Un algorithme génétique pour le forecasting financier

Houcine Senoussi

October 30, 2018

Introduction

- Nous décrivons un algorithme génétique conçu pour prévoir l'évolution des cours en bourses (Forecasting).

Définitions

- On considère une série temporelle c_t représentant le cours en bourse d'une action.
- L'objectif est de détecter des **régularités** dans l'évolution de ce cours et de les utiliser pour **prédire** le futur.
  - Il s'agit donc d'**apprendre** un **modèle** : l'AG est utilisé ici dans une tâche de machine learning.
- On peut étudier le comportement indépendamment de tout autre cours ou par rapport à un indice de référence (*S&P500* ou *CAC40* par exemple).
  - Avantage de la deuxième option : elle permet d'identifier les actions les plus profitables.

Définitions

- Soit i_t la série temporelle formée par les valeurs de l'index de référence considéré. Nous définissons une nouvelle série temporelle v_t de la manière suivante :

$$v_t = (\frac{c_t}{c_{t-1}} - 1)\text{-}(\frac{i_t}{i_{t-1}} - 1)$$

- puis la série

$$z_t = s(v_t)$$

où s est la fonction définie par :

  - $s(x) = 0$ pour $x \leq 0$
  - $s(x) = 1$ pour $x > 0$

Définitions

- Nous organisons ensuite nos données sous la forme de l'ensemble de couples

$$E = \{(X_1, Y_1), ..., (X_N, Y_N)\}$$

dans lequel

  - $X_d = (z_1, ..., z_d)$.
  - $Y_d = s((\frac{c_{d+k}}{c_d} - 1) - (\frac{i_{d+k}}{i_d} - 1))$

- Autrement dit :
  - X_d est une sous-suite de z_t qui représente l'historique à l'instant d.
  - Y_d représente la performance future à un horizon k.

Définitions

- Objectif : Construire un modèle permettant pour toute valeur de X de prédire la valeur de Y correspondante.
- Bien noter que nous sommes en train de résoudre un problème d'**apprentissage supervisé inductif** :
  - Point de départ : l'ensemble E, ensemble d'exemples/d'apprentissage.
  - Output : un classifieur.
- Ce problème sera résolu à l'aide d'un algorithme génétique.

Définitions

- Le modèle prendra donc la forme d'un ensemble de règles permettant de prédire la valeur de Y en fonction de X.
  - Une règle r sera donc définie par un couple (C, Y), où C est une condition portant sur les éléments de X et Y la valeur attendue lorsque cette condition est vérifiée.
- Les conditions C portent sur des sections X (). Nous leur donnons la forme suivante :
  - $C = ((z_d = s_d) \wedge (z_{d-1} = s_{d-1}) \wedge ... \wedge (z_{d-m} = s_{d-m}))$. où m est la longueur de la section de x à analyser et $s_i \in \{0, 1, *\}$.

Définitions

- Le classifieur R sera donc un ensemble de règles $\{r_1, ..., r_N\}$ avec $r_k = (C_k, Y_k)$.
- La valeur de Y attribuée à une donnée X sera obtenue par un vote majoritaire des règles qui le composent.
- **Remarques :**
  1. Que se passe-t-il lorsque le résultat du vote est $50 - 50$ (ou presque) ?
  2. Que se passe-t-il lorsque les règles formant R n'"ont pas d'avis" sur une situation donnée ?

Définition de l'algorithme génétique

- L'espace de recheche est l'ensemble de classifieurs possibles.
- Un individu est un classifieur, càd un ensemble de couples $\{r_1 = (C_1, Y_1), ..., r_N = (C_N, Y_N)\}$
- Définition de la fitness :
  - Appliqué à un ensemble des données de test, un classifieur R peut donner une bonne réponse, une mauvaise réponse ou pas de réponse. Appelons $nb(R)$, $nm(R)$ et $np(R)$ les nombres de réponses dans chacune de ces trois catégories.
  - La fonction fitness doit augmenter avec $nb(R)$ et baisser avec $nm(R)$. Par exemple

$$fitness(R) = \frac{1}{2}(1+\frac{nb(R)-nm(R)}{nb(R)+nm(R)+np(R)})$$

Opérateurs génétiques : la sélection

- Dans cet algorithme la sélection est déterministe. Le nombre d'exemplaires de chaque individu produit par cette opération est le même que la moyenne de celui produit par la "roue de la fortune".
- Éviter les cas limites tout en maintenant la diversité.

Opérateurs génétiques : la sélection

- Soit donc M le nombre d'individus à sélectionner dans une population de N individus R_1, ..., R_N.
  - Pour commencer, nous calculons $p_i^1 = M \frac{fitness(R_i)}{\sum_{i=1}^{N} fitness(R_i)}$, pour $i = 1, ..., N$.
  - Nous sélectionnons alors $n_i = \lfloor p_i^1 \rfloor$ exemplaire des l'individu R_i. Soit $M_1 = \sum_{i=1}^{N} n_i$.
  - Ensuite nous calculons la valeur $p_i^2 = n_i - p_i^2$. Nous avons donc $0 \leq p_i^2 < 1$.
  - Soit $M2 = M$ -$M1$. Nous sélenctionnons un exemplaire de chacun des $M2$ individus ayant les valeurs les plus élevées de p_i^2.

Opérateurs génétiques : le croisement

- Le croisement se fait en deux temps : un croisement à deux points appliqué aux classifieurs puis un croisement à un point appliqué aux règles.
- Soit donc deux individus $A = A_1 ... A_K$ et $B = B_1 \ ... \ B_L$.
  - Les A_i et B_j sont les règles càd les couples (C, Y). Bien noter que ces règles peuvent être aisément représentées à l'aide de l'alphabet $\{0, 1, *\}$.
  - Bien noter que les classifieurs n'ont pas forcément le même nombre de règles , càd que les individus n'ont pas forcément la même longueur (ici on peut avoir $K \neq L$).

Opérateurs génétiques : le croisement

- Première étape du croisement :
  - Deux positions (éventuellement confondues) sont sélectionnées aléatoirement dans chacun des deux individus.
  - Soient i_1, i_2, j_1 et j_2 ces positions.
  - Le résultat de cette première étape est le suivant :
    - $A' = A_1 \ ... \ (A^1_{i1} B^2_{j1}) B_{j1+1} \ ... \ B_{j2-1} (B^1_{j2} A^2_{i2}) \ ... \ A_K$
    - $B' = B_1 \ ... \ (B^1_{j1} A^2_{i1}) A_{i1+1} \ ... \ A_{i2-1} (A^1_{i2} B^2_{j2}) \ ... \ B_L$
  - $(A^1_{i1} B^2_{j1})$, $(B^1_{j2} A^2_{i2})$, $(B^1_{j1} A^2_{i1})$ et $A^1_{i2} B^2_{j2})$ sont le résultat des croisements à un point de la deuxième étape.

Opérateurs génétiques : le croisement

- Deuxième étape du croisement :
  - Soient donc deux règles $r_1 = (C1, Y1)$ et $r_2 = (C2, Y2)$
  - Nous savons que les C_i portent sur des indices successifs $d - m, ..., d$, la valeur de m pouvant être différente entre les deux règles.
  - Nous savons aussi que les valeurs utilisées dans ces conditions sont 0, 1 et $*$ et que la valeur d'un Y est 0 ou 1.
  - Les règles peuvent donc être décrites à l'aide de mots construits sur l'alphabet $\{0, 1, *\}$.
  - Avant de croiser deux règles, nous commençons par leur imposer une longueur commune *mcommun* puis nous tronquons ou complétons par des $*$ chacune des deux règles pour atteindre la longueur *mcommun*.
  - Ensuite un croisement à un point est appliqué.

Opérateurs génétiques : le croisement

- Définir la longueur commune *mcommun* :
  - Nous utilisons une méthode probabiliste. Pour cela nous définissons une distribution normale $N(\mu, \sigma^2)$.
    - Les paramètres de la loi sont la moyenne et la différence des deux longeurs des règles à croiser.
  - Soit donc $f(x)$ la densité de probabilité correspondant à cette distribution. Nous utilisons la distribution suivante pour la longueur des règles :

$$P(longueur = x) = \frac{f(x)}{\sum_{j=1}^{lmax} f(l)}, \text{ pour } x = 1, ..., lmax.$$

  lmax est la longueur maximale autorisée pour les règles.

Opérateurs génétiques : la mutation

- Exercice :
  - Proposer une (ou plusieurs) définition(s) de la mutation dans cet algorithme génétique.

Résultats expérimentaux

- Travail datant de 1999.
- Testé sur la bourse de Lisbonne.
- Description des données d'apprentissage et de test :
  - Utilisation des prix de fermeture.
  - Données quotidiennes s'étendant sur 4-5 jours.
  - Les données sont réparties sur deux ensembles : l'ensemble d'apprentissage (70%) et l'ensemble de test (30%).
- Le résultat de l'apprentisage (le classifieur ayant la fitness la plus élevée) est appliqué aux données de test et ses performances sont comparées à celles de plusieurs stratégies de classification "naïves".
- Plusieurs test utilisant des valeurs différentes avec l'horizon k ont été réalisés.
  - Les meilleurs résultats ont été obtenus avec $k = 10 jours$.

Résultats expérimentaux

- (Critère d'arrêt) : On constate une convergence de la meilleure fitness et de la fitness moyenne au bout de quelques dizaines de génération.

Conclusion

- Un algorithme génétique appliqué à l'apprentissage d'un classifieur pour la prévision de l'évolution des cours en bourse.
- Critiques ? Améliorations possibles ?

References

- Le travail présenté ici est décrit dans l'article "A GENETIC LEARNING ALGORITHM APPLIED TO FINANCIAL FORECASTING" de J. Araujo, P. Bernardo et A. Rosa.