УДК 004.415.25

Соловей О.Л., канд. техн. наук, доцент
Київський національний університет будівництва і архітектури, solovey.ol@knuba.edu.ua
Соловей Б.А., магістр
Національний технічний університет України «КПІ ім. Ігоря Сікорського», bsolovei25@gmail.com

ВКЛЮЧЕННЯ ІНФОРМАЦІЙНОГО КРИТЕРІЮ БАЄСА В МАТЕМАТИЧНУ МОДЕЛЬ ВИБОРУ ОПТИМАЛЬНОГО НАБОРУ ОЗНАК

На сьогодні є добре задокументованим той факт, що модель машинного навчання, яка побудована на підмножині ознак, а не повному їх наборі має кращі властивості. В даній роботі визначено недолік математичної моделі методу обгортки (wrapper) та запропоновано зміни, з метою усунення визначеного недоліку. Задача вибору оптимальної підмножини ознак за методом обгортки для моделей машинного навчання, що належать до класу «навчання з вчителем» формулюється наступним чином. Задано:

1) набір даних який складається з ознак $X = \{x_1..x_k\}$ та набір «вірних» значень ознаки-класу $y = \{y_1..y_n\}$, де *k, n* – кількість ознак та спостережень в заданому наборі даних.

2) набір методів відповідно до яких проходить визначення можливих альтернатив наборів ознак де $l = |P|, o = |F|$.

3) набір критеріїв $C = \{c_1..c_q\}, q = |C|$ згідно з якими проходить оцінка альтернатив з набору *F*.

Розв'язок задачі полягає у:

1) визначенні альтернатив $f \in F$, виконанням методів *P* з набором ознак *X* та ознаки-класу *y*, тобто $\{P, X, y\} -> \{f, F\}$.

2) проведені оцінки альтернатив з набору *F* на заданій множині критеріїв *C(f)* і вибору оптимального варіанту f_{opt}.

Математична модель такої задачі описується виразом (1)

$$f_{opt} = \arg opt\{C(f)\} \tag{1}$$

Для моделей, що належать до класу навчання з вчителем набір *C* складається з метрик якості побудованої моделі: точність (Accuracy), влучність (Precision), повнота (Recall), площа під ROC кривою (AUC), MCC-міра, F-міра (F1 score, Fb score), коефіцієнт Жаккара (Jaccard score), при цьому найчастіше точність має найвищій ваговий коефіцієнт, при цьому модель (1) описується виразом (2):

$$f_{opt} = \arg\max\{C(f)\} \tag{2}$$

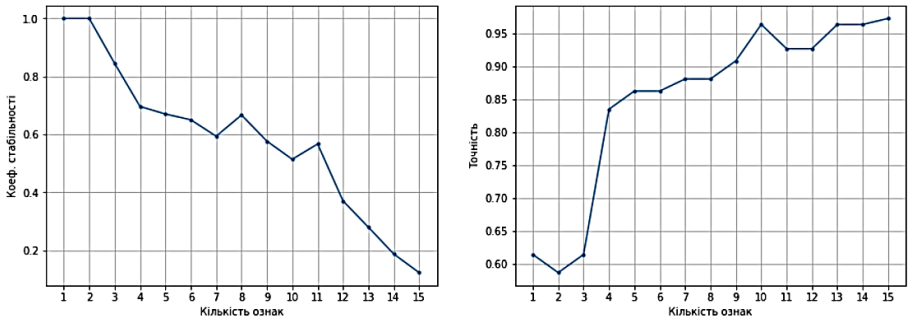
Відповідно з (2), для кожної альтернативи $f \in F$ будується модель і обчислюється точність її прогнозу. Оптимальною вважається підмножина ознак при якій точність моделі приймає максимальне значення.

В роботі [2] було визначено, що результат моделі (2) не є досить надійним через те, що оцінка точності моделі залежить від того, як вихідний набір даних поділено на дані для навчання та дані для тестування моделі. Ненадійність обумовлюється варіативністю оцінки точності. Для усунення зазначеного недоліку, запропоновано додатковий критерій – коефіцієнт стабільності I_s альтернативи $f \in F$, який включено до нової групи набору критеріїв *G(F)*, яка оцінює $f \in F$, ще до етапу побудови моделі машинного навчання. Відповідну математичну модель було формалізовано системою (3)

$$\begin{cases} D = \{d \mid 1 \le d < |G(F) > \theta|, G(F) > \theta \\ f_{opt} = \arg\max\{C(D)\} \end{cases} \tag{3}$$

За результатами експериментальних тестів в роботі [2] визначено, що обчислювати критерії групи *C* має сенс тільки для *d* з коефіцієнтом стабільності $I_s \geq 0.5$ ($\theta \geq 0.5$). Таким чином математичній моделі (3) дозволяє зменшити плив варіативності оцінки точності на результати оптимального вибору підмножини ознак.

Поза увагою в математичній моделі (3) залишено, той факт, що точність моделі класифікації є величина прямо пропорційна кількості ознак *k*, які включено в модель, таким чином, відповідно (3) перевага буде надана альтернативам з *D*, для яких справедливо: 1) $I_s \geq 0.5$; 2) $k->|X|$. На рисунку 1 для набору даних congressEW, який описано в [1] нерівність (1) моделі (3) виконується при $k \in [1..11]$, а умова (2) - при *k=10* тоді, за моделлю (3) буде обрана оптимальна підмножина, яка включатиме 10 із можливих 15 ознак.



а) б)

Рис. 1 – а) коефіцієнт стабільності набору ознак залежно від її кількості та значенню; б) точність моделі для кожної альтернативи $f \in F$

Але, чим більше ознак включено в модель, тим складнішою та залежною від спостережуваних даних стає модель, тобто може стати «перенавченою». Така модель описує не існуючі в наборі взаємозв'язки та має слабку здатність прогнозувати «клас» на даних для тестування [3]. Таким чином, математична модель (3) має бути доповнена критерієм, яких обмежить складність моделі. Таким критерієм може бути Баєсівський інформаційний критерій $BIC_d = -2\log L_d + n\log k$, де $\log L_d$ – максимізоване значення функції правдоподібності моделі *m*, побудованої на підмножині ознак $d \in D$; *k* – кількість ознак які включені в *d*; *n* – кількість спостережень в заданому наборі даних[4]. Чим складніше *m(d)*, тим більше значення BIC_d, його мінімальне значення визначає просту модель не схильну до перенавчання. Математична модель (3) вибору оптимального набору ознак з урахуванням BIC_d прийме вигляд (4)

$$\begin{cases} D = \{d \mid 1 \leq d <| G(F) > \theta |, G(F) > \theta \\ M = \{m \mid 1 \leq m <| D |, C(m(d)) > \rho \\ d_{opt} = \arg\min\{ BIC(M(D))\} \end{cases} \quad (4)$$

Список посилань

1. Brown G. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection / G. Brown, A. Pocock, MJ. Zhao, M. Luján //. The journal of machine learning research. – 2012. – Issue 13. – pp. 27-66.

2. Kuncheva L. A stability index for feature selection. /L. Kuncheva // In Artificial intelligence and applications. – 2007. – Issue 12. – pp. 421-427.

3. Harrell, F. E. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis /F. E. Harrell//, Springer, New York, 2001. – 582 p.

4. van de Schoot R. Bayesian statistics and modelling /R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens // Nature Reviews Methods Primers. – 2021. – Issue 14.