La Programmation Génétique

BELBACHIR Assia
DEAU Raphaël
LENNE Renaud
SNOUSSI Jihene

Définition :

La programmation génétique est une méthode inspirée par la théorie de l'Evolution telle qu'elle a été définie par Darwin et notamment ses mécanismes biologiques. Elle a pour but de trouver des programmes qui répondent au mieux à une tache définie. Pour ce faire, elle permet à la machine d'apprendre, en utilisant un algorithme évolutionniste afin d'optimiser la population de programmes.

Historique :

Afin de bien comprendre d'où vient la programmation génétiques, nous allons tout d'abord identifier quelques date importante pour cette recherche :

- **1958 – Friedberg :** Mutation aléatoire d'instruction dans un programme génétique, attribution de « crédit » aux instructions des programmes les plus efficaces.
- **1963 – Samuel :** Utilisation du terme « machine learning » dans le sens de programmation automatique.
- **1966 – Fogel, Owen & Walsh :** Automates à états finis pour des tâches de prédiction, obtenus par sélection de parents efficaces auxquels on applique des mutation : « evolutionary programming »
- **1985 – Cramer :** Utilisation d'expression sous forme d'arbre. Cross-over entre sous-arbres.
- **1986 – Hicklin :** Evolution de programmes de jeux en LISP. Sélection des parents efficaces, combinaisons des sous-arbres communs ou présents dans un des parents et de sous-arbres aléatoires.
- **1989/1992 – Koza :** Systématisation et démonstration de l'intérêt de cette approche pour de nombreux problèmes.Definitions d'un paradigme standard dans le livre « Genetic programming. On the programming of computers by means of natural selection » [Koza, 1992]. Ce paradigme inclus plusieurs concepts : programmation structurée en expression arborescentes, définition d'une grammaire de langage, type de retour unique pour chaque fonction, définition des proportions de mutation et de cross-over pour chaque génération, etc.

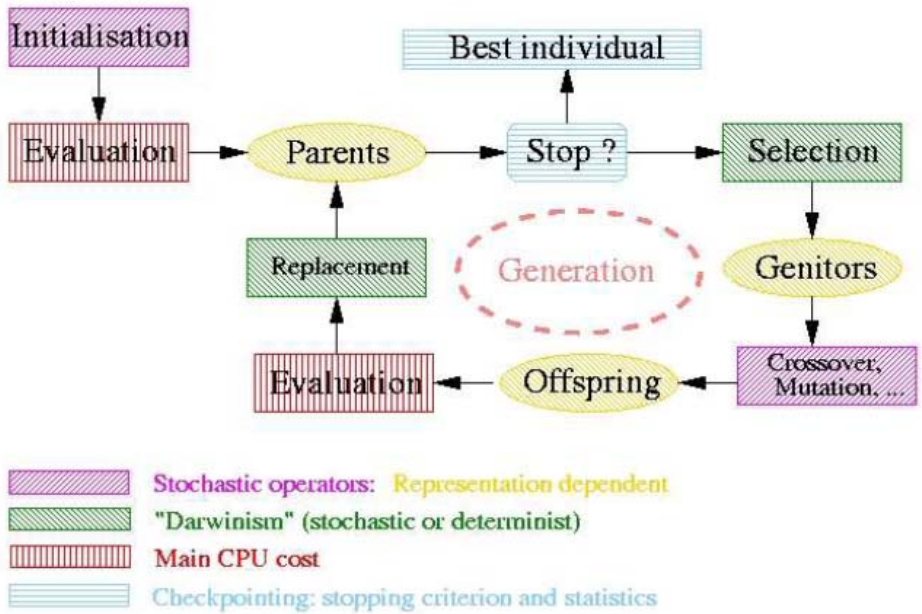
Applications de la programmation génétique :

Il y a de nombreuses applications de la programmation génétique :

- **problèmes de « magie noire »**, tels que la synthèse automatisée des circuits électriques analogues, des contrôleurs, des antennes, des réseaux des réactions chimiques, et d'autres secteurs de conception,
- **la « programmation de l'improgrammable »** comportant la création automatique des programmes machine pour les dispositifs de calcul peu usuels tels que les automates cellulaires, systèmes de multi-agent, systèmes parallèles, rangées de porte field-programmable, rangées field-programmable d'analogue, colonies de fourmi, l'intelligence d'essaim, à distribué des systèmes, et des semblables,

- **« nouvelles inventions commercialement utilisables »** (CUNI) comportant l'utilisation de la programmation génétique comme « machine d'invention » automatisée pour créer de nouvelles inventions commercialement utilisables.
- Reconnaissance d'images (Robinson et al.), classification d'images (Zao), traitement d'images satellites (Daïda), ...
- Prédiction de séries temporelles (Lee), génération d'arbres de décisions (Koza), datamining (Freitas), …
- Classification de segments d 'ADN (Handley), de protéines (Koza et al.), ...
- ……

Les phases de la programmation génétique :



Ce schéma représente le cycle de fonctionnement d'un programme génétique. Dans un premier temps, il nous faut une base pour pouvoir commencer a générer des programmes (phase initialisation). On obtiendra donc un certain nombre d'individus qui permettrons de générer les générations futures. A ce moment, on vérifie si l'une des solutions que nous offrent ces individus est satisfaisante (Bloc Evaluation). Si aucune solution ne convient, on va alors procéder a une sélection des meilleurs afin de générer par différentes techniques les descendants. (Phase sélection & crossover/mutation). Enfin, ces descendants vont venir remplacer la génération précédente en étant à leur tour les parents, et le cycle recommence alors par le bloc Evaluation.

1) Génération de la population initiale

Dans le cadre de la programmation génétique, la génération de la population initiale est une étape décisive et compliquée. L'espace de recherche étant infini, il est très difficile de répartir plus ou moins équitablement la population initiale sur l'espace de recherche, de sorte que cet espace soit au maximum parcouru.

Pour générer cette population initiale, la première chose à faire est de limiter la profondeur de l'arbre de recherche, limitant ainsi l'espace de recherche (qui ne sera donc plus infini) et assurant une convergence du processus.

Ensuite, un certain nombre d'individus doivent être générés aléatoirement (étant donné qu'il n'existe aucune méthode d'initialisation autre). Dans tous les cas, la génération d'un individu se déroulera selon un algorithme simple :

- une fonction *f* est choisie parmi l'ensemble des fonctions, elle sera la racine de l'arbre
- pour chaque fonction, chacun de ses arguments est choisi
  - soit parmi l'ensemble des fonctions, auquel cas, on recommence la boucle avec ses arguments
  - soit parmi l'ensemble des symboles terminaux. Dans ce cas, c'est une feuille de l'arbre, il n'y a pas à continuer pour cette branche.

A chaque étape, pour choisir dans quel ensemble seront choisis les arguments, trois méthodes sont à envisager :

1. **la méthode croissante :**
   A chaque niveau, le prochain symbole est pris aléatoirement soit dans les symboles terminaux, soit dans les fonctions (sauf au niveau de profondeur maximum, où seuls des symboles terminaux sont autorisés). Les arbres représentant les individus sont donc de forme et de taille très irrégulières.

2. **la méthode complète :**
   Les symboles sont tous choisis dans l'ensemble des fonctions, sauf au niveau de profondeur maximal. Ainsi, tous les individus sont équilibrés, complets et ont la même taille, correspondant à la taille maximale autorisée.

3. **une synthèse des deux précédentes, la méthode « ramped half & half » :**
   La profondeur maximale varie entre les différents individus générés. Pour chaque profondeur possible entre 2 et la profondeur maximale fixée au début, deux individus sont générés : un premier par la méthode croissante, le second par la méthode complète. La population est ainsi diversifiée en forme (la moitié des individus ne sont pas équilibrés) et en taille (il est assuré qu'il y a au moins un individu de chaque profondeur possible).

2) Evaluation de la qualité d'une solution

Lorsqu'une solution est générée, il faut pouvoir évaluer sa qualité (appelé également « fitness »). Pour pouvoir la déterminer, un certain nombre d'exemples sont fournis au programme généré. Ces exemples sont également appelés « fitness case ».

La façon de la déterminer est dépendante du problème. Ainsi, pour un problème de classification, plus le nombre d'exemples bien classés par le programme généré est élevé, plus sa qualité sera grande. Pour un problème de régression, un écart entre les valeurs générés par le programme et les valeurs exemple est calculé (par la variance, l'écart-type, la somme des erreurs absolue, ou une autre méthode) et plus cet écart sera élevé, moins la qualité du programme sera grande.

La valeur de ce calcul de la qualité est très dépendant du problème. Par exemple, l'écart entre deux points, dans le cas d'un problème de régression, étant un réel le plus proche de 0 possible alors que le nombre de nourriture avalé par un programme de vie artificielle est un entier le plus grand possible. Ainsi, ces valeur de « fitness » sont incomparables entres elles. Pour palier à ce problème, il existe plusieurs représentation standardisées de cette qualité :

1. **« fitness » standardisé**
   La meilleure valeur possible est 0 et toutes les valeurs sont positives.

2. **« fitness » normalisé**
   Similaire à la représentation standardisée, mais les valeurs sont comprises entre 0 et 1.

3. **« fitness » ajusté**
   Similaire à la représentation normalisée mais où le meilleur score possible vaut 1. Souvent calculé par $\frac{1}{(1+\ fitnessstandardisé)}$

3) La sélection

1. Définition & Principe

Les modifications successives des générations dans les populations naturelles sont orientées par les **pressions intérieures** (séduction, compétition dans l'espèce) et **extérieures** à l'espèce (limitation des ressources, modifications de l'environnement, prédateurs, parasites...), ce qui influence la survie et la reproduction des individus. Pour déterminer quels individus sont plus enclins à obtenir les meilleurs résultats, une sélection est opérée. Ce processus est analogue à un processus de sélection naturelle, les individus les plus adaptés gagnent la compétition de la reproduction tandis que les moins adaptés meurent avant la reproduction, ce qui améliore globalement l'adaptation.

Elle apparaît quand les conditions suivantes sont réunies :

- renouvellement d'une population d'individus par mortalité et reproduction;
- variabilité de caractères au sein des individus d'une population à un instant donné;
- héritabilité de certains de ces caractères variables, c'est-à-dire corrélation forte entre ces caractères chez un individu et ces caractères chez ses parents, ou plus généralement, ses ancêtres;
- variabilité du nombre de descendants;
- interaction non aléatoire entre les caractères variables héritables et l'environnement pour déterminer statistiquement l'importance de la descendance d'un individu.

Il en découle alors que l'environnement détermine une orientation des modifications successives des générations.

Selon Darwin la sélection désigne, dans le domaine de la biologie:

- soit la **survie** et la **reproduction** différentielles des **phénotypes**.
- soit le **système** pour **isoler** ou **identifier** des **génotypes particuliers** dans une population mélangée.

Et on distingue :

- La **sélection naturelle** : survie et reproduction différentielles des organismes, suite à des différences dans les caractéristiques qui affectent leur capacité à utiliser les ressources environnementales. L'adjectif ***naturel*** s'oppose chez **Darwin** au concept de sélection ***artificielle***. Elle se compose :

- d'une sélection de survie (atteindre les proies, échapper aux prédateurs, gérer les parasites et germes de maladie)
- d'une sélection sexuelle (obtenir une descendance).

- La **sélection artificielle** : pratique de sélection d'individus dans une population pour la reproduction, généralement parce que ces individus possèdent un ou plusieurs caractères désirés.
- La **sélection assistée par marqueurs** (abréviation : SAM) : **utilisation de marqueurs** d'**ADN** pour améliorer la réponse à la sélection dans une population. Les marqueurs seront étroitement liés à un ou plusieurs loci cibles, qui peuvent être souvent des loci à effets quantitatifs.
- La **sélection génétique** : processus de sélection de gènes, de cellules, de clones, etc. au sein d'une population ou entre des populations ou des espèces. La sélection génétique résulte généralement de la différence des taux de survie des génotypes variés, reflétant plusieurs variables, y compris la pression sélective et la variabilité génétique présentes dans les populations.
- La **sélection massale** : pratique utilisée dans l'amélioration génétique des plantes et des animaux. C'est la sélection d'un nombre d'individus, selon leur phénotype, qui vont former la génération suivante par inter croisement.
- La **sélection cellulaire**
- La **sélection clonale**
- La **sélection d'hybrides**
- ……

2. Les différentes méthodes de sélection

On distingue plusieurs méthodes de sélection, mais les deux les plus utilisées sont **la sélection par roulette proportionnelle** et **la sélection par tournoi.**

Sélection par roulette (*Roulette Wheel selection*) :
Pour chaque individu, la probabilité d'être sélectionné est proportionnelle à son adaptation au problème. Le principe de Roulette Wheel selection est celui de la roue de la fortune biaisée. Cette roue est une roue de la fortune classique sur laquelle on associe à chaque individu un segment dont la longueur est proportionnelle à sa fitness. On effectue ensuite un tirage aléatoire utilisé dans les roulettes de casinos avec une structure linéaire. Avec ce système, les grands segments, c'est-à-dire les bons individus, seront plus souvent adressés que les petits.

Sélection par tournoi :
La sélection par tournoi consiste à sélectionner n individus au hasard et à prendre le meilleur parmi ces n individus. On organise autant de tournois qu'il y a d'individus à repêcher. Le nombre n permet de donner plus ou moins de chance aux individus peu adaptés. Avec un nombre élevé de participants, un individu faible sera presque toujours sûr de perdre. Le nombre d'individus par tournoi détermine les paramètres d'exploration (n petit) et d'exploitation (n grand) du bassin génétique.

Sélection par rang :
La sélection par rang trie d'abord la population par fitness. Ensuite, chaque chromosome se voit associé un rang en fonction de sa position. Le plus mauvais chromosome aura le rang 1, le suivant 2, et ainsi de suite jusqu'au meilleur chromosome qui aura le rang N (pour une population de N chromosomes). La sélection par rang d'un chromosome est la même que par roulette, mais les proportions sont en relation avec le rang plutôt qu'avec la valeur de

l'évaluation, c'est à dire les individus choisis sont ceux qui possèdent les meilleurs scores d'adaptation (meilleur rang), le hasard n'entre donc pas dans ce mode de sélection.

Sélection « steady-state » :
L'idée principale est qu'une grande partie de la population puisse survivre à la prochaine génération. A chaque génération sont sélectionnés quelques chromosomes (parmi ceux qui ont le meilleur coût) pour créer des chromosomes fils. Ensuite les chromosomes les plus mauvais sont retirés et remplacés par les nouveaux. Le reste de la population survie à la nouvelle génération.

Elitisme :
A la création d'une nouvelle population, il y a de grandes chances que les meilleurs chromosomes soient perdus après les opérations d'hybridation et de mutation. Pour éviter cela, on utilise la méthode d'élitisme. Elle consiste à copier un ou plusieurs des meilleurs chromosomes dans la nouvelle génération. Ensuite, on génère le reste de la population selon l'algorithme de reproduction usuel. Cette méthode améliore considérablement les algorithmes génétiques, car elle permet de ne pas perdre les meilleures solutions.

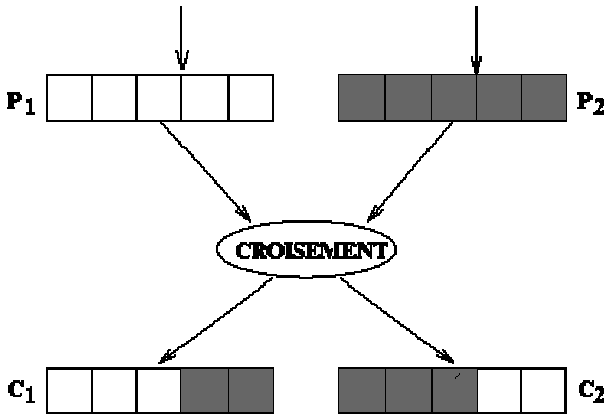
Sélection uniforme:
La sélection se fait aléatoirement, uniformément et sans intervention de la valeur d'adaptation. Chaque individu a donc une probabilité 1/P d'être sélectionné, où P est le nombre total d'individus dans la population.

4) Croisement

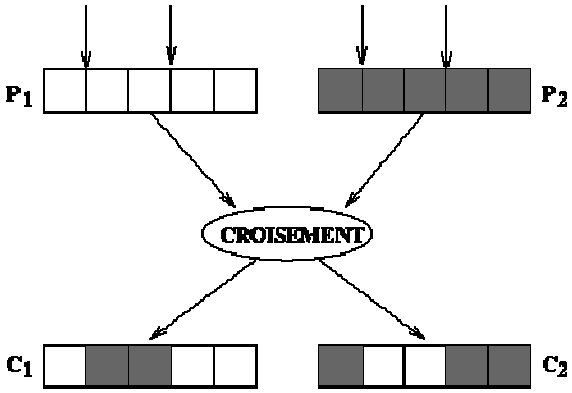
Après avoir fait une sélection des individus « Genitor », il nous faut générer une nouvelle population, afin d'enrichir la diversité de la population. Le croisement permet cela en manipulant la structure des chromosomes.

Il existe plusieurs manières d'effectuer un croisement soit par cross-over où l'on a besoin de deux parents « Genitor » qui génèrent à la fin du croisement deux enfants, soit par copie où l'on n'a besoin que d'un seul parent « Genitor » qui nous donnera à la fin du croisement un enfant qui est le parent lui-même (sa copie).

1. ***cross-over*** : d'une manière générale le chromosome de chaque parent est découpé en deux parties qui sont recombinées pour former les descendants (voir exemple1), mais il se peut qu'il y ait plusieurs points de croisement (voir exemple2).



Exemple1. cross-over avec un seul point de croisement (mono-point).



Exemple2. cross-over avec plusieurs points de croisement (multi-points de croisement)

Le choix du nombre de points de croisement diffère selon l'algorithme.

2. ***copie*** : appelé aussi cross-over uniforme, où le nouvel individu hérite du même gène que le parent « Genitor ».

5) Mutation

Permet de changer (permuter) un gène d'un chromosome par un autre d'une manière aléatoire, ce qui est à première vue très ressemblant avec un cross-over. La différence est que le cross-over essaie de converger vers une solution qui lui parait la meilleure (s'intéresse à la qualité), mais que la mutation permet la diversité. En quelque sorte, la mutation sert à éviter une convergence prématurée de l'algorithme. Par exemple lors de la recherche d'une solution optimum la mutation sert à éviter la convergence vers un optimum local.

On doit définir un taux de mutation lors des changements de population qui est généralement compris entre 0,001 et 0,01. Il est nécessaire de choisir pour ce taux une valeur relativement faible de manière à ne pas tomber dans une recherche aléatoire et conserver le principe de sélection et d'évolution.

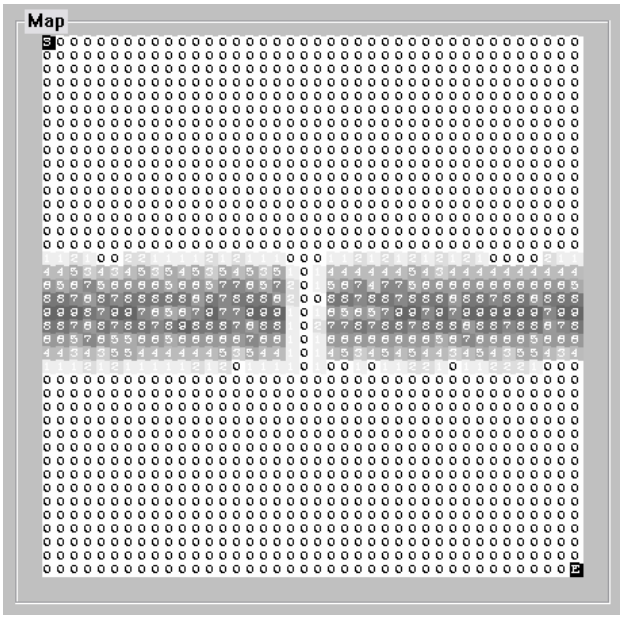
Exemple

Le « pathfinding » est un exemple parfait de programmation génétique. En effet, afin d'optimiser le chemin le plus rapide pour aller d'un point à un autre, il est, la plupart du temps, obligatoire de parcourir toutes les solutions possible. Si l'on part de l'euristique qu'un chemin trouvé est une bonne base pour la solution, alors la génération de recherche de chemin à partir des recherches précédentes est idéale.

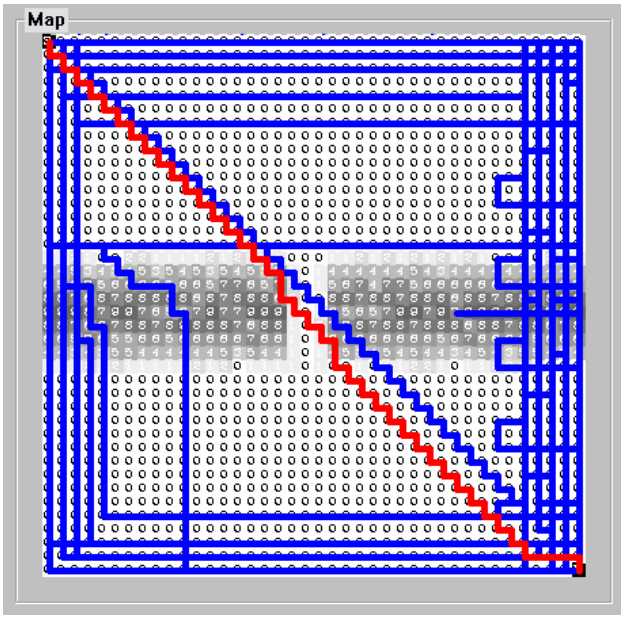
Un exemple de programmation génétique de pathfinding est disponible à l'adresse suivante : http://www.gamasutra.com/features/20060626/strom_01.shtml

Dans cet exemple, le chemin a parcourir peut contenir plusieurs variations (eau, montagne, ennemis, murs, …) Le programme met a notre disposition plusieurs façons de trouver le chemin idéal : grâce à la méthode « A* » bien connu dans le domaine du « pathfinding » que nous ne développerons pas ici et la méthode par programmation génétique. On s'aperçoit qu'après plusieurs itérations de la méthode par programmation génétique, le chemin devient optimal relativement rapidement.

Pour cet exemple, nous prendrons un terrain contenant un passage entre deux montagnes. Plus la zone est foncée, plus la montagne est élevée.



On constate que le programme trouve une solution acceptable (non loin de l'optimum) au bout d'une centaine de seconde (environ 400 itérations). La meilleure solution trouvée est celle en rouge, les autres sont toutes celles testées :



Conclusion

Nous avons étudié la programmation génétique avec ses différentes phases puis nous avons observé son fonctionnement sur un exemple : le « pathfinding ». Cette méthode de recherche de solutions est très efficace dans certains domaines où aucune autre méthode ne peut fournir de bons résultats. Néanmoins, son paramétrage est très délicat et des paramètres trop moyens peuvent mener à une très mauvaise solution.

Les recherches sur ce domaine restent ouvertes, comme par exemple la découverte de nouveaux algorithmes de sélection, de croisement et de mutation.

Bibliographie

Guillaume Beslon
"Algorithmes génétiques et methods évolutives"
ALAB Team, Computer Science Dept. INSA Lyon

John R. Koza
"Genetic Programming: a paradigm for genetically breeding populations of computer programs to solve problems"
Computer Science Department Stanford University, juin 1990

Jean-Sébastien Lacroix, Stéphane Terrade
"Algorithmes Génétiques"
MATH 6414, 17 novembre 2004.

Jean-Baptiste Mouret
"Concepts fondamentaux des algorithmes évolutionnistes"
15 novembre 2005

Denis Robilliard
"Programmation Génétique & Apprentissage par Algorithmes Evolutionnaires"
LIL, ULCO

William B Langdon, Riccardo Poli
"Foundations of genetic programming"
Springer, cop. 2002.