Twierdzenie o schematach

Próbą wyjaśnienia dużej skuteczności algorytmów genetycznych w roli uniwersalnej metody odnajdywania optymalnego rozwiązania bez znajomości charakterystyki zadania jest koncepcja schematu. Schemat jest to wzorzec opisujący podzbiór ciągów podobnych ze względu na ustalone pozycje. Użycie takiej definicji schematu wymaga dodania do alfabetu genów specjalnego symbolu * (nieistotne) reprezentującego wartość nieokreśloną. Schemat reprezentuje wszystkie łańcuchy będące podzbiorem przestrzeni poszukiwań, które zgadzają się z nim na wszystkich pozycjach innych niż *.

Pojęcie schematu jest obowiązujące dla alfabetu genów o dowolnej długości. Przy założeniu, że ciągi kodowe reprezentujące genotypy składają się z symboli alfabetu dwójkowego, na każdej pozycji schematu może pojawić się każdy symbol należący do zbioru V = { 0, 1, * }. Na przykład schemat H1 = (*111100100) złożony z symboli alfabetu V może być reprezentowany przez dwa następujące ciągi:

{ (0111100100),(1111100100) }

natomiast reprezentantami schematu H2 = (*1*1100100) są wyłącznie cztery następujące ciągi:

{ (0101100100),(0111100100),(1101100100),(1111100100)}.

Każdy schemat reprezentowany jest przez 2^r ciągów binarnych, gdzie r jest liczbą symboli nieistotne * w szablonie schematu. Z drugiej strony każdy łańcuch binarny o długości m jest reprezentantem 2^m schematów. Na przykład łańcuch binarny (101001) jest reprezentantem 26 schematów:

(1 0 1 0 0 1)

(* 0 1 0 0 1)

(1 * 1 0 0 1)

(1 0 1 0 0 *) (* * 1 0 0 1) (* 0 * 0 0 1) (1 0 1 0 * *) (* * * 0 0 1)

(* * * * * *

Dla łańcuchów o długości m istnieje 3^m możliwych schematów. W populacji o liczebności n może być reprezentowanych od 2^m do n x 2^m różnych schematów. Już te proste obliczenia dają pewne wyobrażenie o ogromie informacji przetwarzanej przez algorytmy genetyczne; aby jednak zrozumieć schematy – cegiełki, z których powstają przyszłe rozwiązania trzeba wyodrębnić różne typy schematów. Do sformułowania twierdzenia o schematach potrzebne będą dwa atrybuty charakteryzujące schemat: rząd i rozpiętość nazywana również długością definiującą.

Rzędem schematu H, oznaczonym przez o(H) jest nazywana liczba ustalonych pozycji we wzorcu. Dla ciągów binarnych jest to liczba zer i jedynek. Np. o(011*1**) = 4 oraz o(0******) = 1.

Rozpiętość schematu H, którą oznaczamy przez 5(H), to inaczej odległość między dwiema skrajnymi pozycjami ustalonymi. Tak więc, schemat (011*1**) ma rozpiętość równą 4, gdyż jego ostatnia ustalona pozycja ma numer 5, a pierwsza – numer 1, zatem odległość między nimi wynosi 5 – 1 = 4. Schemat który ma tylko jedną ustaloną pozycję, taki jak (0*******), ma rozpiętość 5 = 0.

Pierwszym i najłatwiejszym elementem teorii schematów jest ustalenie wpływu selekcji, czy inaczej reprodukcji na oczekiwaną liczbę reprezentantów określonego schematu. Wyrażenie m = m(H, t) oznacza liczbę reprezentantów schematu H w chwili t. Podczas selekcji genotypy trafiają do puli rodzicielskiej z prawdopodobieństwem proporcjonalnym do wskaźnika przystosowania, a dokładnie z prawdopodobieństwem równym:

$$p_i = f_i / \sum f_i \quad (1.4.)$$

Po skompletowaniu puli rodzicielskiej nowego pokolenia złożonej z n ciągów wybranych w chwili t, można oczekiwać obecności m = m(H, t +1) reprezentantów schematu H w chwili t +1, przy czym zachodzi zależność:

$$E[m(H,t+1)] = m(H,t) \cdot n \cdot f(H) / \sum f_i \quad (1.5.)$$

Liczba f(H) określa średnie przystosowanie ciągów będących reprezentantami schematu H w chwili t. Po uwzględnieniu ilorazu sumy wartości przystosowania wszystkich chromosomów populacji i ich liczby jako średniego przystosowania całej populacji t oraz pominięciu symbolu wartości oczekiwanej po lewej stronie równania 1.5. otrzymuje się słuszną dla dużych populacji aproksymację:

$$m(H,t+1) = m(H,t) \frac{f(H)}{\overline{f}}$$

Oznacza to, że liczebność reprezentacji danego schematu w następnym pokoleniu zmienia się proporcjonalnie do stosunku średniego przystosowania schematu i średniego przystosowania całej populacji; a zatem schematy o przystosowaniu wyższym niż średnie populacji będą miały większą liczbę reprezentantów w następnym pokoleniu, podczas gdy schematy o przystosowaniu niższym niż średnie z populacji otrzymają mniej reprezentantów niż miały dotychczas.

Przy założeniu, że pewien schemat H przewyższa średnią o jakąś jej stałą część równą c, równanie schematów można zapisać tak jak na wzorze 1.7., a startując od t=0 i zakładając, że c nie zmienia się w czasie, otrzymuje się zależność przedstawiona na wzorze

$$m(H,t+1) = m(H,t)(\overline{f} + c\overline{f})/\overline{f} = (1+c)\cdot(H,t)$$

$$m(H,t) = m(H,0)\cdot(1+c)^t \qquad (1.7.)$$

(1.8.)

Otrzymana ocena ilościowa efektu reprodukcji pozwala stwierdzić, że schematy lepsze od przeciętnej są wybierane w liczbie przypadków rosnącej wykładniczo w czasie, natomiast gorsze od przeciętnej wybierane są w liczbie wykładniczo malejącej w czasie.

Kolejnym krokiem jest zbadanie wpływu operatora krzyżowania na oczekiwaną liczbę reprezentantów określonego schematu. Operacja krzyżowania dokonuje uporządkowanej wymiany informacji pomiędzy dwoma chromosomami. Przy pomocy krzyżowania powstają nowe struktury w sposób minimalnie zaburzający wzorzec propagacji dyktowany przez reprodukcję. W końcowym efekcie frekwencje różnych schematów reprezentowanych w populacji rosną lub maleją wykładniczo w kolejnych pokoleniach.

Oszacowanie wrażliwości schematu na krzyżowanie rozpocząć można od rozważenia przykładowego ciągu kodowego A o długości l = 33 oraz dwóch reprezentatywnych schematów H1 i H2 , pasujących do tego ciągu:

A = (11101111101000100011000000100011 0)
H1 = (****111**************************)
H2 = (111**************************10)

Przy założeniu, że łańcuch został wybrany do krzyżowania i losowo wybrano punkt cięcia na pozycji 20, łatwo można zauważyć, iż schemat H1 przetrwał krzyżowanie, czyli że jeden z potomków ciągu A nadal jest reprezentantem schematu, natomiast drugi schemat nie przetrwał. Przyczyna leży w tym, że punkt cięcia w pierwszym schemacie nie trafił pomiędzy

ustalone pozycje, natomiast ustalone pozycje 111 na początku szablonu H2 oraz 10 na jego końcu znalazły się w różnych potomkach. Jest oczywiste, że rozpiętość schematu spełnia istotną rolę w prawdopodobieństwie jego przetrwania lub zniszczenia.

Dla każdego schematu można podać dolne oszacowanie prawdopodobieństwa przeżycia podczas krzyżowania. Kiedy punkt krzyżowania nie trafia pomiędzy ustalone pozycje schemat przeżywa krzyżowanie, więc prawdopodobieństwo przeżycia ps wynosi co najmniej 1 – 5(H)/(l-1) . Jeśli dla określonej pary ciągów sama operacja krzyżowania jest wykonywana z pewnym prawdopodobieństwem pc, to prawdopodobieństwo przeżycia schematu H spełnia nierówność:

$$p_s = 1 - p_c \cdot \frac{\delta(H)}{l-1}$$

która sprowadza się do poprzedniego oszacowania, gdy $p_c = 1$.

Przy założeniu niezależności operacji selekcji i krzyżowania średnia liczba reprezentantów danego schematu H w następnym pokoleniu po łącznym efekcie działania obu operatorów wynosi:

$$m(H,t+1) \geq m(H,t)\frac{f(H)}{\overline{f}}\left[1 - p_c \cdot \frac{\delta(H)}{l-1}\right]$$

Analiza powyższego wzoru pozwala wnioskować, że schematy, które mają przystosowanie wyższe od średniej i małą rozpiętość będą rozprzestrzeniać się zgodnie z wykładniczym prawem wzrostu.

Ostatnim elementem wchodzącym w skład teorii o schematach jest analiza wpływu na wzrost lub spadek reprezentacji schematów operatora mutowania. Operator ten zmienia zawartość każdej pozycji z prawdopodobieństwem równym pm. Aby schemat H przeżył mutację, muszą się zachować wszystkie jego pozycje ustalone. A

zatem, ponieważ pojedyncza pozycja chromosomu pozostaje bez zmian z prawdopodobieństwem $(1 - p_m)$ i ponieważ mutacje na poszczególnych pozycjach są statystycznie niezależne, dany schemat H przeżyje mutację z prawdopodobieństwem $(1 - p_m)^{o(H)}$, gdzie o(H) jest liczbą pozycji ustalonych. Dla stosowanych w programach ewolucyjnych małych wartości p_m, prawdopodobieństwo przeżycia można aproksymować za pomocą wyrażenia $1 - o(H) \times p_m$. Można więc ostatecznie stwierdzić, że oczekiwana liczba reprezentantów schematu H w następnym pokoleniu, otrzymanym w wyniku reprodukcji, krzyżowania i mutacji, spełnia następującą nierówność:

$$m(H,t+1) \geq m(H,t)\frac{f(H)}{\overline{f}}\left[1 - p_c\frac{\delta(H)}{l-1} - o(H)p_m\right]$$

Wzór 1.11. pozwala na sformułowanie Twierdzenia o schematach albo Podstawowego twierdzenia algorytmów genetycznych: schematy dobrze przystosowane, niskiego rzędu i o małej rozpiętości rozprzestrzeniają się w kolejnych pokoleniach zgodnie z wykładniczym prawem wzrostu.