УДК 004.8

М. І. Кривошея

МІНІМАКСНІ ПІДХОДИ ДЛЯ ВІДБОРУ ОЗНАК

Вінницький національний технічний університет

Анотація

У роботі розглядаються мінімаксні підходи до відбору ознак у задачах інтелектуального аналізу даних. Проаналізовано основні існуючі методи відбору ознак, зокрема filter-, wrapper- та embedded-підходи, а також визначено їхні обмеження в умовах шуму, структурної неоднорідності та зсуву розподілу даних. Основну увагу приділено мінімаксній постановці задачі відбору ознак, у якій оптимізується не середній ризик моделі, а максимальна похибка на множині сценаріїв. Показано, що такий підхід може забезпечити підвищення робастності моделі, зменшення впливу спуріозних ознак і покращення якості в найгірших випадках.

Ключові слова: *відбір ознак, мінімаксна апроксимація, інтелектуальний аналіз даних, найгірший випадок, оптимізація*

Abstract

The paper considers minimax approaches to feature selection in data mining tasks. The main existing feature selection methods, including filter, wrapper, and embedded approaches, are analyzed, and their limitations under noise, structural heterogeneity, and distribution shift are identified. Special attention is paid to the minimax formulation of the feature selection problem, where not the average model risk but the maximum error over a set of scenarios is optimized. It is shown that such an approach can improve model robustness, reduce the impact of spurious features, and enhance performance in worst-case situations.

Keywords: *feature selection, minimax approximation, data mining, worst-case optimization*

Вступ

У сучасних задачах інтелектуального аналізу даних та машинного навчання одним із ключових етапів побудови ефективних моделей є відбір ознак (feature selection). У багатьох прикладних галузях, зокрема в біоінформатиці, фінансовому аналізі, системах кібербезпеки та обробці соціальних даних, кількість доступних ознак може сягати сотень або тисяч. При цьому значна частина таких ознак є шумовою, надлишковою або статистично залежною, що негативно впливає на якість моделі та її узагальнювальну здатність.

Основною метою відбору ознак є формування компактної підмножини інформативних ознак, яка забезпечує високу точність прогнозування, зменшення обчислювальної складності алгоритмів та підвищення інтерпретованості моделей. У літературі виділяють три основні групи методів відбору ознак: filter-методи, wrapper-методи та embedded-методи. Filter-методи оцінюють інформативність ознак незалежно від конкретної моделі, wrapper-методи здійснюють пошук оптимальної підмножини ознак шляхом багаторазового навчання моделі, а embedded-методи виконують відбір ознак безпосередньо в процесі навчання моделі.

Попри значну кількість досліджень у цій галузі, більшість існуючих підходів орієнтована на оптимізацію середніх показників якості, наприклад середньої помилки або середньої точності моделі. Такий підхід є ефективним у стандартних умовах, однак може виявитися недостатньо надійним у реальних задачах, де дані часто характеризуються високим рівнем шуму, структурною неоднорідністю та можливими зсувами розподілу. У таких випадках модель, оптимізована за середнім критерієм, може демонструвати прийнятну середню якість, але значно гірші результати на окремих підвибірках або групах даних.

Одним із перспективних напрямів подолання цієї проблеми є використання мінімаксних підходів до відбору ознак. Основна ідея мінімаксної оптимізації полягає в мінімізації максимальної можливої похибки моделі на множині можливих сценаріїв. Такий підхід дозволяє орієнтувати процес відбору ознак не лише на покращення середньої якості моделі, а й на забезпечення її стійкості в найгірших випадках.

Таким чином, застосування мінімаксних підходів у задачах відбору ознак є актуальним напрямом досліджень, спрямованим на підвищення робастності моделей інтелектуального аналізу даних. Метою даної роботи є дослідження можливостей застосування мінімаксних критеріїв для формування стійких підмножин ознак та порівняння їх ефективності з класичними методами відбору ознак.

Результати дослідження.

У межах дослідження було створено експериментальне середовище для перевірки ефективності мінімаксного підходу до відбору ознак у задачах інтелектуального аналізу даних. Основною ідеєю експерименту було порівняння класичних алгоритмів відбору ознак, які оптимізують середній ризик моделі, з мінімаксним підходом, що орієнтується на мінімізацію максимальної похибки моделі у різних сценаріях.

Для тестування було використано синтетичний набір даних з контрольованими параметрами. Генерувалася вибірка, що містила 4000 об'єктів і 200 ознак, з яких 10 були інформативними, а решта 190 – шумовими. Така структура даних дозволяє моделювати реальні задачі інтелектуального аналізу даних, оскільк в них присутня значна частина ознак не містить корисної інформації.

Задача відбору ознак було сформульовано таким чином. Нехай задано вибірку

$$D = \{(x_i, y_i)\}_{i=1}^{N}\ x \in R^m$$

де

x_i - вектор ознак

y_i - цільова змінна

m - загальна кількість ознак

Необхідно знайти підмножину ознак

$$S \subseteq \{1, \dots, m\}, |S\} = k$$

яка забезпечує найкращу якість моделей.

У класичній постановці задача відбору ознак формулюється як мінімізація середнього ризику

$$S = arg \min_{|S|=k} R(S)$$

де

$$R(S) = \frac{1}{N}\sum_{i=1}^{N} L(f_s(x_i), y_i)$$

L - функція витрат

f_s - модель, побудована на підмножині ознак

Було використано мінімаксну постановку задачі

$$S = arg \min_{|S|=k} \max_{r=1,\dots,M} R_r(s)$$

де

$R_r(s)$ - ризик моделі для r- го сценарію

M - кількість сценаріїв або підгруп

Такий підхід дозволяє мінімізувати максимально можливу помилку моделі та підвищити її надійність.

Узагальнений критерій оптимізації має вигляд

$$J(S) = \max_{r=1,\dots,M} R_r(s) + \lambda|S| + \gamma \sum_{i<j} p_{ij}^2$$

де

λ- штраф за кількість ознак

γ- штраф за надлишковість ознак

p_{ij}- коефіцієнт кореляції між ознаками

Для оцінки результатів були використані такі метрики оцінювання

Accuracy

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

де

TP – істинно позитивні класифікації

TN – істинно негативні класифікації

FP – хибно позитивні класифікації

FN – хибно негативні класифікації

Precision

$$P = \frac{TP}{TP + FP}$$

де
P – точність позитивних передбачень

Recall

$$R = \frac{TP}{TP + FN}$$

де
R- повнота

F1-score

$$F1 = \frac{2PR}{P + R}$$

де
F1-score - гармонійне середнє precision і recall.
Worst-group accuracy

$$WGA = \min_{r=1,\dots,M} Acc_r$$

Acc_r- точність моделі на r-й групі даних

У дослідженні були використані такі алгоритми:

- Mutual Information
- Random Forest importance
- LASSO
- запропонований мінімакс-алгоритм

Метод	**Accuracy**	**F1**	**Worst-group accuracy**
Mutual Information	0.812	0.808	0.741
Random Forest importance	0.807	0.804	0.748
LASSO	0.804	0.801	0.753
Мінімакс-підхід	0.798	0.795	**0.781**

Отримані результати показують, що класичні алгоритми забезпечують дещо вищу середню точність моделі. Однак запропонований мінімакс-підхід демонструє найкращий результат за метрикою worst-group accuracy, що свідчить про більш стабільну роботу моделі у складних сценаріях.

Важливим фактором є співвідношення між кількості відібраних ознак k та кількістю інформативних ознак m_{inf}

Залежність Accuracy від кількості відібраних ознак k

(k)	**Mutual Information**	**Random Forest**	**LASSO**	**Minimax**
5	0.7100	0.7293	0.7293	0.6664
10	0.7007	0.7543	0.7543	0.7579
15	0.7007	0.7500	0.7521	0.7557
20	0.7050	0.7529	0.7500	0.7550
30	0.7279	0.7543	0.7421	0.7600

Залежність worst-group accuracy від кількості відібраних ознак k

(k)	**Mutual Information**	**Random Forest**	**LASSO**	**Minimax**
5	0.6856	0.7060	0.7060	0.6332

10	0.6769	0.7540	0.7540	0.7540
15	0.6681	0.7482	0.7391	0.7496
20	0.6812	0.7349	0.7419	0.7496
30	0.7118	0.7377	0.7394	0.7540

В вибірці, яка розглядалась $m_{inf} = 10$

Випадок $k < m_{inf}$

алгоритми фізично не можуть включити всі корисні ознаки. Через це модель втрачає частину важливої інформації. Це добре видно для $k = 5$: усі методи працюють гірше, а мінімаксна стратегія виявляється особливо чутливою, бо намагається збалансувати якість між групами ще до того, як у модель потрапив повний набір інформативних ознак.

Випадок $k \approx m_{inf}$

Коли кількість відібраних ознак наближається до кількості інформативних, алгоритми отримують можливість включити більшість корисних ознак без значного домішування шумових. Це зазвичай є найкращою зоною компромісу між повнотою інформації та стійкістю.

У нашому експерименті саме біля $k = 10$мінімаксна стратегія починає різко покращувати результат і досягає такого ж або трохи кращого рівня worst-group accuracy, ніж класичні підходи.

Випадок $k > m_{inf}$

у модель починають потрапляти додаткові, частково шумові або надлишкові ознаки. Для класичних методів це часто означає:

- зростання ризику перенавчання;
- більшу чутливість до спуріозних кореляцій;
- нестабільність на окремих групах.

Для мінімаксного підходу додаткові ознаки теж не завжди корисні, але він зазвичай краще контролює найгірший сценарій, оскільки підбір ведеться з орієнтацією на

$$\max_r R_r(S).$$

У наведеному експерименті при $k = 15,20,30$саме мінімаксна стратегія демонструє найстабільніші значення worst-group accuracy.

Висновки

У роботі було досліджено застосування мінімаксного підходу до задачі відбору ознак у системах інтелектуального аналізу даних. Проведений аналіз показав, що більшість класичних алгоритмів відбору ознак оптимізують середню якість моделі та можуть бути чутливими до шуму і структурної неоднорідності даних.

Запропонований мінімаксний підхід базується на мінімізації максимальної помилки моделі в різних сценаріях і дозволяє підвищити робастність моделей машинного навчання. Проведене експериментальне дослідження показало, що хоча середня точність моделей, отриманих за допомогою мінімаксної стратегії, може бути дещо нижчою порівняно з класичними алгоритмами, цей підхід забезпечує найкращі результати за метрикою worst-group accuracy.

Також встановлено, що кількість відібраних ознак істотно впливає на якість роботи алгоритмів. Найкращі результати досягаються, коли кількість відібраних ознак близька до кількості інформативних. Перевищення цього значення призводить до включення шумових ознак та зниження стабільності моделі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Swaroop M., Krishnamurti T., Wilder B. Distributionally Robust Feature Selection // arXiv preprint. [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/2510.21113 (дата звернення: 12.03.2026).
2. Quinzan F., Soleymani A., Jaillet P., Bauer S. Doubly Robust Causal Feature Selection // arXiv preprint. [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/2306.07024 (дата звернення: 12.03.2026).
3. Ahadzadeh B., Abdar M., Safara F., Khosravi A., Menhaj M., Suganthan P. SFE: A Simple, Fast and Efficient Feature Selection Algorithm for High-Dimensional Data // arXiv preprint. [Електронний ресурс] – Режим доступу:

https://arxiv.org/abs/2303.10182 (дата звернення: 12.03.2026).
4. Kuhn D., Esfahani P. Distributionally Robust Optimization: A Review // Acta Numerica. 2025. Vol. 34. [Електронний ресурс] – Режим доступу: https://www.cambridge.org/core/journals/acta-numerica/article/distributionally-robust-optimization/5B4E65E3A5A2AEF24E218A6B34E6EAA2 (дата звернення: 12.03.2026).

Кривошея Михайло Ігорович — аспірант кафедри Автоматизації та інтелектуальних інформаційних технологій, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: mishakryvoshea@gmail.com
Науковий керівник: ***Квєтний Роман Наумович*** – доктор технічних наук, професор, професор кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: rkvetny@vntu.edu.ua

Kryvosheia Mykhailo I. — Postgraduate student at the Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, email : mishakryvoshea@gmail.com
Supervisor: ***Kvyetnyy Roman N.***– Doctor of Technical Sciences, Professor, Professor of the Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: rkvetny@vntu.edu.ua