УДК 519.8

І. М. Мельник

ГЕНЕТИЧНИЙ МЕТОД РІШЕННЯ ЗАДАЧІ ПОБУДОВИ ОПТИМАЛЬНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ

В статі розглядається задача побудови оптимальної регресійної моделі складної системи, яка характеризується n *вхідними (незалежними) змінними та одною вихідною (залежною) зміною , що мають стохастичний характер.. Задача побудови оптимальної регресійної моделі полягає в виборі з всієї множини вхідних незалежних змінних підмножину, яка оптимізує заданий функціонал оцінки вибору моделі. В статі ця задача формулюється як задача дискретної оптимізація на спеціальному графі. Пропонуються методи розв'язання цієї задачі як задачі пошуку найкоротшого шляху на цьому графі. Особливу увагу приділяється використанню ідей генетичного алгоритму (як евристичного) пошуку глобального оптимуму для цієї складної задачі.*

Вступ. Задача побудови оптимальної регресійної моделі полягає у виборі з загальної множини незалежних змінних (регресорів) відповідної підмножини змінних для включення цих змінних у кінцеву регресійну модель. Задача є багатоекстремальною задачею дискретної оптимізації перебірного типу. На сьогодні достатньо ефективних алгоритмів розв'язання цієї задачі немає.

В статті задача побудови оптимальної регресійної моделі формулюється як задача дискретної оптимізації на спеціальному графі. Пропонується методи розв'язання цієї задачі як спеціальної задачі пошуку найкоротшого шляху на графі. Особливу увагу приділяється використанню ідеї генетичного алгоритму (як евристичного) для пошуку глобального оптимуму для цієї складної задачі [1,2].

Постановка задачі. 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 деякий функціонал оцінки моделі F(.), тобто функціонал вибору оптимальної моделі. Цей функціонал може залежить від складності моделі (числа оцінюваних

параметрів, які включаються в модель) та/або нев'язок рівняння регресії на деякої підмножині статистичних вибірок та інших параметрів моделі.

Отже, задача побудови оптимальної регресійної моделі: необхідно з усієї множини вхідних змінних $x_1, ..., x_i, ..., x_n$ вибрати таку підмножину змінних, щоб регресійна модель, яка побудована на основі цієї підмножини, давала би оптимум (мінімум або максимум) заданого функціоналу оцінки моделі F(.). Відповідну регресійну модель, в яку включені тільки змінні (регресори) з вибраної таким чином підмножини змінних, будемо називати оптимальною моделлю.

Вирішення задачі. Пропонується задачу вибору оптимальної регресійної моделі формулювати і розв'язувати як задачу дискретної оптимізації на спеціальному графі $(I, \ U)$. Множина вершин I цього графа формується так. Кожній незалежній змінній $x_i \quad (i = 1, ..., n)$ 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 $i \quad (i = 1, \cdots, n)$ 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 μ 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.

Пропонуються два методу точного розв'язання задачі побудови оптимальної регресійної моделі як задачі знаходження найкоротшого шляху на відповідному графі. Знаходження найкоротшого шляху на графі здійснюється процедурою послідовного відмічування вершин цього графа спеціальними позначками з можливим відсівом неперспективних варіантів на етапі конструювання варіантів рішення. У першому методу функціонал оцінки вибору регресійної моделі включає як оцінки складності моделі, так і квадрати нев'язок. рівняння регресійної моделі на всієї множині статистичних вибірок. У другому методі, з розбиттям всієї множини статистичної вибірки на дві підмножини (що є типовим для МГУА [3]), розв'язується мінімаксна задача дискретної оптимізації для квадратів нев'язок рівняння регресійної моделі на одної з підмножини статистичних вибірок .

Перший варіант методу це метод розв'язання задачі побудови оптимальної регресійної моделі для функціоналу оцінки моделі, який залежить від складності моделі (числа оцінюваних параметрів, які включаються в модель, помножених на ваги, та квадратів нев'язок рівняння регресії на множині статистичних вибірок), тобто необхідно вибрати таку підмножину J *з* множини $\{1, ..., n\}$, що

$$\sum_{i \in J} C_i + \sum_{k=1}^{m} (y^k - \sum_{i \in J} a_i\ (J)\ x_i^k)^2 \Rightarrow \min, \qquad (1)$$

де через $a_i(J),\quad i \in J$, 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 $\{1, ..., n\}$ та X_1 відповідна підмножина вхідних змінних, а C_i - "витрати" на отримання статистичної інформації про значення незалежної змінної x_i $(i \in J)$.

Нехай маємо дві підмножині J_1 і J_2 з множини $\{1, ..., n\}$ номерів незалежних змінних x_i, $i = 1,...,n$, тобто $J_1 \subseteq \{1,...,n\}$, $J_2 \subseteq \{1,...,n\}$. .

Лема 1. Якщо маємо $J_1 \subseteq J_2$, то для цих підмножин виконується таке співвідношення:

$$\sum_{k=1}^{m}(y^k - \sum_{i \in J_1} a_i(J_1)\, x_i^k)^2 \geq \sum_{k=1}^{m}(y^k - \sum_{i \in J_2} a_i(J_2)\, x_i^k)^2 \quad (2)$$

Будемо позначати варіант вибору регресійної моделі (варіант вибору сукупності незалежних змінних x_i для моделі (1)), як вибір значення булевих змінних δ_i, $i = 1, ..., n,$ у векторі $\delta = (\delta_1, ..., \delta_n)$, де $\delta_i = 1$, якщо незалежна зміна x_i включається в регресійну модель (тобто включається як регресор в модель (5)), та $\delta_i = 0$ у протилежному випадку.

Отже, задача (1) вибору варіанта регресійної моделі формулюється як задача дискретної оптимізації, або задачі цілочисельного булевого програмування: необхідно знайти мінімум функціоналу

$$\sum_{i=1}^{n} c_i \delta_i + \sum_{k=1}^{m}(y^k - \sum_{i \in J} a_i(J)\, x_i^k \delta_i)^2 \quad (3)$$

за умов

$$\delta_i \in \{0,1\}, \quad (4)$$

де $J = \{i \mid \delta_i = 1\}$, а коефіцієнти $a_i(J)$, $i \in J$, знаходяться за методом найменших квадратів.

Ця задача і розв'язується як задача знаходження відповідного "найкоротшого" шляху з вершини 0 у вершину $n+1$. Для варіанту функціоналу оцінки моделі (3) кожному шляху μ з вершини 0 у вершину $n+1$ будемо ставити у відповідність його "довжину", яка дорівнює

$$\sum_{l=1}^{p} C_{i_\ell} + \sum_{k+1}^{n} (y^k - \sum_{\ell=1}^{p} a_{i_\ell}(\{i_1, ..., i_\rho\}) x_{l_\ell})^2 \qquad (5)$$

Якщо визначити, що "найкоротший" шлях з вершини 0 у вершину $n+1$ є таким, якому відповідає мінімальне значення функціоналу (5), то початкова задача побудови регресійної моделі (задача вибору регресорів) зводиться до находження "найкоротшого" шляху з вершини 0 у вершину $n+1$ на графі $(I, \quad U)$.

Як сказано вище, задачу знаходження найкоротшого шляху з вершини 0 у вершину $n+1$ на графі будемо здійснювати процедурою послідовного відмічання вершин цього графа спеціальними позначками - відмітками (тобто процедурою побудови цих відміток вершин) [4].

Для випадку функціоналу оцінки (1) (або (5)) кожна відмітка P_i вершини $i, \quad i \in I$, якій відповідає шлях $\mu = [i_0 = 0, \quad i_1, ..., i_\ell = i]$ з вершини 0 у вершину *i*, складається з *чотирьох* ознак, тобто $P_i = (P_i^1, \quad P_i^2, \quad P_i^3, \quad P_i^4)$, де

P_i^1 - порядковий номер цієї відмітки в вершини i;

P_i^2 - вершини з множини $\{1, ..., n\}$, через які проходить шлях $\mu_i = [i_0, \quad i_1, ..., \quad i_\ell]$, тобто $P_i^2 = \{i_1, ..., i_\ell\}$, якщо $i_\ell \neq n+1$, та $P_i^2 = \{1, ..., i_{\ell-1}\}$, якщо $i_\ell = n+1$;

$P_i^3 = \sum_{r=1}^{\ell} C_{i_r}$ - тобто ознака P_i^3 дорівнює сумі C_i для всіх вершин, через які проходить шлях μ;

$P_i^4 = \sum_{k=1}^{m} (y^k - \sum_{r=1}^{\ell} a_{i_r}(\{i_1, ..., i_\ell\}) x_{i_r}^k)^2$ - тобто P_i^4 дорівнює значенню квадратів суми нев'язок рівняння регресії в

функціоналі (4) на множині вершин шляху μ_i (на множині вершин $\{i_1 ..., i_\ell\}$).

Множину варіантів шляхів з вершини 0 у вершину $n+1$ на графі (I, U) (тобто множину варіантів рішення задачі вибору оптимальної моделі за функціоналом оцінки (1)) здійснюється спеціальною технологією (процедурою) послідовного відмічання вершин цього графа спеціальними позначками - відмітками (тобто процедурою побудови цих відміток вершин) [7] з відсівом, на етапі конструювання шляхів, завідомо неперспективних варіантів за допомогою твердження Леми 1(формули (2)).

Процедура побудови множини відміток вершин графа (I, U) здійснюється у декілька етапів.

Перший (початковий) етап полягає у призначенні початкових значень позначок для відміток вершин графа.

Основний, другий етап, як раз і полягає в побудові всієї множини відміток вершин графа. Кожна відмітка вершини i графа – це відповідний шлях через вершини цього графа з вершин 0 в вершину i, та початковий відрізок шляху (шляхів) з вершин 0 в вершину $n+1$. По кожні відмітки вершину i здійснюється продовження відрізку шляху з вершин 0 в вершину i в сусідні, з вершиною i, вершини. Це здійснюється процедурою присвоєння нових відміток цим сусіднім вершинам. Причому на цьому етапі суттєво використовується твердження Леми 1 для конструювання тільки „перспективних" продовжень шляху з вершини i в сусідні з нею вершини. Процедура породження нових відміток продовжується доти, доки з'являються нові відмітки, тобто до тої пори поки будуватися варіанти перспективних шляхів в вершину $n+1$. Вершина $n+1$ помічається відмітками відповідно до вибраного правилу формування кінцевого набору варіантів рішення задачі побудови оптимальної регресійної моделі.

В останньому, заключному етапі, по відмітки (відміткам) вершини $n+1$ визначаються номера вершин графа, через які проходе саме «найкоротший» шлях з вершин 0 в вершину $n+1$. Ці

номера вершин і дають множину номерів змінних, які необхідно включити в оптимальну регресійну модель як регресорів.

Другий варіант методу розв'язання задачі побудови оптимальної регресійної моделі це метод рішення цієї задачі для минимаксного функціоналу оцінки моделі.

Розглядається наступний критерій вибору регресійної моделі. Множина точок вибірки $S = \{1, ..., m\}$ розбита на дві частини: S_1 та S_2, що є типовим для МГУА [3]. На першій частині вибірки (підмножини) S_1 будується регресійна модель, а на другі частині (підмножини) S_2 як раз обчислюється значення функціоналу оцінки побудованої моделі

$$F(J) = \max_{k \in S_2} (y^k - \sum_{i \in J} a_i\ (J)\ x_i^k)^2, \qquad (6)$$

де $J = \{i / x_i \in X_1\}$ – підмножина з множини номерів незалежних змінних $I = \{1, ..., n\}$, а X_1 - підмножина вхідних змінних, на яких будується регресійна модель. Значення функціоналу (6) обчислюється як максимальне значення квадрату нев"язок між статистичним значенням вихідної змінної y та значенням регресійної моделі по всім k - им, $k \in S_1$, спостереженням значень вхідних змінних з X_1.

Необхідно вибрати таку регресійну модель (тобто таку підмножину номерів незалежних змінних J з множини $I = \{1, ..., n\}$, для якої значення функціоналу (6) буде приймати мінімальне значення, тобто необхідно розв'язати задачу:

$$MIN_{J \subseteq I} \quad \max_{k \in S_2} (y^k - \sum_{i \in J} a_i\ (J)\ x_i^k)^2, \qquad (7)$$

по всім підмножинам J з множини $\{1, ..., n\}$.

Алгоритм розв'язання задачі (7) вибору оптимальної регресійної моделі як задачі дискретної оптимізації на графі $(I,\ U)$ знаходження відповідного найкоротшого шляху, визначеного значеннями функціоналу (6), аналогічний алгоритму для

першого варіанту, з деякими обчислювальними модифікаціями. Наприклад, буде відсутня ознака P_i^3 , а ознака P_i^4 буде обчислюватись відповідно до формули (6). Оскільки аналог Леми 1 для цього варіанта функціоналу вибору моделі можна брати як гіпотезу, то алгоритм забезпечує ефективне знаходження відповідного наближення до оптимального розв'язку початкової задачі.

Зробимо таку ремарку. Що до функціоналу оцінки моделі (6) - він має детерменірованний (не стохастичний) характер, хоча сам економічний процес має стохастичний характер. Цікавим може бути представлення функціоналу оцінки моделі у стохастичному вигляді, Наприклад, з використанням вірогідних обмежень: $\min \quad b$ при умові

$$P\{(y - \sum_{J \subseteq I} a_i(J)x_i)^2 \le b) \ge p,$$

де b - і є оцінка значення квадратів нев'язок рівняння регресії на підмножині статистичних вибірок, на якої обчислюється значення функціоналу оцінки моделі, а p - відповідний поріг вірогідності для цієї оцінки. Розгляд та дослідження цього підходу буде окремою публікацію, але зауважимо, що алгоритмічний підхід для розв'язання цієї задачі є [3].

Але методи точного розв'язання задачі побудови оптимальної регресійної моделі ефективні для не надто значних розмірностей, тобто коли кількість всіх визначально заданих змінних (регресорів) не є великою.

Для довільного, в тому числі для достатньо великого, числа первинних змінних (регресорів) пропонується спеціальний генетичний алгоритм розв'язання задачі вибору оптимальної регресійної моделі, який не гарантує точного розв'язання, але дає достатньо задовільне рішення.

Ідеї класичного генетичного алгоритму, були запропоновані в Мічиганському університеті Дж. Холландом [6]. Цей алгоритм часто дозволяє найти задовільне розв'язання аналітично не вирішуваних (або важко вирішуваних) проблем засобом підбору та конструювання відповідних процедур обчислювального процесу цього алгоритму з використанням механізмів, які аналогічні генетичним процесам еволюції у світі живих організмів. Головна особливість генетичного алгоритму є в тому, що на кожному його кроці (ітерації) аналізується

не одно рішення (хромосома), а деяка їх підмножина. Кожне рішення (хромосома) представляється у вигляді булевого вектора, компоненти якого приймають значення 0 або 1 і називаються генами. Ця підмножина (часто достатньо "добрих") хромосом (хромосом) називається "популяцією". На кожному циклі генетичного алгоритму на базі поточної популяції здійснюється за допомогою основних процедур алгоритму – кроссоверу (схрещування), мутації та селекції – процес генерування додатково нових хромосом, які називаються "нащадками". При цьому ітераційний процес будується так, що кожна послідуюча популяція повинна бути "кращою" в порівнянні з попередньою.

Отже, відповідно до термінології та визначень теорії генетичних алгоритмів [6,7], кожний шлях з вершини 0 в вершину $n+1$ на графі $(I,\ U)$ (тобто відповідний варіант вибору регресійної моделі), якому відповідає булевий вектор розмірності n, є хромосома відповідної популяції (підмножини рішень), а кожний елемент цього булевого вектора – ген цієї хромосоми.

Визначимо для кожної хромосоми μ, на основі значень цільових функції F(.) всіх хромосом популяції, функцію пристосовності (фітнесс – функція) P(μ) таким способом:

$$P(\mu) = F(\mu) / ((\text{СУМА } F(.) - F(\mu))), \qquad (6)$$

де СУМА **F(.)** – оператор сумування по всіма хромосомам популяції.

Лема 2. Якщо деяка хромосома дає локальний (глобальний) оптимум для основної задачі вибору оптимальної регресійної моделі (для цільової функції **F(.))**, то ця хромосома є локальним (глобальним) оптимумом для функції пристосовності **P(.)**, і навпаки.

Лема 3. Якщо μ_1 та μ_2 - хромосоми відповідної популяції та $F(\mu_1) < F(\mu_2)$, то $\mathbf{F(\mu_1) / F(\mu_2) < P(\mu_1) / P(\mu_2)}$.

Визначимо, що тип популяції k -ого типу (k -ий тип популяції) - ця множина хромосом (шляхів з вершини 0 в вершину $n+1$) які обов'язково містять вершину з номером k ($k > 0$) (тобто змінну x_k) **,** а також можливо ще вершини, номери яких більше k. Отже, загальна множина популяцій розбивається на n підмножин (n типів популяцій), які не пересікаються і які в сумі дають множину загальної популяції. Це означає, що в графі $(I,\ U)$

виділяються n частинних графів $(I_k,\ U_k)\ k=1,...,n,$ де I_k - підмножина вершин з множини вершин I основного графа $(I,\ U)$, $I_k = \{0,k,i>k\}$. А U_k - підмножина дуг з множини дуг U основного графа $(I,\ U)$, $U_k = \{(0,k);(k,j), j\in I_k, j>k;(i,j),i,j\in I_k, i>k, i<j\}$.

Кожний шлях в графі $(I_k,\ U_k)$ $(k=1,...,n)$ з вершини 0 в вершину $n+1$ (він обов'язково проходе через вершину з номером k)) є деяка популяція (хромосома) з популяції k -ого типу та є відповідним рішенням (яке обов'язково включає змінну x_k) для основної задачі по вибору оптимальної регресійної моделі.

Запропонований генетичний метод є ітераційним алгоритмічним процесом. На кожні поточні ітерації t $(t=0,1,..)$ поточна підмножина популяцій k – ого $W_k(t)$ піддається дій основних процедур генетичного алгоритму: кроссоверу – тобто виробленню нових популяцій k – ого типу шляхом схрещуванню, а також мутації та селекції, і додатково процедурі обміну (міграції) між підмножинами популяцій різних типів. В результаті одержується нова поточна підмножина популяцій k – ого типу $W_k(t+1)$, яка, як правило, є "кращою" („поліпшеною") в порівнянні з попередньою $W_k(t)$ і яка, в свою чергу, в подальшому, піддається діям обчислювальних процедур запропонованого генетичного алгоритму на $t+1$ ітерації. Цей ітераційний процес продовжується до тих пір, поки поточна підмножина популяцій k – ого типу поліпшується. У цієї моделі генетичного алгоритму важливе місце займає процедура побудування початкової підмножини популяцій k – ого $W_k(0)$ ($k=1,...,n$).

Вибір початкової підмножини k – ого типу популяції здійснюється спеціальним варіантом метода міток на графі $(I_k,\ U_k)$. У кожної вершини i може бути одна або декілька міток. Кожна мітка вершини i має чотири ознаки (числа):

(N_i, P_i^1, P_i^2, L_i), де N_i - номер попередній вершини , P_i^1 - номер метки у попередній вершини, P_i^2 - поточний номер цієї метки вершини i , L_i - значення функціоналу оцінки відповідної регресійної моделі. Для кожної дуги $(i, j) \in U_k$ графа $(I_k, \ U_k)$ вводиться вірогідний датчик с значеннями 0 або 1 $\xi_{ij} = \{0,1\}$ з дискретним розподілом вірогідностей $P_{ij} = (p_{ij}, 1 - p_{ij})$, де $0 \le p_{ij} \le 1$.

ПРОЦЕДУРА визначення початкової підмножини популяцій k – ого типу $W_k(0)$.

Початковий (нульовий) етап. Вводимо лічильник P_l поміток вершини l ($l = k+1,...,n,n+1,$), який буде визначати на кожні ітерації кількість вже зроблених поміток для цієї вершин. На початку покладаємо значення цих лічильників рівним 0. Вводимо для дуги $(i, j) \in U_k$ ($i = k, k+1,...,n$) вірогідний датчик $\xi_{ij} = \{0,1\}$ з дискретним розподілом вірогідностей $(p_{ij}, 1 - p_{ij})$, де вірогідність $0 < p_{ij} < 1$ задається.

Перший етап (помічування вершин 0 та k). На цьому етапі помічуємо вершини 0 та k, які будуть мате тільки по одній помітки. Помічаємо вершину 0 міткою $R_0^1 = (N_0 = 0, P^1{}_0 = 0, P^2{}_0 = 1, L_0 = 0)$. По помітки R_0^1 вершини 0 помічаємо вершину k міткою $R_k^1 = (N_k = 0, P^1{}_k = 1, P^2{}_k = 1, L_k)$, де $L_k = \max_{s \in S_2} (y^s - a_k x_k{}^s)^2$, а a_k визначено методом МНК, коли набір регресорів складається з одної змінної x_k на статистичної вибірки S_1 значень змінної x_k.

Основний етап помічування вершин графа $(I_k,\ U_k)$ (помічування вершин $k+1$, $k+2$, …, n, а також $n+1$). Помічування вершин $k+1$, $k+2$, …, n, а також $n+1$, виділено в окремий (основний) етап, тому що у них (крім вершини $k+1$) може бути більш чим одна помітка. Оскільки помічування цих вершин проводиться з урахуванням результатів „розиграшу" вірогідних датчиків $\xi_{ij}=\{0,1\}$, $(i,j)\in U_k$, то ці вершини можуть і не мати ні одної помітки.

Відпрацювання вершини k з її поміткою $R_k^1=(N_k=0, P^1{}_k=1, P^2{}_k=1, L_k)$. Здійснюємо по помітки R_k^1 вершини k процедуру помічування для вершини $k+1$. Для цього „розігруємо" для дуги (i_k, i_{k+1}) вірогідний датчик $\xi_{k,k+1}=\{0,1\}$. Якщо при цьому „випало" значення датчика $\xi_{k,k+1}$=1, то помічаємо вершину $k+1$ поміткою $R_{k+1}^1=(N_{k+1}=k, P_{k+1}^1=P_k^2, P_{k+1}^2=P_{k+1}+1, L_{k+1})$, де $L_{k+1}=\max_{s\in S_2}(y^s-\sum_{l=k}^{k+1}a_l x_l{}^s)^2$. Коефіцієнти a_k та a_{k+1} визначаються МНК для набору регресорів з x_k та x_{k+1} на статистичної вибірки S_1 значень цих змінних (регресорів). Збільшуємо значення лічильника P_{k+1} поміток вершини $k+1$ на 1. Покладаємо для лічильника $\xi_{k,k+1}$ новий розподіл вірогідностей $(p_{k,k+1}=1,0)$ і переходимо до розгляду процедури помічування для вершини $k+2$. А якщо, при розиграшу вірогідного датчика $\xi_{k,k+1}=\{0,1\}$ „випало" значення $\xi_{k,k+1}$ рівне 0, то переходимо зразу до розгляду процедури помічування для вершини $k+2$.

Здійснюємо по помітки R_k^1 вершини k процедури помічування для вершини $k+2$. „Розігруємо" для дуги (i_k, i_{k+2}) вірогідний датчик $\xi_{k,k+2}=\{0,1\}$ з дискретним розподілом вірогідностей $(p_{k,k+2}, 1-p_{k,k+2})$. Якщо при цьому „випало"

значення датчика $\xi_{k,k+2}$ рівне 1, то помічаємо вершину $k+2$ поміткою R^1_{k+2} = $(N_{k+2}=k, P^1_{k+2}=P^2_k, P^2_{k+2}=P_{k+2}+1, L_{k+2})$, де $L_{k+2}=\max_{s\in S_2}(y^s-(a_k x_k{}^s+a_{k+2}x^s_{k+2}))^2$. Коефіцієнти a_k та a_{k+2} визначаються МНК для набору регресорів з x_k та x_{k+2} на вибірки S_1. Збільшуємо значення лічильника P_{k+2} поміток вершини $k+2$ на 1. Покладаємо для датчика $\xi_{k,k+2}$ новий розподіл вірогідностей $(p_{k,k+2}=1,0)$ і переходимо до розгляду процедури помічування для вершини $k+3$. А якщо, при розиграшу вірогідного датчика $\xi_{k,k+2}=\{0,1\}$ „випало" значення $\xi_{k,k+2}$ рівне 0, то переходимо зразу до розгляду процедури помічування для вершини $k+3$.

Здійснюємо процедури помічування для вершини $k+3$ и т. д., поки не дійдемо до вершини n. „Розігруємо" для дуги (i_k, n) датчик $\xi_{k,n}=\{0,1\}$ з дискретним розподілом вірогідностей $(p_{k,n}, 1-p_{k,n})$. Якщо при цьому „випало" значення датчика $\xi_{k,n}$ рівне 1, то по помітки R^1_k вершини k помічаємо вершину n поміткою R^1_n = $(N_n=k, P^1_n=P^2_k, P^2_n=P_n+1, L_n)$, де $L_n=\max_{s\in S_2}(y^s-(a_k x_k{}^s+a_n x^s_n))^2$. Коефіцієнти a_k та a_n визначаються МНК для набору регресорів з x_k та x_n на статистичної вибірки S_1 значень x_k та x_n. Збільшуємо значення лічильника P_n поміток вершини n на 1. Покладаємо для датчика $\xi_{k,n}$ новий розподіл вірогідностей $(p_{k,n}=1;0)$. і переходимо до розгляду процедури помічування для вершини $n+1$.А якщо, при "розиграшу" вірогідного датчика $\xi_{k,n}=\{0,1\}$ „випало" значення $\xi_{k,n}$ рівне 0, то переходимо зразу, без помічування вершини n, до помічування вершини $n+1$.

По помітки $R_k^1 = (N_k = 0, P^1{}_k = 1, P_k^2 = 1, L_k)$ вершини k помічуємо вершину $n+1$ поміткою $R_{n+1}^1 = (N_{n+1} = k, P_{n+1}^1 = P_k^2, P_{n+1}^2 = P_{n+1} + 1, L_{n+1} = L_k)$. Збільшуємо значення лічильника P_{n+1} поміток вершини $n+1$ на 1.

Отже, на цьому кроці основного етапу повністю провели використання помітки вершини k для процедурі помічування вершин $k+1$, $k+2$,..., n, а також $n+1$.

Далі працюємо з вершиною $k+1$ з її поміткою (якщо вона є) $R_{k+1}^1 = (N_{k+1}, P_{k+1}^1, P_{k+1}^2, L_{k+1})$ (у цієї вершини може бути тільки одна помітка) для використання процедури помічування по неї для вершин $k+2$, $k+3$, ..., n, а також $n+1$.

„Розігруємо” для дуги (i_{k+1}, i_{k+2}) вірогідний датчик $\xi_{k+1,k+2} = \{0,1\}$ з дискретним розподілом вірогідностей $(p_{k+1,k+2}, 1 - p_{k+1,k+2})$. Якщо при цьому „випало” значення датчика $\xi_{k+1,k+2}$ рівне 1, то помічаємо вершину $k+2$ поміткою $R_{k+2}^q = (N_{k+2} = k+1, P_{k+2}^1 = P_{k+1}^2, P_{k+2}^2 = P_{k+2} + 1, L_{k+2})$, де $L_{k+2} = \max_{s \in S_2} (y^s - \sum_{l=k+1}^{k+2} a_l x_l{}^s)^2$, а $q = P_{k+2} + 1$. Коефіцієнти a_{k+1} та a_{k+2} визначаються МНК для набору регресорів з x_{k+1} та x_{k+2} на статистичної вибірки S_1 значень x_{k+1} та x_{k+2}. Збільшуємо значення лічильника P_{k+2} поміток вершини $k+2$ на 1. Покладаємо для датчика $\xi_{k+1,k+2}$ новий розподіл вірогідностей $(p_{k+1,k+2} = 1,0)$ і переходимо до розгляду процедури помічування вершини $k+3$. А якщо, при розиграшу вірогідного датчика $\xi_{k+1,k+2} = \{0,1\}$ „випало” значення цього датчика рівне 0, то переходимо зразу до розгляду вершини $k+3$. Розглядаємо використовування помітки R_{k+1}^1 вершини $k+1$ для процедуру помічування вершину $k+3$ и т. д.,

поки не дійдемо до вершини n. Розігруємо" для дуги (i_{k+1}, n) її датчик $\xi_{k+1,n} = \{0,1\}$ з дискретним розподілом вірогідностей $(p_{k+1,n}, 1 - p_{k+1,n})$. Якщо при цьому „випало" значення датчика $\xi_{k+1,n}$ рівне 1, то по помітки $(N_{k+1}, P^1_{k+1}, P^2_{k+1}, L_{k+1})$ вершини $k+1$ помічаємо вершину n поміткою R^q_n = $(N_n = k+1, P_n{}^1 = P^2_{k+1}, P^2_n = P_n + 1, L_n)$, де $q = P_n + 1$, а $L_n = \max_{s \in S_2} (y^s - (a_{k+1} x_{k+1}{}^s + a_n x^s_n))^2$. Коефіцієнти a_{k+1} та a_n визначаються МНК для набору регресорів з x_{k+1} та x_n на статистичної вибірки S_1. Збільшуємо значення лічильника P_n поміток вершини n на 1. Покладаємо для датчиика $\xi_{k+1,n}$ новий розподіл вірогідностей $P_{k+1,n} = (p_{k+1,n} = 1;0)$ і переходимо до помічування вершини $n+1$ по помітки R^1_{k+1} вершини $k+1$. А якщо, при "розиграшу" вірогідного датчика $\xi_{k+1,n} = \{0,1\}$ „випало" значення рівне 0, то переходимо зразу, без помічування вершини n, до помічування вершини $n+1$. Помічаємо по мітки $(N_{k+1}, P^1_{k+1}, P^2_{k+1}, L_{k+1})$ вершини $k+1$ вершину $n+1$ поміткою $(N_{n+1} = k+1, P^1_{n+1} = P^2_{k+1}, P^2_{n+1} = P_{n+1} + 1, L_n = L_{k+1})$. Збільшимо значення спеціального лічильника P_{n+1} поміток вершини $n+1$ на 1.

Далі працюємо з помітками вершини $k+2$ (якщо воне є) для використання процедури помічування по цим поміткам для вершин $k+3, k+4, ..., n$, а також $n+1$. У цієї вершини можуть бути не більш двох поміток (одна по помітки $(N_k, P_k{}^1, P_k{}^2, L_k)$ вершини k, а друга по помітки $(N_{k+1}, P_{k+1}{}^1, P_{k+1}{}^2, L_{k+1})$ вершини $k+1$ (якщо вершина $k+1$ була помічена). Беремо першу помітку вершини $k+2$ $R^1_{k+2} = (N_{k+2}, P^1_{k+2}, P^2_{k+2}, L_{k+2})$. По нї проводимо усі процедури помічування від вершини $k+3$ до вершини $n+1$. Потім

аналогічно працюємо з другою поміткою вершини $k + 2$ (якщо воне є).

Далі працюємо з помітками вершини $k + 3$ (їх може бути не більш трёх) і так до тех пір, поки не дійдемо до розгляду вершини $n-1$.

По поміткам вершини $n-1$ (їх може бути не більш чім $n-k-1$), почергово у відповідності з „розіграшем" вірогідного датчика $\xi_{n-1,n} = \{0,1\}$ помічаємо (або не помічаємо) вершину n та обов'язково помічаємо вершину $n+1$.

Далі переходимо до розгляду поміток вершини n (їх може бути не більш чім $n-k$). По її поміткам (якщо вони є) помічаємо почергово відповідними помітками вершину $n+1$.

Етап формування початкової підмножини популяцій k – ого типу $W_k(0)$. Провіряємо по поміткам вершини $n+1$: чи початкова підмножина популяцій k - ого типу містить відповідну кількість осіб? Якщо ні, то переходимо на начало **ПРОЦЕДУРИ** визначення початкової підмножини популяцій k – ого типу для отримання нових поміток вершини $n+1$ (а значить і нових популяцій k – ого типу). А якщо, вершина $n+1$ має відповідну кількість поміток, то, по суті, початкову підмножину популяцій k - ого типу побудовано. Кожні помітки вершини $n+1$ відповідає шлях в графі $(I_k,\ U_k)$ з вершини 0 в вершину $n+1$, який відповідає одному з рішень основної задачі вибору регресійної моделі.

Кроки процесу обчислень одного циклу генетичного алгоритму для розв'язання задачі вибору оптимальної регресійної моделі.

Великий цикл.

Крок I. Вибір почергово номера типу популяції **k** (k = 1, … ,m).

Крок II. Малий цикл.

Крок 1. Генерація направленим випадковим способом підмножини популяцій (рішень) для вибраного k – ого типу популяцій. Робиться це **ПРОЦЕДУРОЮ** визначення початкової підмножини популяцій k – ого типу $W_k(0)$.

Будемо вважати, що цей тип популяції повинен містити N_k осіб, де N_k визначається відповідним засобом.

Крок 2. Обчислення значень цільової функції **F(.)** і функції пристосовності **P(.)** для всіх хромосом генерованої популяції.

Крок 3. Вибір "батьків" з популяції хромосом типу **k** на основі поєднання двух принципів: інбридингу та аутбридингу, тобто на основі близької й далекої "спорідненості" батьків.

Крок 4. Кроссовер (схрещування) пар хромосом -"батьків" для генерації хромосом -"нащадків".

Крок 5. Обчислення значень цільової функції **F(.)** та функції пристосовності **P** (.) для хромосом - "нащадків".

Крок 6. Проведення процедури мутації для хромосом - "нащадків" для генерації хромосом - "мутантів".

Крок 7. Обчислення значень цільової функції **F(.)** і функції пристосовності **P (.)** для хромосом "мутантів".

Крок 8. Перевірка закінчення процесу обчислень для популяцій типу **k**. Якщо ні, то перехід до *Кроку 3* малого циклу. А якщо так, то перехід до наступного кроку.

Крок *9* . Проведення процедури міграції між підмножиною популяцією k - ого типу та підмножинами популяцій інших типів.

Крок 10. Проведення процедури селекції для генерації наступної підмножини популяції k - ого типу.

Крок 11. Перехід до **Кроку I** великого циклу для вибору наступного номера типу популяції (з перевіркою умови $k \le m$). Якщо виконується умова $k = m$, то здійснюється перехід на проведення обчислювань великого циклу для наступної (t+1) - ой ітерації.

Запропонована в статті модель генетичного алгоритму відноситься до, так званих, *острівних моделей* (island model) [8] – моделей паралельного генетичного алгоритму, в яких основна популяція розбивається на декілька різних типів (видів) популяцій. Кожна з типів популяції буде розвиватися окремо за допомогою деякого генетичного алгоритму (можливо що до різних типів популяцій можуть використовуватися різні варіанті цього алгоритму). Отже, можна образно сказати, що особі загальної популяції розселені, якщо загальна популяція відповідно до деякого правилу сегментовано на деяке число типів (видів) часткових популяцій, по декілька ізольованими островам. Зрідка може проходити міграція між островами – типи популяцій (острова) міняються деякими „добрими"

особами. Оскільки у обчислювальному аспекти генетичний алгоритм має стохастичний характер, то при різних його запусках популяція може „сходитися" до різних добрих рішень. Острівна модель дозволяє паралельно запустити генетичний алгоритм зразу декілька раз и сумістити (співставити) „досягнення" на різних островах для отримання найліпшого загального рішення.

Висновки. Представлення задачі побудови оптимальної регресійної моделі як задачі дискретної оптимізації пошуку найкоротшого шляху на спеціальному графі дозволяє плідно використовувати інструментарій технології рішення цієї задачі для вирішення початкової задачі.

Розглядаються два критерії оцінки побудови оптимальної регресійної моделі, з яких мінімаксний критерій оцінки є найоригінальнішим і представляє значний практичний інтерес. Пропонуються методи розв'язання задачі побудови оптимальної регресійної моделі як спеціальної задачі пошуку найкоротшого шляху. Особливий інтерес з них представляє запропонований генетичний алгоритм для вирішення цієї спеціальної задачі пошуку найкоротшого шляху як метод вирішення початкової задачі.

Література

- І.М. Мельник, В.С. Степашко. Про один підхід по вибору оптимальної регресійної моделі методом дискретної оптимізації // Міжнародний семінар з індуктивного моделювання. Збірник праць. // Відповідальний редактор В.С.Степашко - Київ: МННЦ ІТС НАН та МОН України, 2005. – С. 223-229.
- І.Мельник, Генетичний метод структурно-параметричної ідентифікації регресивної моделі складної системи //Матеріали XIV міжнародної з автоматичного управління (Автоматика – 2007), м. Севастополь,10-14 вересня 2007 року. – Ч.1 – Севастополь: СНУЯЄтаП, 2007, с. 80-82.
- А. Г. Ивахненко, В.С. Степашко, Помехоустойчивость моделирования. - К.: Наукова думка, 1985. – 216 с.
- Ю. М. Ермольев, И. М. Мельник, Экстремальные задачи на графах. – К.: Наукова думка, 1967. – 189.
- Ю. М. Ермольев, И. М. Мельник, О методах стохастического программирования с конечним числом испытаний //журнал «Кибернетика», 1974, № 4, с. 82 – 84.
- Holland J.H. Adaptation in Natural and Artificial Systems. Ann Arbor: The University of Michigan Press, 1975.

- Goldberg D.S. Genetic algorithms in search, optimization, and machine learning. – Reading, MA: Addsson-Wesley.1989.
- W.D. Whitley, S.B. Rana, and R.B. Heckendorn, “Island model genetic algorithms and linearly separable problems”, in Selected Papers from AISB Workshop on Evolutionary Computing. London, UK: Springer-Verlag, 1997, pp. 109-125.