**УДК 519.1**

**І.М. Мельник**

# Застосування методу гілок та границь для вибору оптимальної регресійної моделі для мінімаксного функціоналу оцінки моделі

*В статті пропанується стохастичний метод гілок та границь для рішення дискретної задачі оптимізації по вибору оптимальної регресійної моделі для мінімаксного функціоналу оцінки моделі. Для розбивки поточної множини рішень задачі на підмножини розгалуження використовується принцип дихотомії. Для підмножин розгалуження спеціальною формулою обчислюються оцінки знизу цільової функції задачі вибору оптимальної моделі. Вибір поточної підмножини рішень для проведення процедури розгалуження здійснюється стохастичною процедурою.*

***Ключові слова**: задача вибору оптимальної регресійної моделі, метод гілок та границь, підмножина розгалуження, оцінка знизу для цільової функції.*

*The article offered stochastic method of branches and boundaries to resolve a discrete task of optimization on the choice of optimum regressive model for mini-max function of the mode assessment. For break down of solution set of task into parts of fork principle subset the dichotomy principle is used. For subset of fork by the special formula estimation calculation is made for assessment of decrease by special function of optimum model having. The choice of current subset of decisions for conducting of fork procedures is carried out by stochastic procedure.*

***Keywords:** task of choice of optimum regressive model, method of branches and boundaries, subset of fork, assessment of decrease by special function.*

**Вступ.** Для більшість задач економічного характеру параметри (змінні), які їх характеризують мають дискретний характер. Отже пошук рішення цих задач мають дискретний (переборний) характер. З математичної

точки зору рішення цих задач відноситься до так званих задач дискретної оптимізації.

Множина $X$ є *дискретною* множиною, якщо для будь-якого її елементу $x(x \in X)$ існує число $\varepsilon(x) > 0$ таке, що для кожного елементу $y$ з цієї множині $X$ має місто нерівність $\|x - y\| > \varepsilon(x)$. Тобто, для кожного елемента з множини $X$ існує область (сфера) в просторі дійсних чисел, в які крім цього елементу не існує інших елементів цієї множини.

Вперше метод гілок і границь був запропонований Лендом і Дойгом [1] в 1960 році для вирішення загальної задачі цілочисельного лінійного програмування. Інтерес до цього методу і фактично його "друге народження" пов'язано з роботою Літтла, Мурті, Суїні і Керела [2], присвяченій задачі комівояжера . Починаючи з цього моменту, з'явилося велике число робіт, присвячених методу гілок і границь та різних його модифікацій. Такий великий успіх пояснюється тим, що ці автори першими звернули увагу на широту можливостей цього методу, відзначили важливість використовування при цьому специфіки вирішуваної задачі та самі продемонстрували це, скориставшись для розробки спеціального ефективного алгоритму рішення задачі комівояжера специфікою цієї задачі.

В основі алгоритмічної схеми методу гілок і границь лежить ідея послідовного розбиття поточної множини допустимих рішень на її підмножини розгалуження. На кожному кроці цього методу елементи розбиття (тобто підмножини) піддаються перевірці для з'ясування, чи може містить дана підмножина оптимальне рішення або явно ні. Перевірка здійснюється за допомогою обчислення

значення оцінки (оцінки знизу для задачі мінімізації) для цільової функції на даній підмножині і співставлення цього значення оцінки з значенням на цей момент *рекорду*. **Рекорд** - це значення цільової функції для найкращого на даний момент із знайдених рішень. Якщо оцінка знизу цільової функції на підмножини рішень, яка оцінюються, не менше (більше) рекорду, то ця підмножина може бути явно відкинута. Підмножина рішень, яка перевіряється, може бути відкинутий ще і у тому випадку, коли вдається знайти на якомусь кроці найдеться рішення, яке найкраще оцінки знизу цільової функції на ці підмножині. Якщо значення цільової функції для знайденого рішення менше раніше обчисленому рекорду, то значення рекорду змінюється на знайдене значення цільової функції. Коли, на якомусь кроці, вдається відкинути всі елементи (підмножини рішень) розбиття, то рекорд — це оптимальне значення рішення початкової задачі. В іншому випадку, з не відкинутих підмножин вибирається найбільш перспективне (наприклад, з якнайменшою оцінкою знизу цільової функції), і воно піддається розбиттю. Нові підмножини знов піддаються перевірці і т.д. до тих пір, поки значення рекорду не буде менше (не більше) значень ніжних оцінок для усіх підмножин розгалуження. По закінченню обчислювального процесу алгоритму поточне значення рекорду є оптимальне значення цільової функції, а відповідне рішення і є оптимальне рішення задачі.

У загальному випадку задача дискретної оптимізації формулюється так: необхідно найти оптимум (мінімум) функції $F(x)$, де змінна (параметр) $x$ вибирається з деякої дискретної множини $X$, тобто розглядається така задача:

$$F(x) \rightarrow \min, \quad (1)$$

$$x \in X, \quad (2)$$

де $X$ - дискретна множина.

У статті розглядається задача вибору оптимальної регресійної (економетричної) моделі. Задача полягає у вибору з множини факторів, які описують економічний процес, таку підмножину факторів з усієї множини , яка достатньо адекватно представляють функціонування цього процесу і при цьому оптимізують вибрану кількість факторів на основі деякого функціоналу оцінки моделі .Ця задача має дискретній характер і відноситься до так званих $NP$ повних задач. Пошук її рішення носить переборний характер і представляє собою дуже трудомісткий обчислювальний процес.

**Основний матеріал.** Розглядається складна система, яка характеризується $n$ вхідними (незалежними) змінними $x_1, ..., x_i, ..., x_n$ та одною вихідною (залежною) зміною $y$, що мають стохастичний характер і задані вибіркою з $m$ статистичних спостережень цих змінних. В процесі ідентифікації генеруються структури лінійних регресійних моделей, параметри яких оцінюються за методом найменших квадратів (МНК). Заданий функціонал оцінки моделі $F(.)$, тобто функціонал вибору оптимальної моделі, який має мінімаксний характер.

Розглядається наступний критерій вибору регресійної моделі. Задана множина точок вибірки спостережень $S = \{1, ..., m\}$. Хай маємо деяку підмножину вхідних (незалежних) змінних з загальної множини $\{x_1, ..., x_i, ..., x_n\}$ . На ці підмножини буде будуватися відповідна регресійна модель типу: $y = \sum a_i x_i$ , де сумування проводиться по тім змінним з множини $\{x_1, ..., x_i, ..., x_n\}$ , які включені у вибрану підмножину, а

коефіцієнти $a_i$ обчислені за МНК для цієї підмножини. Визначемо, що $J$ - номера змінних з вибраної підмножини, тобто $J$ це підмножина з множини номерів незалежних змінних $I = \{1, ..., n\}$, $J \subseteq I = \{1,..., n\}$. Обчислюємо значення функціоналу оцінки регресійної моделі $F(.)$ так:

$$F(J) = \max_{l \in S}(y^l - \sum_{i \in J} a_i\ (J)\ x_i^l)^2\ , \qquad (3)$$

де $J = \{i / x_i \in X_1\}$ – підмножина з множини номерів незалежних змінних $I$, а $X_1$ - вибрана підмножина вхідних змінних, на яких буде будуватися регресійна модель. де через $a_i(J),\ \ i \in J$, будемо позначати коефіцієнти регресійної моделі, які знайдені за МНК для набору регресорів $J$, $J = \{i / x_i \in X_1\}$ - підмножина множини номерів вхідних змінних $\{1, ..., n\}$ для відповідна підмножина вхідних змінних $X_1$. Значення функціоналу (3) обчислюється як максимальне значення квадрату нев'язок між статистичним значенням вихідної змінної $y$ та значенням відповідної регресійної моделі по всім $k$ - им, $k \in S$, спостереженням значень вхідних змінних з $X_1$. Отже, для рішення початкової задачі необхідно найти таку підмножину $J^*$ з множини $I$, $J^* \subseteq I$, яка би мінімізувала значення функціоналу (3), тобто

$$F(J^*) = \min \max_{l \in S}(y^l - \sum_{i \in J} a_i\ (J)\ x_i^l)^2\ , \qquad (4)$$

де мінімум беретеся по всім підмножинам $J$ з множини $I$.

Пропонується задачу вибору оптимальної регресійної моделі формулювати і розв'язувати як задачу дискретної оптимізації.

В [3,4] задачу вибору оптимальної регресійної моделі формулювалась і розв'язувалась як задачу дискретної оптимізації на спеціальному графі $(I,\ U)$. Множина вершин $I$ цього графа формується так. Кожній незалежній змінній $x_i$ $(i=1,...,n)$ ставиться у відповідність вершина графа $i$. Додаються додатково ще дві вершини: 0 та $n+1$. Множина дуг $U$ будується таким чином. Вершина 0 з'єднується з кожною вершиною *i* $(i=1,...,n)$ дугою $(0,i)$. Далі, кожна вершина *i* з'єднується з кожною вершиною $j,\ j>i$, дугою $(i,j)$.

Трактується, що вершина $i$ $(i=1,\cdots,n)$ відповідає ситуації, коли незалежна змінна $x_i$ включається в регресійну модель. Тоді кожному шляху з вершини $0$ у вершину $n+1$ відповідає певний варіант побудови регресійної моделі, а саме такої, в яку у вигляді регресорів включаються незалежні змінні $x_i$, що відповідають вершинам графа $(I,\ U)$, через які проходить цей шлях. Кожному шляху $\mu$ з вершини 0 у вершину $n+1$ ставиться у відповідність його "довжина", яка дорівнює значенню функціоналу вибору оптимальної моделі **F(.)**. Отже, початкова задача побудови регресійної моделі (задача вибору набору регресорів) зводиться до знаходження найкоротшого шляху з вершини 0 у вершину $n+1$ на графі $(I,\ U)$.

Граф $(I,\ U)$ дозволяє послідовно, економним способом, організувати перебір варіантів розв'язання задачі вибору оптимальної регресійної моделі та їх співставлення між собою. Знаходження найкоротшого шляху на побудованому графі $(I,\ U)$ і дає розв'язок основної задачі вибору оптимальної регресійної моделі.

Для так сформульованої задачі вибору оптимальної регресійної моделі як задачі озимизації на відповідному графі був запропонований генетичний метод рішення цієї задачі [3,4] як ефективний метод наближеного рішення початкової задачі.

В даної статті пропонується метод точного розв'язання початкової задачі побудови оптимальної регресійної моделі – стохастичний аналог методу гілок та границь рішень задач дискретної оптимізації . Як звісно, для цього методу суттєво необхідно мати процедуру (формулу) обчислення оцінки знизу значення цільової функції задачі, яка підлягає рішенню, на любої довільної підмножини рішень цієї задачі. Для задачі вибору оптимальної регресійної моделі, яка розглядається, оцінка знизу функціоналу оцінки моделі буде визначатися (обчислюватися) так. Хай задана деяка підмножина номерів змінних початкової задачі $J_k = \{k, k+1, ..., n\}$ ($J_k \subseteq I$). Тоді відповідною для $J_k$ підмножиною рішень початкової задачі є всі набори змінних з номерами, компільовані з підмножини $J_k$, тобто всі набори змінних взяті з підмножини змінних $X_k = \{x_i, i \in J_k\} = \{x_k, x_{k+1}, ..., x_n\}$. Тоді **оцінкою знизу** $R(J)$ цільової функції (3) для підмножини рішень, яка

складається з множини всіх наборів змінних з $X_k$, береться таке обчислення:

$$R(J_k) = (\sum_{l=1}^{m} (y^l - \sum_{i \in J_k} a_i(J_k) x_i^l)^2) / (n - k + 1). \qquad (5)$$

Щоб показати що значення функції $R(J_k)$ є нижньою оцінкою для значень функціоналу оцінки моделі $F(X_k^1 \subseteq X_k)$ необхідно показати, що для любої підмножини $X_k^1$ з множини $X_k$ виконується нерівність

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

Доказ проводиться по принципу „припущення від протилежного".

Також можна доказати, що функція $R(J)$ монотонно зростає по мірі проведення процедури розгалуження.

Загальне рішення початкової задачі вибору оптимальної регресійної моделі на послідовне рішення $n$ часткових задач дискретної оптимізації. Кожна часткова задача дискретної оптимізації – це вибір найліпшого рішення на, так званої**,** множина рішень $k$ - ого типу - $I_k$ ($k = 1,...,n$). **Множина рішень** $k$ **- ого типу** - $I_k$ ($k = 1,...,n$), це множина таких рішень, які обов'язково містять змінну з номером $k$, $k > 0$ (тобто змінну $x_k$), а також, можливо, ще всі, або частину змінних $x_i$, номери яких більше $k$ ($i > k$). Таким чином, загальна множина рішень $I$ розбивається на $n$ підмножини $I_k, k = 1,...,n,$ ( рішень $n$ типів), які не перетинаються і в сумі дають загальну множину рішень.

Опишемо використання методу гілок та границь для множини рішень $k$ - ого типу - $I_k$ $(k=1,...,n)$.

Здійснюємо розбивку (розгалуження) цієї множини $I_k$ на дві підмножини по принципу дихотомії – пополам. Перша підмножина $I_k^1$ включає рішення, які містять змінні з наборами номерів $(k,k+1,...)$, $(k,k+2,...),(k,k+3,...),...,$ $(k,k+[(n-k)/2],...)$, де $[(n-k/2]$ - ціла частина значення $(n-k)/2$. Друга підмножина $I_k^2$ включає рішення, які містять змінні з наборами номерів $(k,k+[(n-k)/2]+1,...)$, $(k,k+[(n-k)/2]+2,...)$, $(k,k+[(n-k)/2]+3,...),...,$ $(k,n-2,...)$, $(k,n-1,...)$, $(k,n))$. В подальшому визначимо $p=[(n-k)/2]$. Обчислюємо значення нижньої оцінки цільової функції, аналогічно формулі (5), для першої підмножини $I_k^1$ - це $R_1^{(1)}$. Для підмножини $I_k^1$ рішень формула (5) приймає такий вигляд:

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

де коефіцієнти $a_i(I_k^1),$, $i=k,k+1,...,k+p$, найдені по методу найменших квадратів (МНК) для набору регресорів $x_k, x_{k+1},..., x_{k+p}$.

Аналогічно формулі (5) , обчислюємо значення оцінки знизу цільової функції для другої підмножини рішень $I_k^2$ - це $R_2^{(1)}$. Для цієї підмножини рішень формула (5) приймає такий вигляд:

$$R_2^{(1)} = R(I_k^2) = (\sum_{l=1}^{m} (y^l - \sum_{i=k+p+1}^{n} a_i(I_k^2) x_i^l)^2 / (n-k-p), \qquad (5'')$$

де коефіцієнти $a_i(I_k^1)$, $i = k + [(n-k/2]+1,...,n.$, найдені за методом найменших квадратів для набору регресорів $x_{k+[(n-k)/2]+1},..., x_n$.

Обчислюємо значення вірогідностей для підмножин $I_k^1$ та $I_k^2$. Для першої підмножини $I_k^1$ це буде $P_1^{(1)} = R_2^{(1)} / (R_1^{(1)} + R_2^{(1)})$, а для другої підмножини $I_k^2$ - це $P_2^{(1)} = R_1^{(1)} / (R_1^{(1)} + R_2^{(1)})$. Далі, „розігруємо” з цими вірогідностями вибір підмножини ($I_k^1$ або $I_k^2$) для подальшого розгалуження. З вибраної підмножини рішень поступаємо аналогічно. Наприклад, в результаті „розіграшу” отримали для розгалуження підмножину $I_k^1$. Розділяємо її по принципу дихотомії на дві підмножини: $I_k^{11}$ та $I_k^{12}$. Підмножина рішень $I_k^{11}$ містить змінні з наборами номерів $(k, k+1,...)$, $(k, k+2,...),...,$ $(k, k+[[(n-k)/2]/2],...)$. А підмножина рішень $I_k^2$ містить змінні з наборами номерів $(k, k+[[(n-k)/2]/2]+1,...)$, $(k, k+[[(n-k)/2]/2]+2,...),...,$ $(k, k+[(n-k)/2],...)$. Обчислюємо, відповідно формулам (5), $(5')$ та $(5'')$, значення ніжних оцінок цільових функцій для цих двух підмножин. Хай для підмножини $I_k^{11}$ це буде $R_1^{(2)}$, а для підмножини $I_k^{12}$ - це $R_2^{(2)}$. На основі значення цих оцінок знизу $R_1^{(2)}$ та $R_2^{(2)}$ обчислюємо

відповідні вірогідності вибору підмножини ($I_k^{11}$ або $I_k^{12}$) для продовження процесу розгалуження. Для підмножини $I_k^{11}$ це буде $P_1^{(2)} = R_2^{(2)}/(R_1^{(2)} - R_2^{(2)})$, а для д підмножини $I_k^{12}$ це буде $P_2^{(2)} = R_1^{(2)}/(R_1^{(2)} - R_2^{(2)})$. Проводимо, відповідно цих обчислених вірогідностей, процедуру стохастичного „розиграшу" по вибору підмножини ($I_k^{11}$ або $I_k^{12}$) для подальшого розгалуження. Після вибору підмножини розбиваємо (дробимо) її по принципу дихотомії на дві нові підмножини. Для них знову обчислюємо значення оцінок знизу цільових функцій, обчислюємо відповідні вірогідності для стохастичного вибору наступної підмножини для продовження процесу розгалуження, проводимо відповідний „розиграш" і так далі, поки на черговому етапу розгалуження не одержимо підмножину з одним елементом (рішенням). Обчислюємо значення цільової функції (3) для цього рішення і приймаємо це значення за рекорд. Проводимо співставлення цього значення рекорду з значеннями оцінок знизу цільових функцій для усіх визначених при процесі розгалуження підмножин. Ті підмножини, які ще не піддавалися процедурі подальшого розгалуження і для яких значення оцінок знизу не менше (більше) значення рекорду, є *не перспективними* для подальшого розгалуження і викидаються з подальшого розгляду для цього процесу. З тих підмножин рішень, які лишилися для розгляду процесу розгалуження, вибираємо деяку підмножину (як початкову). Для неї здійснюємо, аналогічно як було описано вище, процес розгалуження на основі принципу дихотомії. При цьому, якщо на якомусь кроці процесу розгалуження виникають підмножини

рішень, для яких ніжні оцінки не менше (або більше) рекорду, то вони викидаються з подальшого розгляду для процесу розгалуження. Цей процес продовжується до тих пір, поки на якомусь кроці не виникни випадок, що всі підмножини, які збудовані при розгалуженні, мають значення оцінок знизу цільових функцій більше значення рекорду, або матиме підмножину яка має лиш одно рішення. В першому випадку процедура рішення закінчується, того що рішення знайдено. В другому випадку порівняємо значення цільової функції знайденого рішення з значенням рекорду, і якщо значення менше рекорду, то це значення беремо за рекорд. Потім знову переходимо на начало процесу розгалуження для початку проведення цього процесу. Весь процес обчислення алгоритму закінчується тоді, коли маємо такий рекорд, що його значення не більше (менше) значення нижчих оцінок для усіх підмножин, які виникли в процесі розгалуження. При цьому рішення, які відповідає цьому значенню рекорду, і є рішенням основної задачі вибору оптимальної регресійної моделі.

У описаної вище алгоритмічної схеми методу гілок та границь рішення задачі вибору оптимальної регресійної моделі послідовно ділили кожну підмножину розгалуження по принципу дихотомії на дві підмножини і з них вибирали одну (по стохастичному принципу), ту, яку на даний момент буде розгалужуватися, тобто по лінійної ієрархічної схемі. Цей процес розгалуження здійснюється до тих пір, поки не буде визначимо підмножину, яка буде містити лиш одне рішення. Значення цільової функції і буде визначатися як значення рекорду. Після цього переходиться на начало процесу розгалуження. При цьому довільно вибираємо початкову підмножину для

розгалуження. Знову, аналогічно здійснюємо, по лінійні схемі, процес розгалуження підмножин до тих пір, поки не визначимо підмножину, яка буде містити лиш одне рішення. Значення цільової функції для цього рішення зіставляється з значенням рекорду, для визначення подальше значення рекорду і так далі, до тих пір поки не буде визначено оптимальне рішення початкової задачі, значення цільової функції якого буде рівнятися значенню рекорду. У цієї схемі методу переслідувалася мета, щоб на кожної ітерації розгалуження дійте на підмножину з одним рішенням, для уточнення ( поліпшення) значення рекорду.

Отже, у описаної вище логічної схеми обчислювального процесу методу гілок та границь рішення задачі вибору регресійної моделі послідовно на кожному кроці розгалуження ділили (дробили) на дві підмножини рішень (стохастичною процедурою) тільки з тієї множини, яку перед цим розгалужувалася. Тобто на кожному кроці проведення процедури розгалуження вибираємо підмножину для подальшого розгалуження тільки з тих двох підмножин, які були визначені на попередньому кроці розгалуження.

Цю послідовну ієрархічну ціпочку (послідовність) розвиваємо до тих пір, поки або не одержимо у результаті процедури розгалуженні підмножини, які мають значення нижньої оцінок більше значення рекорду, або не одержимо підмножину з одним рішенням. В останньому випадку здійснюємо перевизначене значення рекорду. А в першому випадку переходимо на початок процесу розгалуження. Проте, процес розгалуження можна змінити таким образом, щоб на кожному кроці, для вибори кандидатури підмножини рішень для подальшого розгалуження на цьому кроці, розглядати всі підмножини рішень, які були

отримані на попередніх кроках і для яких не здійснювалась процедура розгалуження. Для цього одночасно визначаються значення вірогідностей для всіх цих підмножин рішень, на базі яких проводиться „розиграш" для вибору кандидатури підмножини для подальшого розгалуження. Яка у свою чергу, буде піддаватися процедурі розгалуженні по розбивки її, відповідно принципу дихотомії, на дві підмножини. Для цих двох підмножин визначаємо оцінки знизу значень цільової функції, як і для інших підмножин рішень, відповідно формул $(5')$ та $(5'')$. Ці дві підмножини приєднуємо, замість підмножини, яка була роздроблена на ці дві підмножини, до множини підмножин рішень, які до цього моменту не піддавалися процедурі розгалуження. Для нової множини підмножин рішень, які не піддавалися процедурі розгалуження, на основі значень оцінок знизу обчислюємо відповідні вірогідності. На основі значень цих вірогідностей проводимо стохастичний „розиграш" по вибору поточної підмножини зі всієї множини підмножин рішень для здійснення подальшої процедури розгалуження. Далі для вибраної таким засобом підмножини здійснюється розгалуження на дві підмножини, потім проводиться обчислення для них оцінок знизу значення цільової функції і так далі до тих пір, поки не отримаємо випадок, при якому поточне значення рекорду (тобто значення цільової функції відповідного рішення) не більше значень нижчих оцінок цільової функції для всіх підмножин, які були отримані в процесі проведення процедур розгалуження.

**Література**

1. Land A.H., and Doig A.G. An autmatic method of solving discrete programming problems. Econometrica. v28 (1960), pp. 497-520.

2. Little J.D.C., Murty K.G., Sweeney D.W., and Karel C. An algorithm for the traveling salesman problem. Operations Research. V. 11 (1963), pp. 972-989.
3. Мельник И.М. Генетический алгоритм решения задачи построения оптимальной регрессионной модели как задачи дискретной оптимизации // Проблемы управления и информатики. – 2008. - №3. – С.30-43.
4. Мельник І.М. Генетичний метод рішення задачі побудови оптимальної регресійної моделі / Економіко-математичне моделювання соціально-економічних систем. Збірник наукових праць. Вип. 13 / Відп. ред. академік НАН України О.О.Бакаєв. – Київ: МННЦ ІТС НАНУ, 2008.- С.129-147.