

پردازش تكاملی

برنامه ریزی ژنتیک Genetic Programming

دانشگاه صنعتی مالک اشتر

مجتمع دانشگاهی فن آوری اطلاعات و امنیت

زمستان ۱۳۹۲

مروری سریع بروی برنامه ریزی ژنتیک

- توسعه اولیه: آمریکا ۱۹۹۰
- اولین محققان: J.koza
- استفاده به منظور: وظایف یادگیری ماشین (پیشگویی، کلاسه بندی و ...)
- ویژگی ها :
 - رقیبی برای شبکه های عصبی و کاربردهای مشابه آن
 - نیاز به جمعیت زیاد (هزاران)
 - سرعت پایین
- خصوصیات:
 - کروموزوم های غیر خطی: درختها، گرافها
 - امکان جهش اما نه به صورت ضروری (مورد بحث و اختلاف نظر)

خلاصه و درونمایی از ویژگی های فنی برنامه ریزی ژنتیک

ارائه	ساختار درختی
ترکیب مجدد	مبادله در زیر درخت ها
جهش	تغییر تصادفی در درخت ها
انتخاب والد	متناسب سازی همزمان
انتخاب بازمانده	جایگزینی نسل

مثال مقدماتی: امتیاز دهی اعتبار

- بانک میخواهد وجه تمایزی میان متقاضیان وام خوب و بد داشته باشد.
- مدل ضرورتاً به تاریخچه ای از داده ها نیاز خواهد داشت.

ID	No of children	Salary	Marital status	OK?
ID-1	2	45000	Married	0
ID-2	0	30000	Single	1
ID-3	1	40000	Divorced	1
...				

مثال مقدماتی: امتیاز دهی اعتبار

- یک مدل امکان پذیر:

IF (NOC = 2) AND (S > 80000) THEN good ELSE bad

- در حالت عمومی:

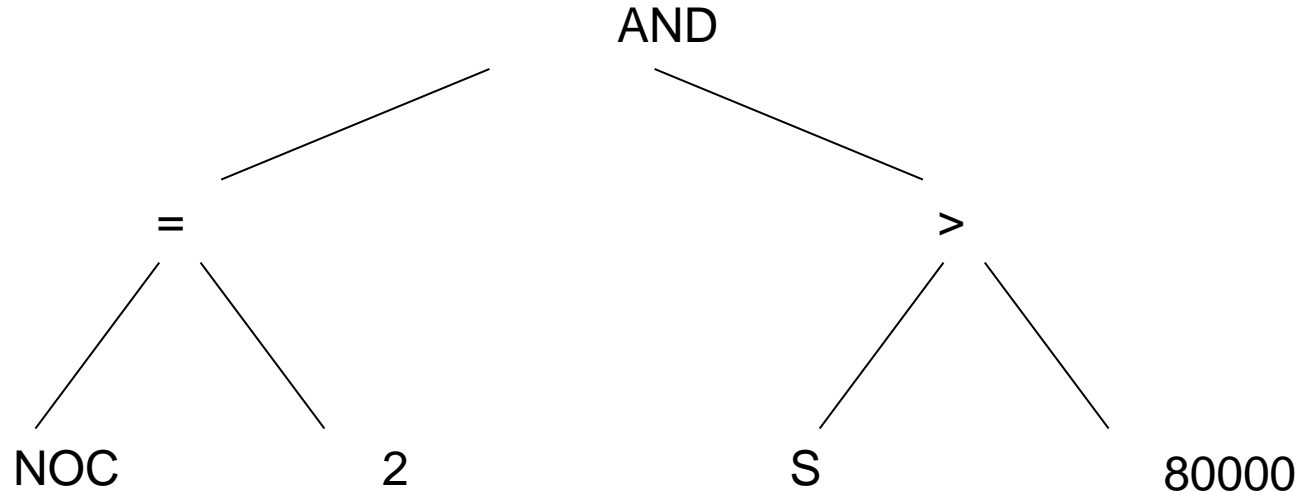
IF formula THEN good ELSE bad

- تنها مجهول بخش سمت راست میباشد، بدین ترتیب
- فضای جستجوی ما (phenotypes) مجموعه ای از فرمول ها میباشد.
- شایستگی عصبی از یک فرمول: درصد توانایی کلاسه بندی مناسب موردها در مدل میباشد که پایه ای است به منظور
- پارس (تجزیه شدن): ارائه عصبی از فرمولها (genotypes) درخت ها می باشند.

مثال مقدماتی: امتیاز دهی اعتبار

IF (NOC = 2) AND (S > 80000) THEN good ELSE bad

به کمک درخت زیر میتوان آنرا ارائه نمود



ارائه مبتنی بر ساختار درختی

- درخت های یک ساختار همه منظوره میباشند.

- فرمول ریاضی

$$2 \cdot \pi + \left((x+3) - \frac{y}{5+1} \right)$$

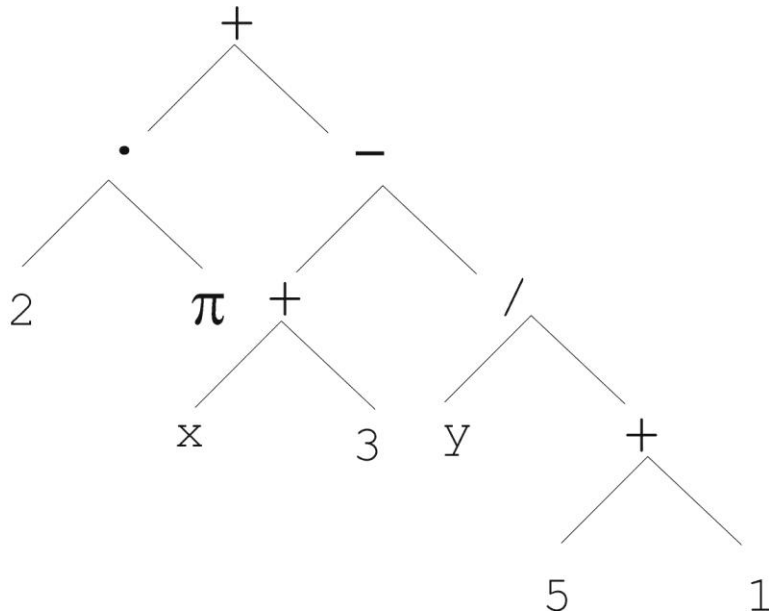
- فرمول منطقی

$$(x \wedge \text{true}) \rightarrow ((x \vee y) \vee (z \leftrightarrow (x \wedge y)))$$

- برنامه

```
i = 1;  
while (i < 20)  
{  
  i = i + 1  
}
```

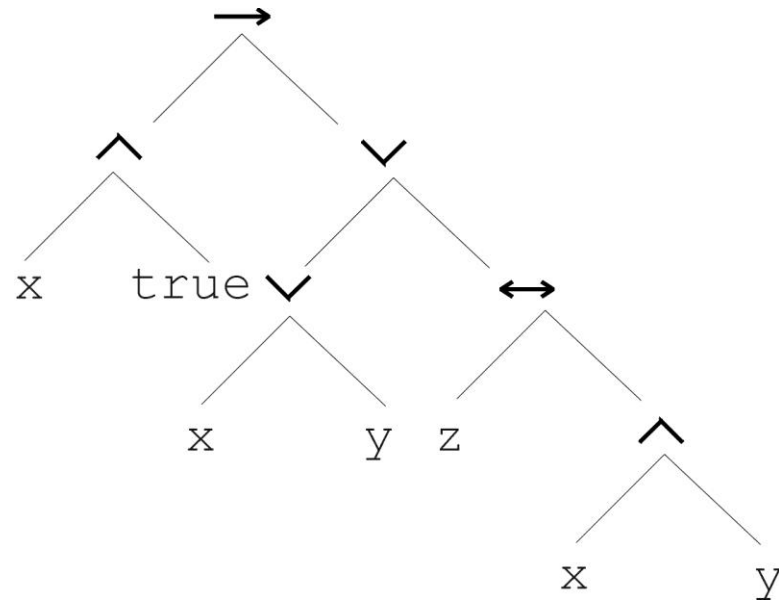
ارائه مبتنی بر ساختار درختی



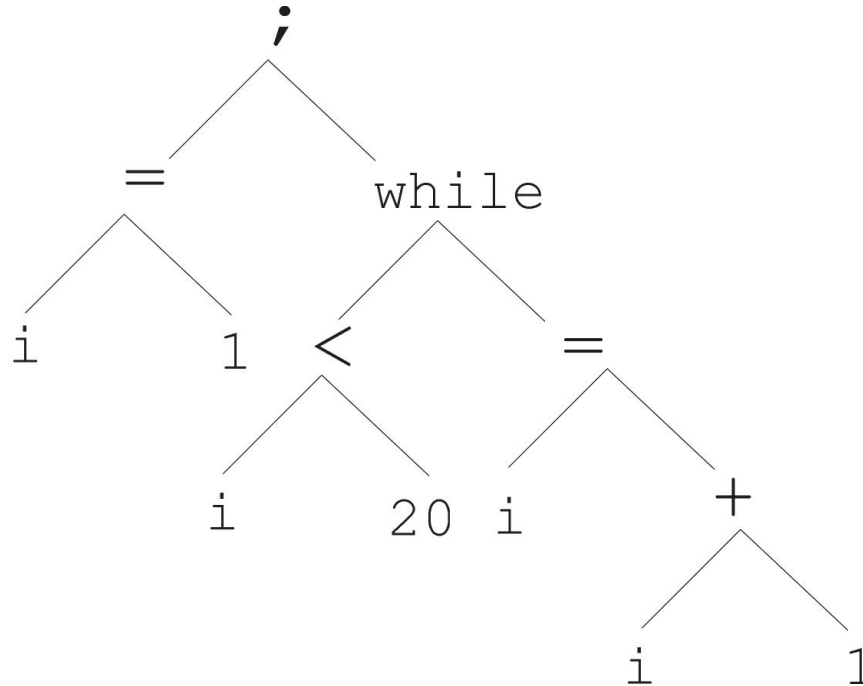
$$2 \cdot \pi + \left((x + 3) - \frac{y}{5 + 1} \right)$$

ارائه مبتنی بر ساختار درختی

$$(x \wedge \text{true}) \rightarrow ((x \vee y) \vee (z \leftrightarrow (x \wedge y)))$$



ارائه مبتنی بر ساختار درختی



```

i = 1;
while (i < 20)
{
i = i + 1
}
    
```

ارائه مبتنی بر ساختار درختی

- در GA, ES, EP کروموزوم ها ساختار خطی دارند (رشته بیت، رشته اعداد صحیح، بردارهای مقادیر حقیقی، جایگشت ها)
- در ساختار درختی کروموزومها ساختار غیر خطی دارند.
- در GA, ES, EP اندازه کروموزوم ها ثابت میباشد.
- درخت ها در GP ممکن است از نظر عمق و پهنا بسیار متغیر باشند.

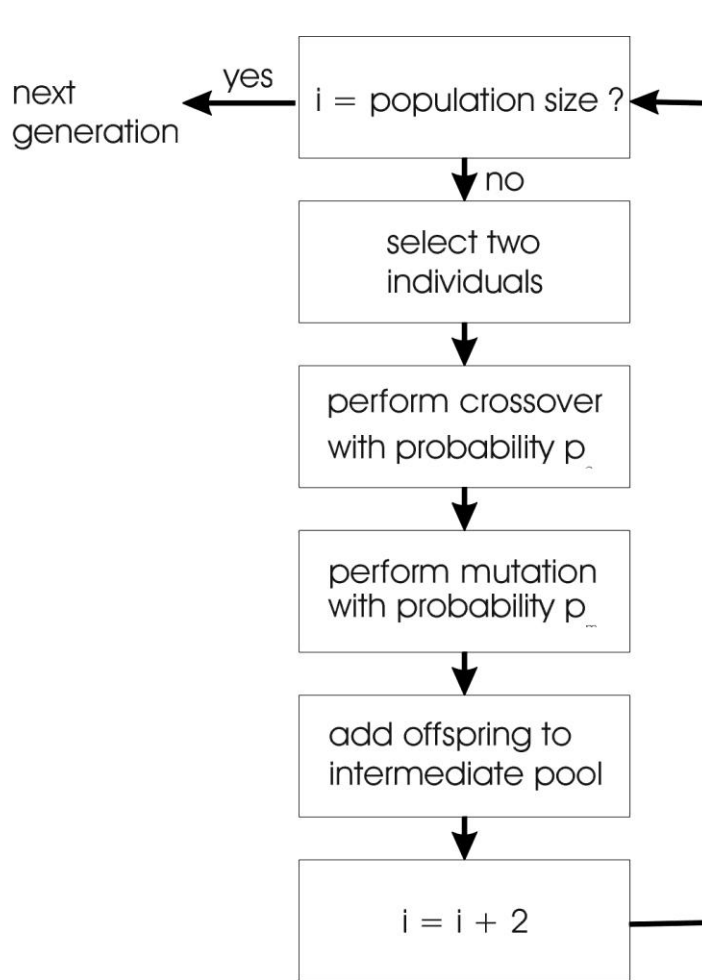
ارائه مبتنی بر ساختار درختی

- نمایش عبارات میتواند توسط موارد ذیل تعریف شود:
 - مجموعه ترمینال (T)
 - مجموعه توابع (F): همراه با نمادهای تابع
- پذیرفتن تعریف بازگشتی زیر:
 1. Every $t \in T$ is a correct expression
 2. $f(e_1, \dots, e_n)$ is a correct expression if $f \in F$, $\text{arity}(f)=n$ and e_1, \dots, e_n are correct expressions
 3. اشکال دیگری از عبارات صحیح وجود ندارد.
- در حالت کلی عبارات در GP نوشته نمی شوند.
(closure property any $f \in F$ can take any $g \in F$ as argument)

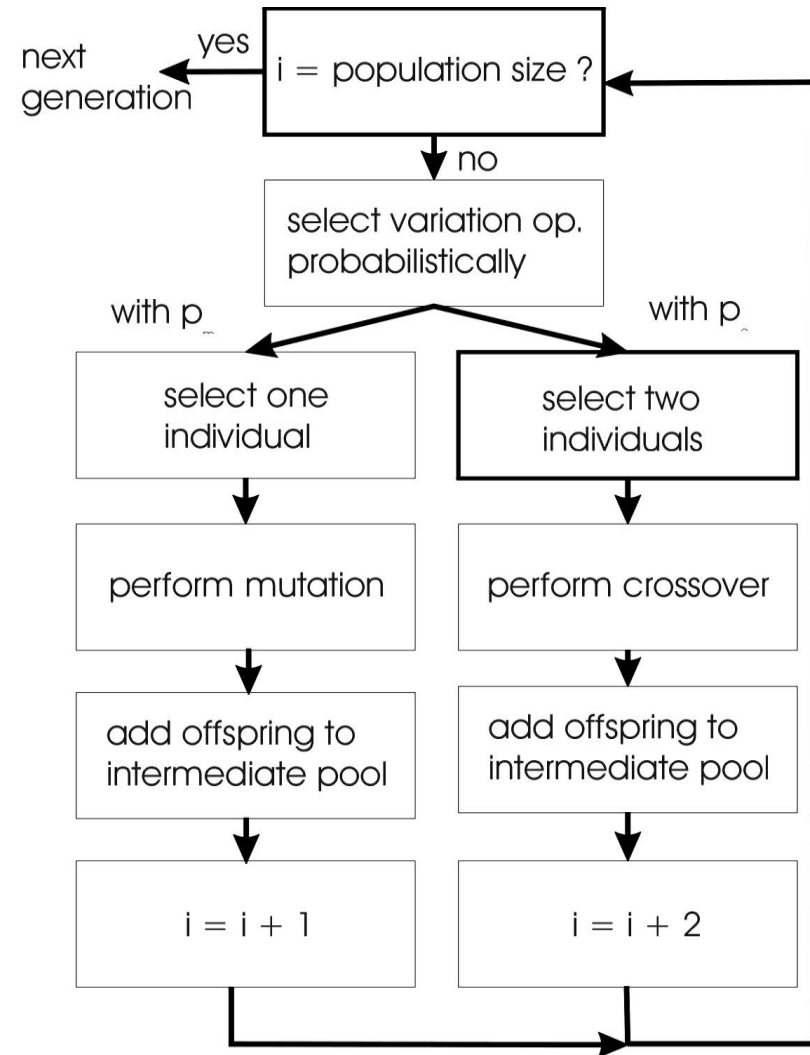
طرح ایجاد زاد ولد ها

مقایسه:

- طرح GA از توالی جهش AND های متقاطع استفاده میکند (be it probabilistically)
- طرح GP از جهش OR متقاطع استفاده میکند (chosen probabilistically)

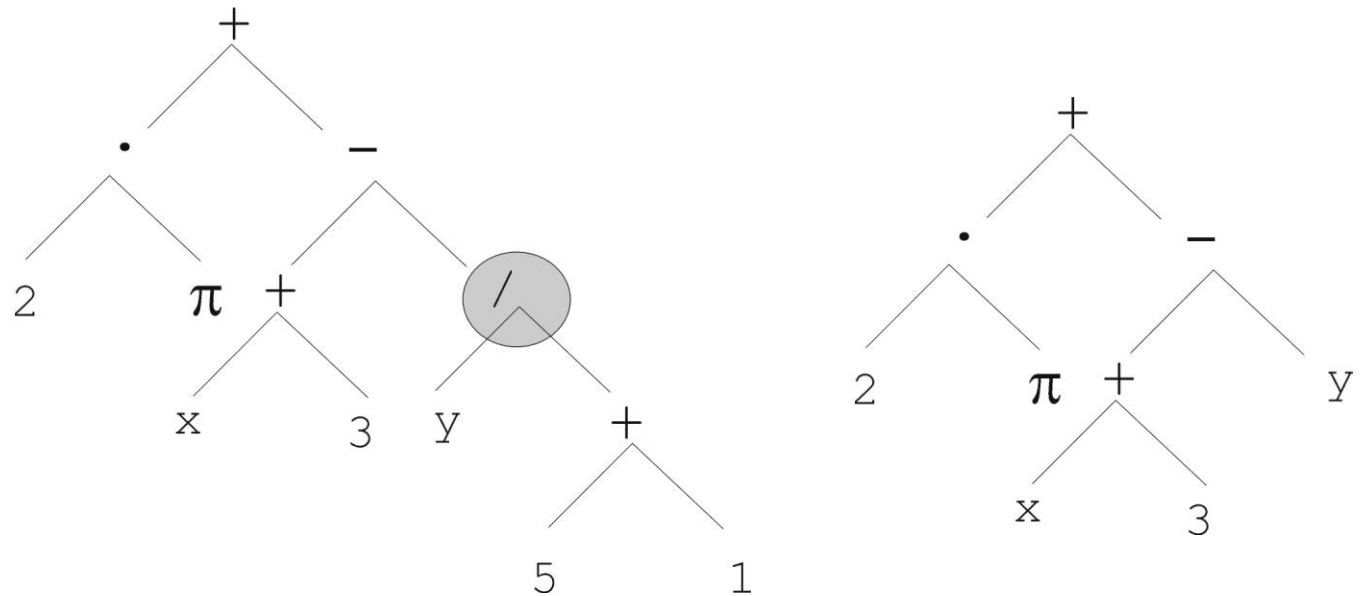


GA flowchart



GP flowchart

- اغلب جهش های معمول: زیر درخت انتخابی را به صورت تصادفی با درخت توسعه یافته جایگزین میکند.

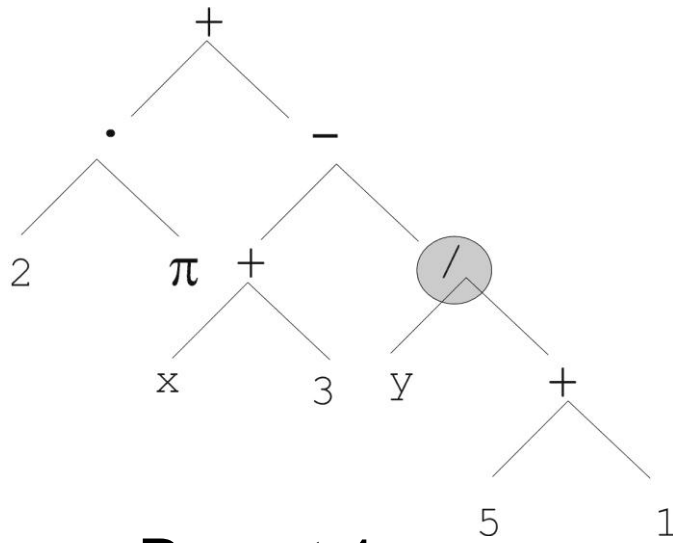


جهش cont'd

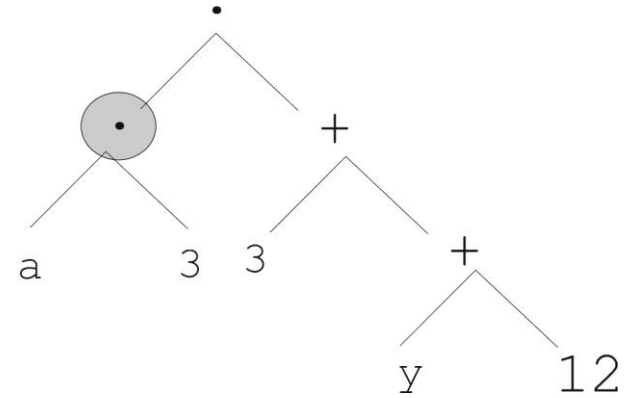
- جهش دارای دو پارامتر میباشد:
 - P_m احتمالی به منظور انتخاب جهش در مقایسه با ترکیب مجدد
 - احتمال در انتخاب یک نقطه داخلی به عنوان ریشه زیر درخت جایگزین شده.
- در مورد P_m د به طور قابل ملاحظه تاکید بر مقدار 0 شده است (Koza'92) و یا اینکه با مقادیر کوچک همانند 0.05 در نظر گرفته شود (Banzhaf et al. '98).
- اندازه فرزند میتواند از سایز والدین تجاوز کند.

ترکیب مجدد

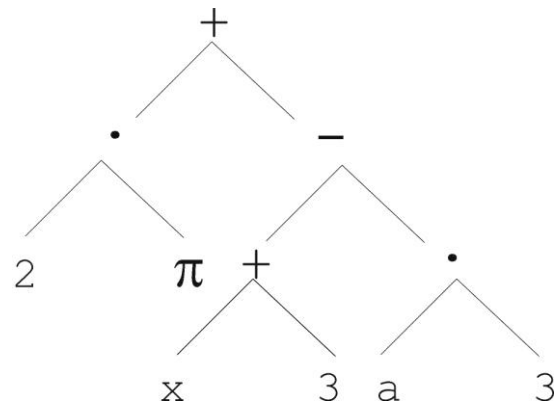
- متداول ترین حالت باز ترکیب: تعویض دو زیر درخت تصادفی انتخاب شده در میان والدین میباشد.
- باز ترکیب دارای دو پارامتر میباشد:
 - مقدار احتمالی p_c به منظور انتخاب باز ترکیب در مقایسه با جهش
 - احتمال انتخاب یک نقطه داخلی در داخل هر والد به عنوان نقطه تقاطع
- اندازه زاد و ولدها نیز میتواند از والدین تجاوز نماید.



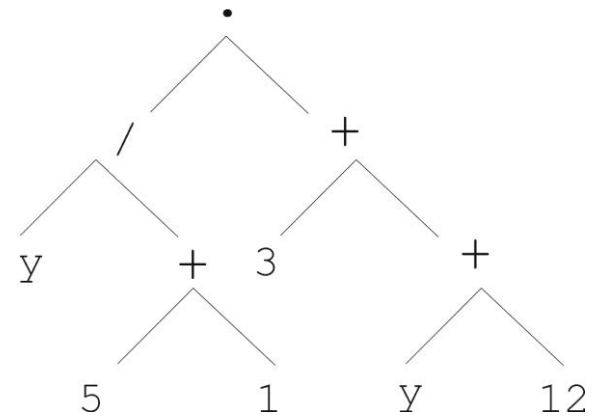
Parent 1



Parent 2



Child 1



Child 2

انتخاب

- انتخاب والدین نوعاً بر اساس تناسب است
- انتخاب مجدد در جامعت های بسیار بزرگ:
 - تعیین رتبه جمعیت با تناسب و تقسیم آن به دو گروه
 - گروه اول : بهترین $x\%$ از جمعیت و گروه دوم $(100-x)\%$
 - 80% عملیات های انتخاب از گروه ۱ انتخاب میکند و 20% از گروه دوم
 - برای جمعیت های با سایز $1000, 2000, 4000, 8000$ مقدار $x = 32\%, 16\%, 8\%, 4\%$
 - جهش: به منظور افزایش راندمان، درصد ها به صورتی سر انگشتی حساب شده است.
- انتخاب باز ماندگان:
 - نمونه : شمای گسترش (thus none)
 - Recently steady-state is becoming popular for its elitism

مقدار دهی اولیه

- بیشترین مقدار اولیه عمق درخت ها D_{max} مقدار دهی میشود .
- شیوه کامل (عمق هر شاخه برابر است با D_{max})
 - گره های واقع در $d < D_{max}$ به صورت تصادفی از مجموعه توابع F انتخاب میشود.
 - گره های واقع در $d = D_{max}$ به صورت تصادفی از مجموعه ترمینال T انتخاب میشود.
- شیوه ی رویدن (هر شاخه درای عمق $depth \leq D_{max}$ میباشد.
 - گره های واقع در عمق $d < D_{max}$ به صورت تصادفی از $F \cup T$ انتخاب میشوند.
 - گره های واقع در عمق $d = D_{max}$ به صورت تصادفی از T انتخاب میشود.
- مقدار دهی اولیه معمول GP: به صورت نیمه به نیمه بالا میرود تا زمانی روش های رشد نمودن و کامل شدن هر کدام نیمی از جمعیت اولیه را بدست میدهد.

Bloat چاقی

- چاقی : ” رهایی جستن از چاق ترین “ به معنی اینکه اندازه های درخت در یک جمعیت و در طول زمان در حال افزایش میباشد.
 - موضوعی است که در تحقیقات اخیر مورد بحث بوده و به دنبال دلایل وقوع آن میباشد.
 - نیاز به اقدامی متقابل خواهد داشت.
 - خود داری از به کار گیری عملگرهایی که فرزندان بسیار بزرگ را تحویل داده و بوجود خواهند آورد.
- فشار صرفه جویی: مجازاتی است برای مقابله با ازدیاد ساز

مسائلی که محیط های فیزیکال را درگیر میکند

- درخت هایی به منظور جاسازی داده در مقایسه با درخت هایی و یا برنامه هایی که واقعا قابل اجرا میباشد.
- اجرا میتواند محیط را دچار تغییر سازد در نتیجه محاسبه تناسب
- مثال: کنترل کننده روبات
- محاسبات تناسب اغلب به وسیله شبیه سازی انجام میشود که بازه ای از پرهزینه تا بشدت پرهزینه (از نظر زمانی) میباشد.
- اما اغلب کنترلرها بسیار خوب هستند.

مثال کاربردی: رگرسیون نمادی

- برخی نقاط را در \mathbf{R}^2 , $(x_1, y_1), \dots, (x_n, y_n)$ مفروض میباشد.
 - تابع یافتن $f(x)$ s.t. $\forall i = 1, \dots, n : f(x_i) = y_i$
 - راه حل امکان پذیر GP
- $$err(f) = \sum_{i=1}^n (f(x_i) - y_i)^2$$
- باز نمایش به وسیله $F = \{+, -, /, \sin, \cos\}$, $T = \mathbf{R} \cup \{x\}$
 - Fitness is the error
 - تمامی عملگر های استاندارد است.
 - pop.size = 1000, ramped half-half initialisation
 - Termination: n “hits” or 50000 fitness evaluations reached
(where “hit” is if $|f(x_i) - y_i| < 0.0001$)

بحث و گفتگو

آیا GP:

هنر نمو دادن برنامه های کامپیوتری می باشد؟
وسیله ای است به منظور برنامه ریزی خودکار سازی شده
کامپیوتر ها؟

همان GA میباشد که به نوع دیگری باز ارائه شده است؟