УДК 519.1

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ МЕТОДУ ГІЛОК І ГРАНИЦЬ ДЛЯ РОЗ'ВЯЗАННЯ ЗАДАЧІ ВИБОРУ ОПТИМАЛЬНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ

І.М. Мельник, Г.А. Піднебесна

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОНМС України, Київ, пр-т Академіка Глушкова, 40

ivanmelnyk@ukr.net, pidnebesna@mail.ru

Запропоновано метод гілок і границь для розв'язання задачі дискретної оптимізації з метою вибору оптимальної регресійної моделі з використанням принципу дихотомії для розбиття поточної множини розв'язків задачі на підмножини розгалуження. Вибір підмножини розв'язків для процесу розгалуження здійснюється за стохастичною процедурою
Ключові слова: метод гілок і границь, вибір оптимальної регресійної моделі

It is proposed the method of branch and bound for solving discrete optimization problem for selecting the optimal regressive model with using the principle of dichotomy for branching the current set to subsets and selecting a subset of solutions for branching process by random procedure.
Keywords: branches and borders method, the problem of choosing optimal regressive model

Предложен метод ветвей и границ для решения задачи дискретной оптимизации с целью выбора оптимальной регрессионной модели. Для ветвления используется принцип дихотомии. Выбор подмножества решений для процесса ветвления осуществляется с помощью стохастической процедуры.
Ключевые слова: метод ветвей и границ, выбор оптимальной регрессионной модели

Вступ. Для багатьох задач оцінки, аналізу та прогнозування в різних сферах діяльності людини певні величини, які характеризують конкретні об'єкти досліджень, мають дискретну природу. З математичної точки зору постановки таких задач виявляються задачами дискретної та комбінаторної оптимізації. До такого класу задач відноситься вибір оптимальної регресійної моделі. Цей клас задач оптимізації має свою специфіку і потребує математичного апарату, спрямованого на те, щоб зменшити трудомісткість алгоритмічного процесу розв'язання таких задач. В основі комбінаторних методів лежить перебір можливих варіантів розв'язків поставленої задачі. Вони характеризуються певною послідовністю перебору можливих варіантів та правилами виключення, що дають змогу ще в процесі розв'язування задачі виявити неоптимальні варіанти без попередньої їх перевірки. Одним з ефективних методів розв'язання задач дискретної та комбінаторної оптимізації є алгоритмічна схема методу гілок і границь, яка була вперше запропонована Лендом і Дойгом [1] в 1960 році для загальної задачі цілочисельного лінійного програмування.

Загальна ідея методу гілок і границь полягає в послідовному конструюванні та переборі тих варіантів рішень задачі, які, згідно з відповідними критеріями, є «перспективними» для подальшого розвитку, і відсіву тих варіантів, які завідомо є «неперспективними». Цей обчислювальний процес продовжується доти, доки не буде знайдено оптимальний варіант розв'язку задачі.

Можна виділити три основних алгоритмічних складових методу гілок і границь:

- отримання оцінки множини, яка розглядається на певному кроці алгоритму,
- визначення за певним критерієм вибору «перспективного» напрямку обчислень при розгалуженні,
- процедура розгалуження – поділ поточної «перспективної» множини на підмножини для подальшого обчислення.

Від конкретного вибору кожної з цих складових залежить ефективність застосування методу для конкретних класів задач.

Опис запропонованого алгоритму

В основі алгоритмічної схеми лежить ідея послідовного розбиття поточної множини допустимих розв'язків на підмножини (підмножини розгалуження). На кожному кроці елементи розбиття (тобто підмножини розв'язків) піддаються перевірці для з'ясування, чи може дана підмножина містити оптимальний розв'язок. Перевірка здійснюється за допомогою обчислення значення нижньої оцінки цільової функції (оцінки знизу для задач мінімізації) на цій підмножині розв'язків і співставлення значення оцінки зі значенням рекорду на цей момент.

Нижня оцінка цільової функції на даній підмножині розв'язків – це таке дійсне число, яке явно не більше (менше) будь-якого значення цільової функції на цій підмножині. *Рекордом* називаємо значення цільової функції задачі для найкращого на даний момент зі знайдених розв'язків. Якщо нижня оцінка цільової функції на даній підмножині розв'язків не менше (більше) рекорду, то вважаємо, що ця підмножина може бути відкинута з подальшого розгляду, оскільки явно не містить оптимального розв'язка. Підмножина розв'язків, яка перевіряється, може бути відкинута ще й у тому випадку, коли на певному кроці знайдено розв'язок, для якого значення нижньої оцінки цільової функції на цій підмножині краще.

Якщо значення цільової функції для знайденого розв'язку менше раніше обчисленого рекорду, то значення рекорду змінюється на це значення. Якщо на якомусь кроці вдається відкинути всі елементи розбиття (підмножини розв'язків) , то значення рекорду — це оптимальне значення розв'язку початкової задачі. В іншому випадку з невідкинутих підмножин розв'язків обирається одне з перспективних (наприклад, з найменшою нижньою оцінкою

цільової функції), і воно розбивається на підмножини розгалуження. Ці нові підмножини розв'язків знову перевіряються і т.д. доти, поки на певному кроці не буде знайдено таке значення рекорду, яке буде меншим (не більшим) значень нижніх оцінок цільової функції на всіх підмножинах розгалуження. По закінченні цього обчислювального процесу поточне значення рекорду і є оптимальним значенням цільової функції, а відповідний розв'язок є оптимальним розв'язком початкової задачі.

Розглянемо докладніше. Нехай розглядається складна система, яка характеризується m вхідними (незалежними) змінними $\{x_1, ..., x_i, ..., x_m\}$ та однією вихідною (залежною) змінною y, що мають стохастичний характер і задані вибіркою з n статистичних спостережень цих змінних $S = \{1, ..., n\}$.

Нехай $I \subseteq MX = \{1,...,m\}$ – довільна підмножина індексів незалежних змінних (регресорів). Регресійна модель на цій підмножині вхідних (незалежних) змінних для вихідної (залежної) змінної y будується в лінійному вигляді [2]:

$$y = \sum_{i \in I} a_i x_i \ .$$

Значення функціоналу оцінки регресійної моделі $F(I)$ обчислюється так:

$$F(I) = \max_{s \in S} (y^s - \sum_{i \in I} a_i\ (I)\ x_i^s)^2 , \qquad (1)$$

де $I = \{i \mid x_i \in X_1\}$ – підмножина з множини номерів незалежних змінних MX, X_1 – вибрана підмножина вхідних змінних, на яких будується регресійна модель,

$a_i(I), i \in I,$ - коефіцієнти регресійної моделі, знайдені за МНК для набору регресорів X_1.

Необхідно знайти таку підмножину I^*, $I^* \subseteq MX$, яка би мінімізувала значення функціоналу (1):

$$F(I^*) = \min \max_{l \in S} (y^l - \sum_{i \in I} a_i\ (I)\ x_i^l)^2 , \qquad (2)$$

де мінімум беретеся за всіма підмножинами I з множини MX.

Будемо розв'язувати загальну задачу як послідовність m часткових задач, тобто виконувати відповідну декомпозицію. Кожна k-та ($k = 1,...,m$) часткова задача полягає в пошуку методом гілок і границь найкращого (субоптимального) розв'язку на відповідній підмножині з усієї множини розв'язків задачі вибору оптимальної регресійної моделі як задачі дискретної оптимізації.

Підмножина номерів змінних k-ї часткової задачі задається так:

$$I_k = \{k, k+1, ..., m\}, \qquad I_k \subseteq MX = \{1, ..., m\}.$$

Відповідно для k-ї часткової задачі розглядаються всі набори змінних з номерами з цієї підмножини $X_k = \{x_i, i \in I_k\} = \{x_k, x_{k+1}, ..., x_m\}$.

Нижньою оцінкою $R(I_k)$ цільової функції (1) для підмножини розв'язків k - ї часткової задачі в [2] позначається таке дійсне число, яке обчислюється за формулою:

$$R(I_k) = \sum_{s=1}^{n} (y^s - \sum_{i \in I_k} a_i(I_k) x_i^s)^2 / n. \qquad (3)$$

Щоб довести, що значення функції $R(I_k)$ є нижньою оцінкою для значень функціоналу оцінки моделі $F(X_k^1 \subseteq X_k)$ необхідно показати, що для будь-якої підмножини X_k^1 з множини X_k виконується нерівність:

$$F(X_k^1) \geq R(X_k).$$

По перше. Зважаючи на означення (1) та (3)

$$R(I_k) = \sum_{s=1}^{n} (y^s - \sum_{i \in I_k} a_i(I_k) x_i^s)^2 / n \leq \max_{s=1,...,n} (y^s - \sum_{i \in I_k} a_i(I_k) x_s^i)^2 = F(I_k). \qquad (4)$$

По друге. Необхідно показати, що нерівність (4) виконується також для довільної підмножини X_k^1 з множини X_k, тобто для будь-якого $X_k^1 \subseteq X_k$. Якщо позначати через $I_k^1 = \{i / x_i \in X_k^i\}$, то виконується нерівність:

$$\sum_{s=1}^{n} (y^s - \sum_{i \in I_k} a_i(I_k) x_s^i)^2 \leq \sum_{s=1}^{n} (y^s - \sum_{i \in I_k^1} a_i(I_k^1) x_s^i)^2.$$

З цієї нерівності і випливає остаточна достовірність теореми. Оскільки маємо:

$$R(I_k) = \sum_{s=1}^{n} (y^s - \sum_{i \in I_k} a_i(I_k) x_s^i)^2 \leq \sum_{s=1}^{n} (y^s - \sum_{i \in I_k^1} a_i(I_k^1) x_s^i)^2 \leq \max_{s=1,...,n} (y^s - \sum_{i \in I_k^1} a_i(I_k^1) x_s^i)^2 = F(X_k^1)$$

Аналогічно доводиться, що функція $R(I)$ монотонно зростає при проведенні процедури розгалуження. Ця умова є необхідною методологічною вимогою для процедурі визначення нижньої оцінки цільової функції в методі гілок та границь.

Розглянемо використання методу гілок і границь для пошуку найкращого розв'язку на множині розв'язків k-го типу I_k $(k = 1, ..., m)$.

Здійснюємо розбиття (розгалуження) множини на дві підмножини за принципом дихотомії (навпіл).

Перша підмножина номерів $I_k^1 = \{k, k+1, k+2, ..., k+[(m-k)/2]\}$ генерує підмножину розв'язків задачі, які містять розв'язки зі змінними типу $(k, k+1, ...)$, $(k, k+2, ...)$,..., $(k, k+[(m-k)/2])$.

Друга підмножина номерів $I_k^2 = \{k, k+[(m-k)/2]+1, ..., m\}$ генерує підмножину розв'язків задачі, які містять розв'язки зі змінними типу $(k, k+[(m-k)/2]+1, ...)$, $(k, k+[(m-k)/2]+2, ...)$, $(k, k+[(m-k)/2]+3, ...)$, ..., (k, m).

Позначимо $p = [(m-k)/2]$.

Значення нижньої оцінки цільової функції обчислюємо аналогічно (3), для першої підмножини I_k^1:

$$R_1^{(k)} = R(I_k^1) = (\sum_{l=1}^{n} (y^s - \sum_{i=k}^{k+p} a_i(I_k^1) x_i^s)^2 / n, \quad (5)$$

де коефіцієнти $a_i(I_k^1)$, $i = k, k+1, ..., k+p$, знайдені за МНК для набору регресорів $\{x_k, x_{k+1}, ..., x_{k+p}\}$.

Аналогічно для другої підмножини розв'язків I_k^2:

$$R_2^{(k)} = R(I_k^2) = \sum_{l=1}^{n} (y^s - \sum_{i=k+p+1}^{m} a_i(I_k^2) x_i^s)^2 / n, \quad (6)$$

де коефіцієнти $a_i(I_k^1)$, $i = k+p+1, ..., m.$, знайдені за МНК для набору регресорів $\{x_{k+p+1}, ..., x_m\}$.

У випадку, коли обидві нижні оцінки менші за рекорд, вибір напрямку розгалуження проводиться наступним чином. Обчислюються значення ймовірностей того, що підмножина містить оптимальний розв'язок. Для цих підмножин I_k^1 та I_k^2 відповідно. Для першої підмножини I_k^1 це буде

$$P_1^{(k)} = R_2^{(k)} / (R_1^{(k)} + R_2^{(k)}),$$

а для другої підмножини I_k^2, відповідно,

$$P_2^{(k)} = R_1^{(k)} / (R_1^{(k)} + R_2^{(k)}).$$

Далі відповідно до отриманого розподілу вибирається підмножина для подальшого розгалуження.

Обчислюється значення цільової функції (1) для цього розв'язку, і це значення приймається за рекорд.

На кожному кроці алгоритмічного процесу проводиться порівняння цього значення рекорду зі значеннями нижніх оцінок цільової функції для всіх визначених у процесі розгалуження підмножин розв'язків. Ті підмножини розв'язків, які ще не піддавалися процедурі подальшого розгалуження і для яких значення оцінок знизу не менше (більше) значення рекорду, є неперспективними для подальшого розгалуження і надалі не розглядаються.

Обчислювальний процес продовжується доти, поки врешті-решт на якомусь кроці не виникне ситуація, коли всі підмножини, побудовані при розгалуженні, мають значення нижніх оцінок цільової функції більше (не

менше) значення рекорду, або матимемо підмножину, яка має лиш один розв'язок. В першому випадку процедура закінчується, розв'язок уже знайдено. В другому випадку порівнюємо значення цільової функції знайденого розв'язку зі значенням рекорду, і якщо це значення менше рекорду, то беремо його за рекорд.

В цій алгоритмічній схемі методу гілок і границь переслідувалася мета, щоб на кожному кроці повної ітерації процесу розгалуження стохастичною процедурою вийти на підмножину з одним розв'язком для уточнення значення рекорду.

Чисельний експеримент

Тестовий приклад. Було згенеровано випадковим чином набір з 5 значень (Табл. 1) для кожного з вхідних незалежних *первинних показників* (регресорів) $\{x_1, ..., x_i, ..., x_m\}$ (m=6). Значення *вихідного показника y* були розраховані за такою формулою:

$$y = 2x_2 - x_3.$$

Таблиця 1

Згенеровані вхідні дані

x_1	x_2	x_3	x_4	x_5	x_6	y
1	5	-3	115	-0,54402	0	13
2	2	0	206	0,912945	-12,6965	4
3	1	5	333	-0,98803	0	-3
4	5	12	502	0,745113	47,20909	-2
5	3	21	719	0,26237	0	-15

В таблиці 2 представлені покроково результати обчислень пропонованого алгоритму.

Для наглядності використано структурні вектори, відповідні до поточних підмножин індексів I_k:

$$D^k = \{d_i\} = \begin{cases} 1, i \in I_k \\ 0, i \notin I_k \end{cases}, \; i = \overline{1, m}$$

Тобто, якщо регресор входить до моделі, то відповідний йому елемент структурного вектора дорівнює одиниці, якщо не входить – нулю.

На першому кроці (k=*1*) обчислюється значення цільової функції для моделі, в яку входять всі регресори (структурний вектор складається тільки з одиниць). Це значення $F(I_1)$ береться за рекорд.

Таблиця 2

		x_1	x_2	x_3	x_4	x_5	x_6	$F(I_k)$	$R(I_k)$
№	*k=1*								
1	d_1	1	1	1	1	1	1	7,34E-28	1,6E-28
2	d_1^1	1	1	1	0	0	0	*2,84E-29*	*7,26E-30*
3	d_1^2	1	0	0	1	1	1	160,16	37,98
4	d_2	1	1	1	0	0	0	*2,84E-29*	7,26E-30
5	d_2^1	1	1	0	0	0	0	34,07	18,22
6	d_2^2	1	0	1	0	0	0	31,15	14,01
	k=2								
7	d_2	0	1	1	1	1	1	7,58E-27	2,50E-27
8	d_2^1	*0*	*1*	*1*	*1*	*0*	*0*	*5,30E-27*	*1,58E-27*
9	d_2^2	0	1	0	0	1	1	272,80	78,11
10	d_2^1	0	1	1	1	0	0	5,30E-27	1,58E-27
11	d_2^{11}	**0**	**1**	**1**	**0**	**0**	**0**	**4,93E-30**	*2,45E-30*
12	d_2^{12}	0	1	0	1	0	0	18,81	8,81
	k=3								
13	d_3	0	0	1	1	1	1	16,21	7,05
	k=4								
14	d_4	0	0	0	1	1	1	211,58	60,27
	k=5								
15	d_5	0	0	0	0	1	1	214,34	81,90
	k=6								
16	d_6	0	0	0	0	0	1	225	82,83

Проводиться розбиття множини I_1 на підмножини, яким відповідають структурні вектори d_1^1 та d_1^2. Для них розраховуються оцінки знизу і порівнюються з рекордом. Значення $R_2(I_1)$ більше за рекорд, тому цей напрямок вважаємо неперспективним. Подальшому розгалуженню підлягає множина, якій відповідає структурний вектор d_1^1. Відповідне значення цільової функції береться за рекорд. Проводиться поділ і обчислення та порівняння відповідних оцінок знизу (рядки таблиці №=5, 6). Значення оцінок знизу більші за рекорд, тобто подальше розгалуження цих напрямків неперспективне. Перехід до наступного *k, (k=2).*

Повторюються процедури обчислень та порівнянь відповідних $F(I_2), R_1(I_1)$ та $R_1(I_2)$. Перспективним напрямком вважається той, якому відповідає структурний вектор d_2^1, проте подальше розгалуженню неможливе. За рекорд береться $F_2(I_1)$ (рядок таблиці №=11). Перехід до наступного *k (k=3).*

При подальших обчисленнях значення нижніх оцінок поточних множин виявились більшими за рекорд. Тобто, не знайшлось перспективного напрямку для подальшого розгалуження. Це означає, що знайдено структуру оптимального рішення для конкретної задачі. Структурний вектор знайденого рішення відповідає істинній моделі.

Зауваження. Наведений приклад є окремим випадком, коли структурний вектор не має нулів між одиницями. Подальша робота має бути направлена на відшукання умов (критеріїв), за яких задача вирішувалась би в повному обсязі.

Висновки

Запропоновано застосування методу гілок і границь для розв'язання задачі дискретної оптимізації з метою вибору оптимальної регресійної моделі.

Для розбиття поточної множини розв'язків задачі на підмножини розгалуження використовується принцип дихотомії. Для підмножин розгалуження обчислюються нижні оцінки значень цільової функції вибору оптимальної моделі. Вибір підмножини розв'язків з двох перспективних для процесу розгалуження може здійснюватися за стохастичною процедурою.

Наведено чисельний приклад роботи алгоритму, який є окремим випадком загальної задачі.

Література

1. Land A.H., and Doig A.G. An automatic method of solving discrete programming problems // Econometrics. V. 28 (1960). – P. 497-520.
2. Мельник І.М. Метод гілок і границь для розв'язання задачі вибору оптимальної регресійної моделі як задачі дискретної оптимізації // Індуктивне моделювання складних систем. Зб. наук. праць. – Київ: МННЦ ІТС НАН України, 2009. – С. 131-139
3. Ivan Melnyk, Galyna Pidnebesna. Features of the Implementation of Branch and Bound Method for Solving the Problem of Selecting Optimal Regression Model // Proceedings of the IV International Workshop on Inductive Modelling IWIM-2012, 8 -15 July 2012, Zhukyn-Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, 2012. – P. 19-22.