УДК 681.513

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ПОШУКУ КРАЩОЇ СТРУКТУРИ ЛІНІЙНОЇ РЕГРЕСІЇ

В.С. Степашко, Я.А. Бондарська
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України
yana2301@gmail.com

Найбільш об‘ємна за обчисленнями та часом частина алгоритмів моделювання – пошук кращої структури. Можливий шлях зменшення кількості структур, що треба перевірити – попереднє виключення гірших структур з перебору. Один з алгоритмів, в якому використовується цей підхід – алгоритм Ла Мотта-Хокінга. В статті представлені результати порівняння початкового метода Ла Мотта-Хокінга, а також його модифікації. Результати дослідження будуть використані в модифікації комбінаторного алгоритму МГУА.
Ключеві слова: пошук кращих структур, метод Ла Мотта-Хокінга, структурне моделювання

Most computationally critical part of linear regression algorithms is selection of structures that should be checked. One way to reduce the number of such structures is to make some primary estimations of structures quality and divide the structures into groups according to this estimation. One technique that uses this approach is La Motte-Hocking algorithm. This paper represents research results of La Motte-Hocking method efficiency in comparison with the combinatorial algorithm. Results of research and some possible ways to improve results are presented in this paper. The outcomes are planned to be used for updating the combinatorial GMDH algorithm.
Keywords: selection of structures, La Motte-Hocking algorithm, structures modeling

Наиболее объемная по вычислениям часть алгоритмов моделирования – это поиск структуры модели. Возможный подход к сокращению количества структур, которые надо проверить – исключение худших структур из перебора. Один из алгоритмов, где реализуется данный подход – алгоритм Ла Мота-Хокинга. В статье представлены результаты сравнения метода Ла Мота-Хокинга, а также его модификации. Результаты будут использованы при улучшении комбинаторного алгоритма МГУА.
Ключевые слова: поиск лучших структур, метод Ла Монта-Хокинга, структурное моделирование

1. Постановка задачі пошуку кращої структури лінійної регресії

Задача пошуку кращої структури лінійної регресії полягає в переборі моделей з множини $\Im$ з метою пошуку моделі з найкращим значенням заданого показника якості. Формально задачу можна записати так:

$$d^* = \min_{\Omega_m}(CR(M(d)), \qquad (1)$$

де m – загальна кількість аргументів,

$d[1\times m]$ - структурний вектор,

$M(d)$ - частинна модель, що відповідає структурному вектору d,

$CR(M(d))$ - значення критерію якості для моделі $M(d)$,

Ω_m - множина можливих структур.

Структурний вектор d складається з 0 та 1, що визначають, які з виміряних вхідних аргументів будуть включені в модель.

Частинна модель включає не всі вхідні аргументи і визначається вектором структури d, вектором параметрів a, значенням критерію якості для даної виборки. Модель, що включає всі вхідні аргументи, називається повною. Кількість точок вхідних вимірювань є довжиною вибірки.

Мета пошуку кращої структури – визначення структури частинної моделі з оптимальним значенням критерію якості. Найпростіший метод пошуку кращої структури – повний їх перебір. Цей метод працює для кількості аргументів не більше 20, проте реальні задачі частіше містять набагато більше регресорів. Серед більш ефективних підходів до пошуку кращої структури можна виділити:

- поділ множини на групи, виключення груп з гіршими значеннями критеріїв (метод Ла Мотта-Хокінга);
- послідовне включення/виключення регресорів з моделі відповідно до зміни критерію якості моделі. (метод покрокової регресії).

Метод Ла Мотта-Хокінга. Метод побудований на наступній властивості критерію *RSS*:

$$\begin{array}{l} R1,\ R2 - \textit{частинні структури} \\ \textit{якщо}\ \ R1 \subseteq R2,\ \textit{тоді}\ \ RSS_{R1} \geq RSS_{R2}. \end{array} \qquad (2)$$

З (2) очевидно, що якщо є деяка частинна модель R^* і підмножина моделей $\{R\}_{R^*}$, подбудована виключенням з R^* деяких аргументів, значення критерію *RSS* моделі R^* буде нижнєю межею значень *RSS* для моделей з підмножини $\{R\}_{R^*}$.

Перший крок методу Ла Мотта-Хокінга –поділ множини структур на групи. Далі пошук проводиться в кожній групі окремо.

В методі також задаються параметр k та складність моделей. Пошук проводиться на одному рівні складності моделей. Кроки методу наступні:

1. генеруються всі можливі *k*-структури (структури, де виключено *k* аргументів). Знайдені структури впорядковуються за значенням RSS (за зростанням);
2. для кращої *k*-структури будується підгрупа моделей. Для цього з *k*-моделі додатково виключаються (*m*-*k*-*s*) аргументів. Проводиться пошук в групі, знаходиться модель з кращим значенням *RSS;*
3. *RSS* кращої моделі в групі порівнюється з *RSS* наступної *k*-моделі. Якщо *RSS* моделі не більше ніж *RSS* наступної *k*-моделі, пошук завершено. Інакше кроки 2, 3 повторюються для наступної *k*-моделі.

Критерій якості, що використовується в методі – *RSS*. Більшість критеріїв має наступну структуру:

$$QR = g(s)RSS + f(s), \qquad (3)$$

де s – складність моделі, $g(s)$, $f(s)$ – деякі монотонно зростаючі функції.

Якщо складність моделі фіксована, модель з кращим значенням *RSS* буде також мати краще значення *RSS* для будь-яких критеріїв зі структурою (3).

Переваги методу:

- результат методу співпадає з результатом повного перебору;
- переглядається набагато менше моделей, ніж при повному переборі.

Недоліки:

- метод є повільним для малої кількості аргументів;
- немає процедури для визначення параметра k;
- неможливо попередньо оцінити кількість моделей, яка буде згенерована.

Шляхи покращення методу Ла Мотта-Хокінга. Для великих складностей метод сповільнюється дуже швидко. Основні причини цього:

- збільшення кількості k-моделей, що треба згенерувати;
- збільшення кількості моделей в кожній групі.

Можливі шляхи покращення методу Ла Мотта-Хокінга:

1. не будувати всі k-моделі;
2. не проводити повний перебір моделей в групі.

Головним недоліком першого підходу є ймовірність, що пошук доведеться проводити для усіх k-моделей. Іншою проблемою є складність задання параметра k для кожного етапу методу

Другий підхід полягає у використанні більш ефективних методів пошуку в кожній групі, наприклад, покрокової регресії, або покрокового виключення. При покроковому виключенні послідовно виключаються аргументи, виключення яких призводить до найменшого збільшення критерію якості. Проблемою другого підходу є те, що метод покрокової регресії не є точним методом.

В даній роботі був реалізований другий підхід.

2. Результати дослідження

Мета та предмет дослідження. Мета даного дослідження – порівняння різних методів пошуку кращої структури. Порівнювались швидкість методу та точність результату. Методи, що були досліджені: метод повного перебору, метод Ла Мотта-Хокінга, модифікований метод Ла Мотта-Хокінга, метод покрокової регресії. Параметри виборки, на якій проводились дослідження:

- довжина виборки 25;
- рівень шуму 50% від максимального значення вихідного параметру.

Значення параметру k для метода Ла Мотта-Хокінга 3. Максимальна складність моделі, що може бути знайдена методом Ла Мотта-Хокінга (m-k-1).

Порівняння точності методів. Точність методу визначається як різниця результатів методу та повного перебору. Різниця може бути виміряна як кількість різних аргументів або як різниця критеріїв якості, наприклад, *RSS*.

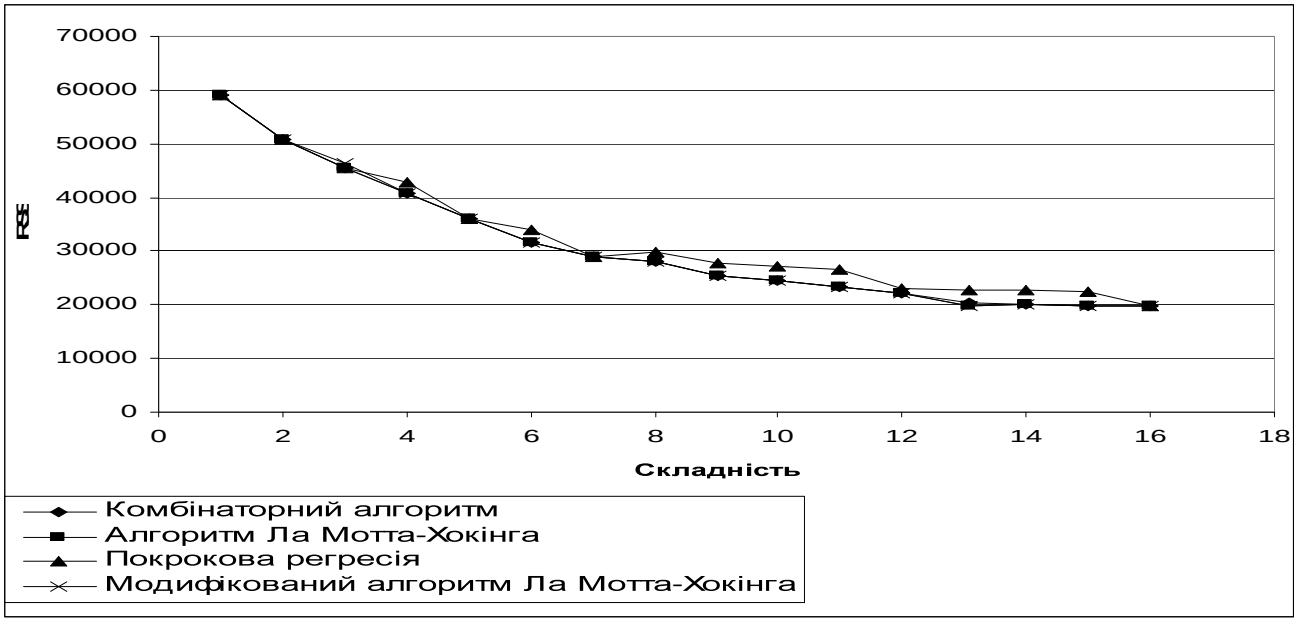

Рис.1. Порівняння точності методів на фіксованих рівнях склад-

Краще порівнювати точність результату для кожного рівня окремо. Оскільки краща модель для всієї множини структур може бути знайдена перебором кращих моделей на кожному рівні складності окремо, і результат є точним для окремих рівней складності, то він буде точним і для повного перебору.

Як видно з рис. 1, для малих складностей модифікація методу Ла Мотта-Хокінга і покрокова регресія не є точними методами. Точність модифікованого методу Ла Мотта-Хокінга може бути покращена підбором параметра *k*. Можливий алгоритм підбору:

1. загальна формула для кращого значення параметру:

$$k = ((m\text{-}s)/2) \pm 2 \qquad (4)$$

2. якщо кількість структур складності $k \geq 2^{20} - 1$, метод не буде ефективним.

З кроку 1 зрозуміло, що алгоритм не буде ефективним для малих складностей (наприклад, якщо треба знайти модель складності 3 для загальної кількості аргументів 20, при повному переборі буде згенеровано 1140 структур, для метода Ла Мотта-Хокінга спочатку будуть побудовані всі структури складності (20-3)/2=8, тобто 125 970 структур).

Порівняння кількості згенерованих моделей. Кількість згенерованих моделей є основним показником швидкості методу. Для повного перебору кількість моделей визначається як:

$$C_m^s = \frac{m!}{(m-s)!s!}, \tag{5}$$

де *m*-загальна кількість аргументів, *s*-складність моделі

Максимальне значення C_m^s досягається при *s=m/2*. Для методу Ла Мота-Хокінга кількість структур не може бути передбаченим. На рис. 2 показана середня кількість переглянутих структур для 1000 різних структур. Пошук проводився на окремих рівнях складності.

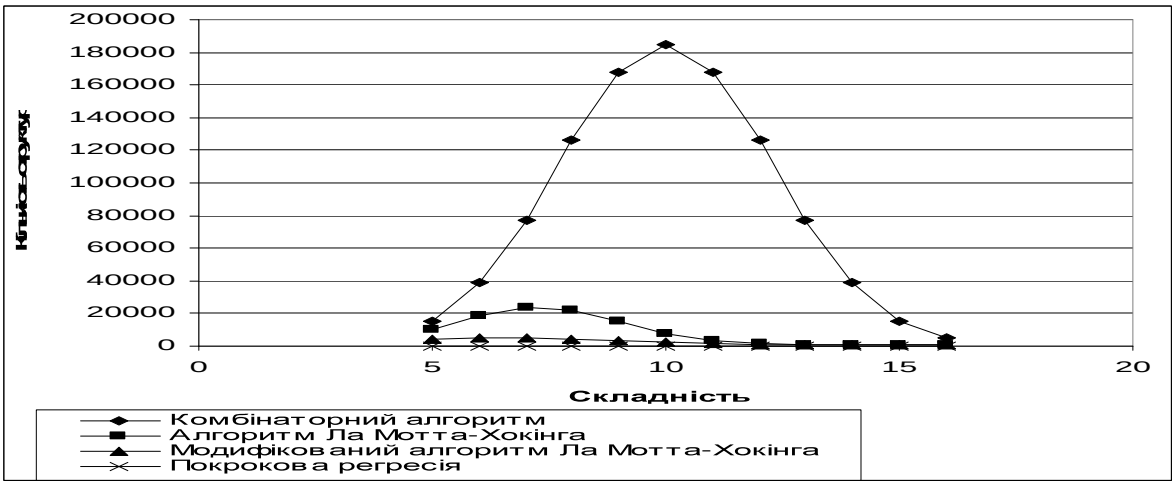

Рис.2 Кількість переглянутих структур для пошуку на окремих рівнях складності

Модифікація методу Ла Мотта-Хокінга є швидшою за стандартний метод Ла Мотта-Хокінга для складностей, близьких до *m/2*. Результати для повного перебору наведено на рис.3.

Рис.3. Кількість структур для повного перебору

Видно, що для повного перебору модифікація методу Ла Мота-Хокінга не є набагато швидшою за стандартний метод. Але для великих складностей є помітне зменшення структур.

Порівняння часу роботи методів. Час роботи методу також є важливим критерієм швидкості метода, оскільки в ньому враховані всі додаткові обчислення, проведені в методі На рис. 4, 5 показано порівняння часу роботи для досліджених методів.

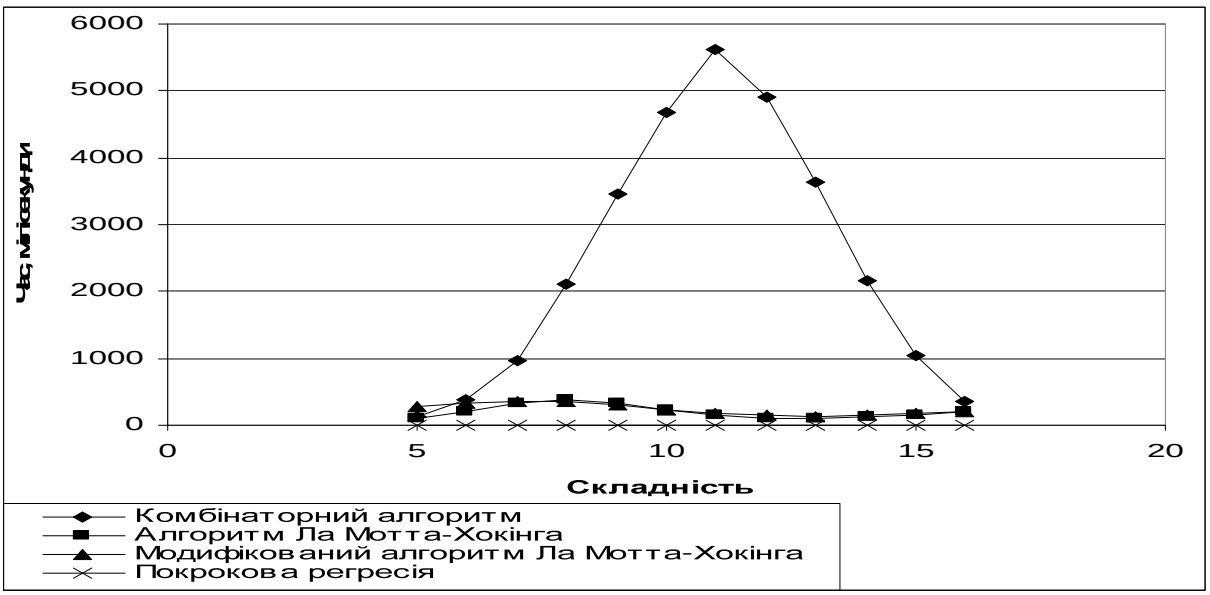

Рис.4 Порівняння часу роботи для окремих рівней складності

Рис.5 Порівняння часу роботи при повному переборі

Графіки є пропорційними графікам кількості переглянутих структур, отже можна зробити висновок, що в методах немає додаткових обчислень, сповільнюючи методи.

Крім того, модифікація методу Ла Мотта-Хокінга не є швидшою за стандартний метод, отже, деяка необхідна додаткова оптимізація обчислень в методі.

Порівняння кількості згенерованих структур для різних значень параметра s. Швидкість методу Ла Мотта-Хокінга залежить також і від значення параметра *k*. Якщо задано велике значення параметра, пошук в групі проходить швидко, але кількість *k*-моделей, що треба згенерувати, теж зростає дуже швидко. Навпаки, якщо задано мале значення параметра *k*, *k*-моделі будуть згенеровані швидко, але пошук в кожній групі буде повільним. Крім того, при завданні параметру *k* треба враховувати, що для завершення пошуку модель, де буде виключено *k* аргументів, буде порівнюватись з моделлю, де виключено більше аргументів. При дуже великій різниці пошук теж сповільниться.

На рис. 6 зображена залежність між різними значеннями параметру *k* та кількістю згенерованих моделей.

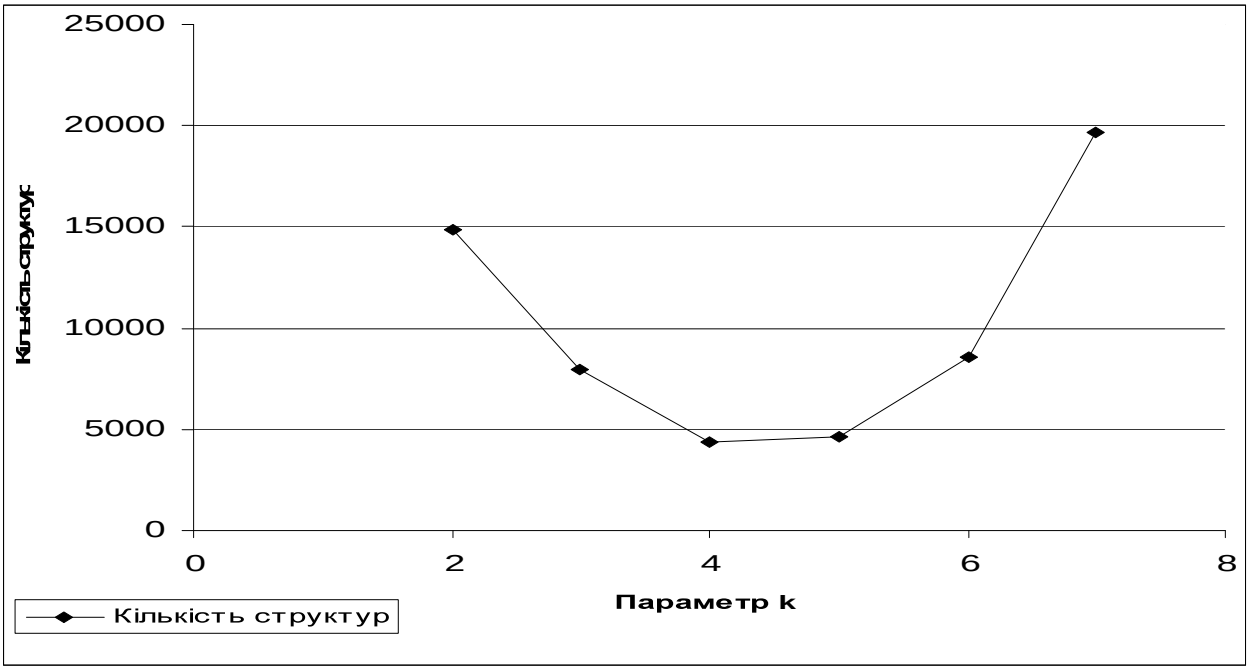

Рис.6 Порівняння кількості структур для різних значень параметра *k*

З рис. 6 видно, що оптимальним значенням параметру *k* є значення близьке до *(m-s)/2*.

Дослідження ефективності методу Ла Мотта-Хокінга та його модифікації. Попередні дослідження дають інформацію щодо точності методів. Швидкість методів для великої кількості аргументів теж є важливою. Дослідження ефективності проводилось наступним чином: максимальна кількість структур досягається при *s=m/2*. Для різних значень m були знайдені структури такої складності. Якщо кількість переглянутих структур більша за $S \geq C_{20}^{10} = 184756$, метод є неефективним. Результати дослідження представлені на рис.7.

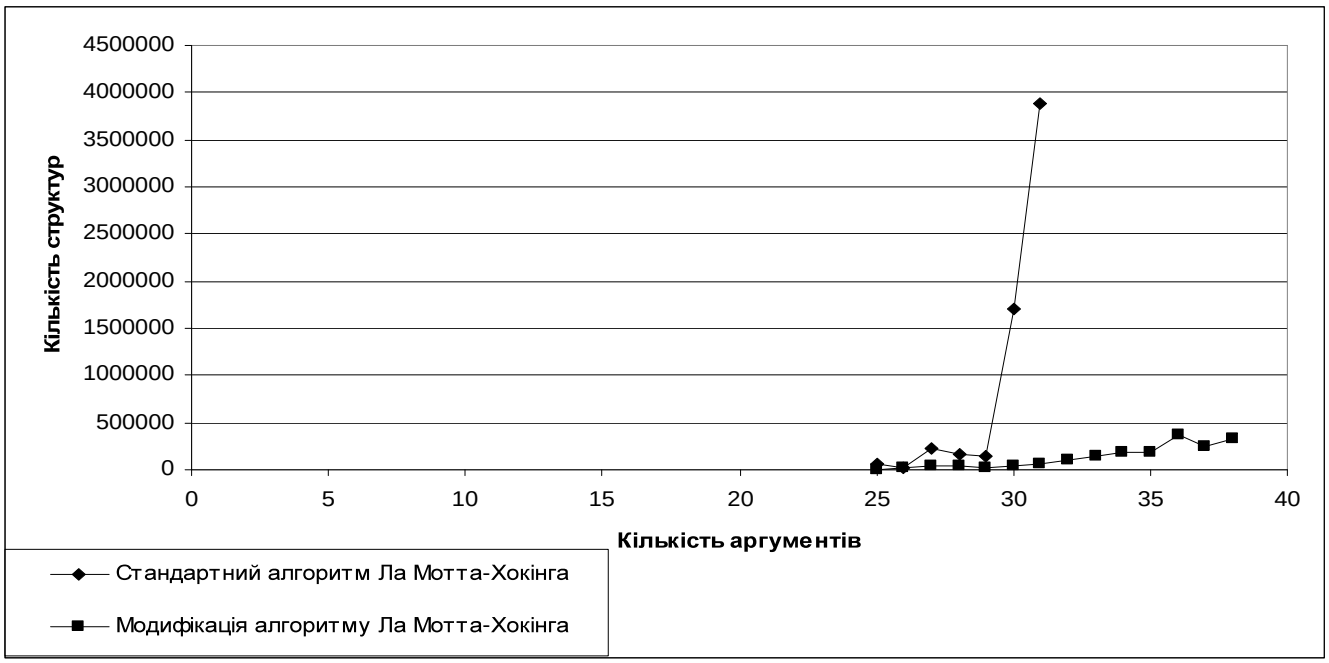

Рис.7 Порівняння кількості згенерованих структур для великої кількості аргументів

Кількість згенерованих структур для модифікованого методу Ла Мотта-Хокінга є набагато меншою, ніж кількість структур для стандартного методу. Крім того, збільшення кількості структур є набагато повільнішим.

Висновки

Результати дослідження методів пошуку кращої структури були представлені в статті. Детальні результати для кожного методу наведені нижче.

Метод Ла Мотта-Хокінга. Метод є точним, результат завжди співпадає з результатом повного перебору, але для малих складностей метод є повільнішим за повний перебір. Є ефективним для складностей до 30 аргументів. Основна причина сповільнення метода – велике збільшення кількості k-моделей.

Модифікація методу Ла Мотта-Хокінга. Метод не завжди дає той же результат, що і повний перебір. Для покращення точності необхідно додатково підбирати параметр *k*. Метод є швидшим за стандартний для складностей близьких до *m/2*. Метод є порівняно швидким для складностей до 37-40 аргументів.

Оптимальний алгоритм пошуку кращої структури може бути знайдений шляхом використання різних методів, в залежності від кількості структур, що треба переглянути та загальної кількості аргументів.

Література

1. Statistical Methods for Digital Computers. Edited by Kurt Enslein, Anthony Ralston and HERBERT S. Wilf. John Wiley, New York.