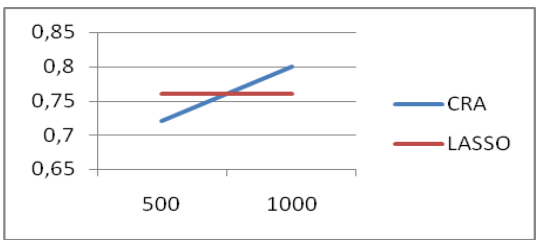




m =100, *mt* =25

№	*n*	*CRA*	*LASSO*
1	500	0,72	0,76
2	1000	0,8	0,76





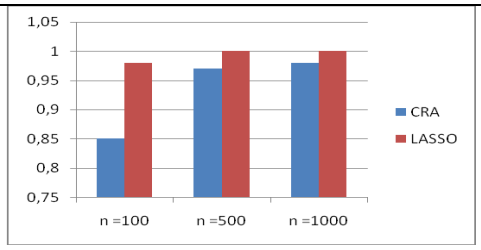
3.2 Порівняння чутливості досліджуваних алгоритмів

Статистична характеристика *чутливості* (яка в деяких галузях також називається справжньою позитивною швидкістю (*true positive rate*) або ймовірність виявлення (*probability of detection*)) вимірює частку фактичних позитивів, які правильно визначені як такі. Тобто, кількість винайдених істинних аргументів, які увійшли до побудованої моделі.

Значення *чутливості*, розраховані в ході експериментів з різними параметрами, подано в таблицях 2–4 та візулізовано відповідними графіками.

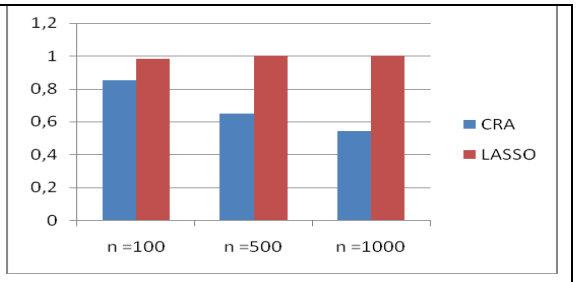
Таблиця 2. Значення чутливості для *m* =10, *mt* =5

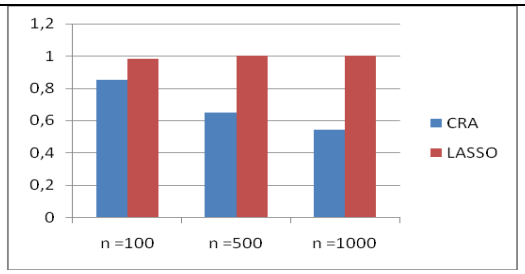
№	m =10, mt =5	CRA	LASSO
1	n =100	0,85	0,98
2	n =500	0,97	1
3	n =1000	0,98	1



Таблиця 3. Значення чутливості для *m* =50

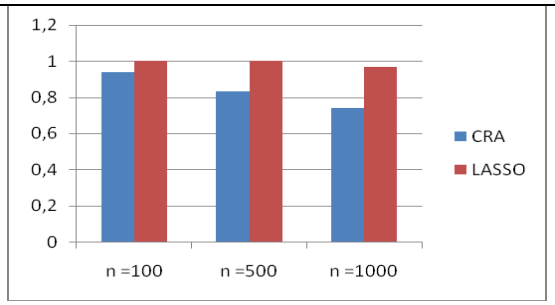
m =50	*n* =100	
	CRA	*LASSO*
mt =5	0,85	0,98
mt =15	0,65	1
mt =25	0,54	1
	n =500	
	CRA	*LASSO*
mt =5	0,9	0,98
mt =15	0,85	1
mt =25	0,78	1

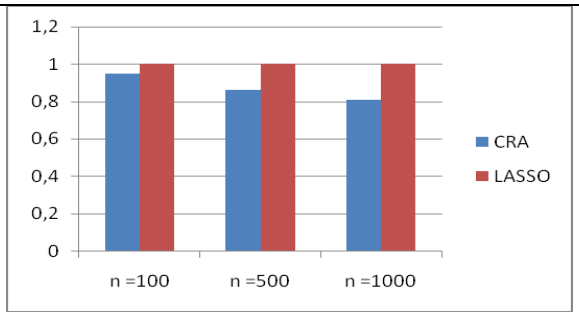




Таблиця 4. Значення чутливості для *m* =100

m =100	*n* =100	
	CRA	*LASSO*
mt =5	0,94	1
mt =15	0,83	1
mt =25	0,74	0,97
m =100	*n* =500	
	CRA	*LASSO*
mt =5	0,95	1
mt =15	0,86	1
mt =25	0,81	1





Проведені експерименти показують, що характеристика *чутливості* запропонованого алгоритму *CRA* дещо поступається алгоритму *LASSO*.

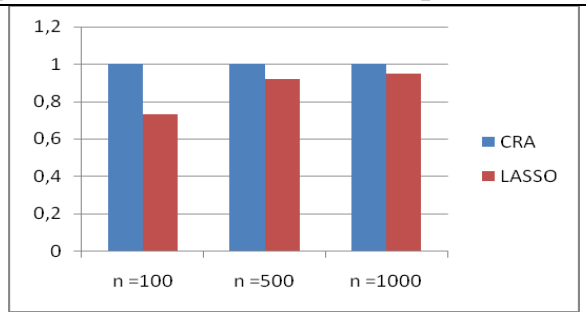
3.3 Порівняння специфічності алгоритмів *CRA* та *LASSO*

Специфічність (яка також називається справжнім негативним показником - *true negative rate*) вимірює частку фактичних негативів, які правильно ідентифіковані як такі (тобто, тих регресорів, які не повинні входити в модель, «неінформативних»). Зауважимо, що терміни «позитиви» і «негативи» означають присутність або відсутність екземпляра (аргумента) у множині.

Значення *специфічності*, розраховані в ході експериментів з різними параметрами, подано в таблицях 5–7 та візулізовано відповідними графіками.

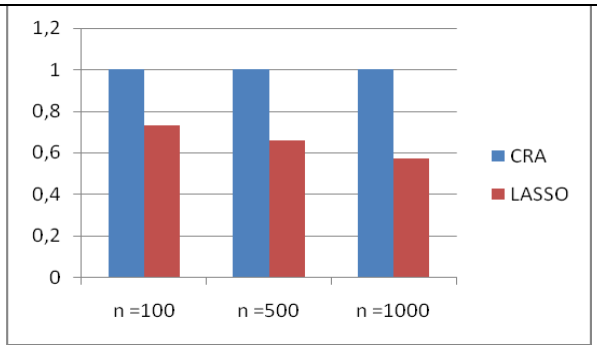
Таблиця 5. Значення чутливості для різної кількості точок вимірювань

m =10, mt =5	*CRA*	*LASSO*
n =100	1	0,73
n =500	1	0,92
n =1000	1	0,95



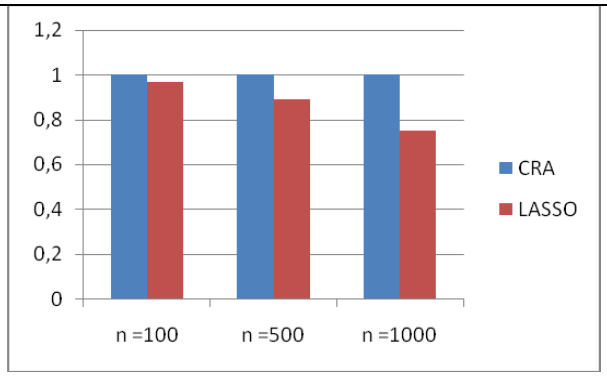
Таблиця 6. Значення чутливості для *m* =50 при різних *mt* та *n*

m =50	*n* =100	
	CRA	*LASSO*
mt =5	1	0,73
mt =15	1	0,66
mt =25	1	0,57



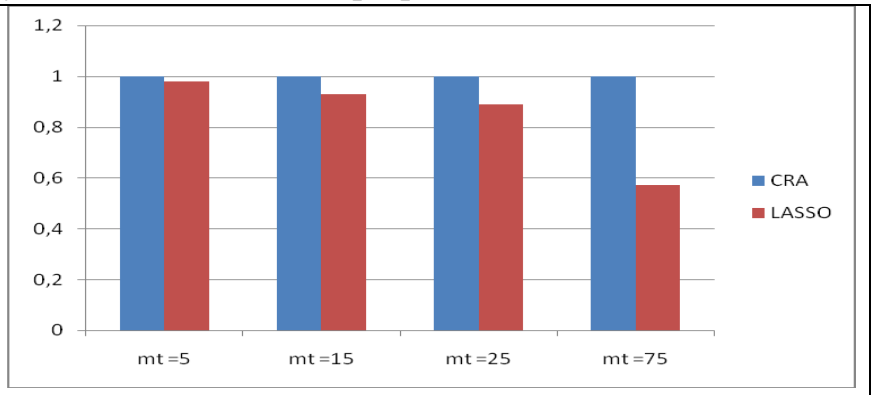
Продовження таблиці 6

m =50	*n* =500	
	CRA	*LASSO*
mt =5	1	0,97
mt =15	1	0,89
mt =25	1	0,75

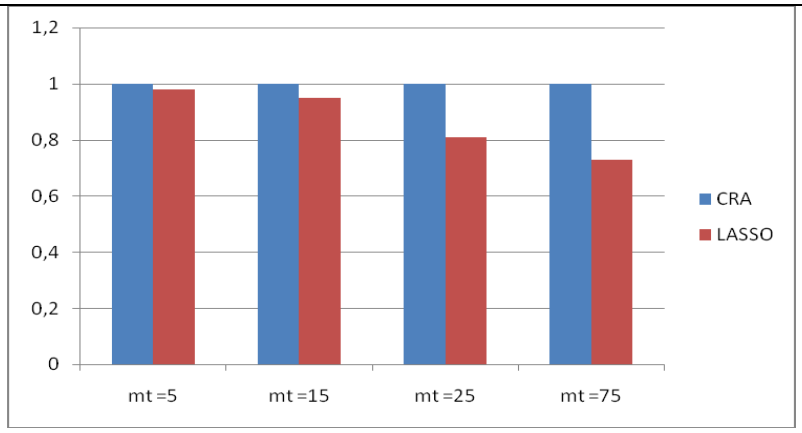


Таблиця 7. Значення чутливості для m =50 при різних mt та n

m =100	n =500	
	CRA	*LASSO*
mt =5	1	0,98
mt =15	1	0,93
mt =25	1	0,89
mt =75	1	0,57



m =100	n =1000	
	CRA	*LASSO*
mt =5	1	0.98
mt =15	1	0,95
mt =25	1	0,81
mt =75	1	0,73



Проведені експерименти показують, що запропонований алгоритм *CRA* має показник *специфічності spc* = 1. Це означає, що неінформативні аргументи метод відкидає. Значення *специфічності* для алгоритму *LASSO* є дещо гіршим, тобто в модель, побудовану цим методом, потрапляють зайві аргументи.

Висновки

В роботі було досліджено властивості кореляційного алгоритму з розрахунком рейтингу регресорів *CRA*, який відноситься до перебірних алгоритмів МГУА. Запропоновано метод побудови по статистичній вибірці даних оптимальної (у сенсі мінімуму зовнішнього критерію) лінійної моделі з найменшою можливою кількістю регресорів. Алгоритм базується на застосуванні парної кореляції та процедури кластеризації для відбору інформативних регресорів, підрахунку та аналізі їхнього рейтингу.

Для оцінки ефективності роботи запропонованого алгоритму було проведено низку числових експериментів зі зміною кількості істиних аргументів, їхньої частки щодо загальної кількості регресорів. Проведено моделювання при різній кількості точок вимірювань. Розглянуто такі статистичні показники моделювання як індекс Жаккара, чутливість та

специфічність. Проведено порівняння з популярним сучасним методом моделювання *LASSO*.

Результати числових експериментів показали, що обидва досліджуваних метода показали високі показники індексу Жаккара, тобто добре виокремлюють інформативні та неінформативні аргументи із загальної множини. Показник чутливості (міра правильно визначених істиних аргументів) у *LASSO* дещо краща. Проте характеристика специфічності алгоритму *CRA* показує, що він на відміну від *LASSO* не включає до складу моделі зайвих, неінформативних аргументів. Це дає підстави вважати перспективними подальші дослідження методу *CRA* з застосуванням різних процедур розбиття вибірки, способів визначення параметрів, критеріїв відбору кращої моделі, тощо.

Література

[1] Степашко В.С., Єфіменко С.М., Савченко Є.А. Комп'ютерний експеримент в індуктивному моделюванні. – Київ: Наукова думка, 2014. – 222 с.

[2] V. Stepashko, "Developments and Prospects of GMDH-Based Inductive Modeling", In: Advances in Intelligent Systems and Computing II. CSIT 2017 / Shakhovska N., Stepashko V. (eds), AISC series, vol. 689, Cham: Springer, 2018, pp. 474-491.

[3] Ivakhnenko A.G., Ivakhnenko G.A., Savchenko E.A., and Wunsch D. Problems of Further Development of GMDH Algorithms: Part 2 // Pattern Recognition and Image Analysis, Vol. 12, № 1, 2002, pp. 6-18.

[4] Г.А. Піднебесна, «Опис методу побудови лінійної моделі на основі аналізу парних кореляцій та рейтингу регресорів». Індуктивне моделювання складних систем. Збірник наук. праць. К.:МННЦІТС, 2018, С. 108-115

[5] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B (Methodological), 1996, vol. 58, no. 1, pp. 267-288.

[6] Gower, J.C. "Measures of Similarity, Dissimilarity and Distance." John Wiley & Sons, Inc. 2004 (https://doi.org/10.1002/0471667196.ess1595).

[7] P. Willett, J.M. Barnard, and G.M. Downs, "Chemical similarity searching." Journal of Chemical Information and Computer Sciences, 38(6): pp. 983–996, 1998.