Estimation de probabilités par algorithme génétique

Anne-Lise Bedenel, Laetitia Jourdan, Christophe Biernacki

Estimation de probabilités par algorithme génétique

Anne-Lise Bedenel[1,2,3], Laetitia Jourdan[2], Christophe Biernacki[3]

[1] MeilleureAssurance, France
anne-lise.bedenel@meilleureassurance.com
[2] Université Lille 1 CRIStAL, UMR 9189, France
laetitia.jourdan@univ-lille1.fr
[3] Université Lille 1, Inria, France
christophe.biernacki@math.univ-lille1.fr

Mots-clés : *Algorithme génétique,BIC,assurance,WEB*

1 Introduction

Dans le cadre de la comparaison d'assurances en ligne, les données évoluent constamment, ce qui rend leur exploitation difficile. En effet, la plupart des méthodes d'apprentissage classiques nécessitent des descripteurs de données identiques pour les échantillons d'apprentissage et de test. Afin de répondre aux attentes métiers, les formulaires en ligne d'où proviennent les données sont régulièrement modifiés. Cette modification régulière des variables et des descripteurs de données complexifie les analyses. Un premier travail à l'aide d'une méthode statistique (MS) basée sur la vraisemblance et la sélection de modèle [1] a été réalisé. Cependant, cette méthode est très coûteuse en temps de calcul et nous contraint à limiter notre espace de recherche.

Dans ce travail, nous proposons d'utiliser un algorithme génétique (AG) afin de pallier aux défauts de notre méthode statistique.

2 Modélisation

Dans le cadre de travaux avec l'entreprise meilleureassurance.com, la question suivante s'est posée : Comment les internautes réagissent-ils lorsque les descripteurs d'une variable changent dans un questionnaire d'assurance automobile ?

Pour répondre à cette question, nous étudions la variable « Niveau de garantie souhaité » que les internautes doivent renseigner. Cette variable avait initialement quatre descripteurs :{Tiers, Tiers++, Intermédiaire, Tous Risques}. Dans un second temps, elle a été décomposée en sept descripteurs : {Tiers, Tiers+, Tiers++, Intermédiaire, Tous Risques, Tous Risques+, Tous Risques++}. Contrairement à la plupart des sites e-commerce, nous n'avons pas de données de navigation à utiliser. En effet, un internaute venant comparer un produit d'assurance, revient rarement sur le site. Notre cas d'usage est représenté par le graphe 1, où les arcs orientés représentent les paramètres que l'on va chercher à estimer.

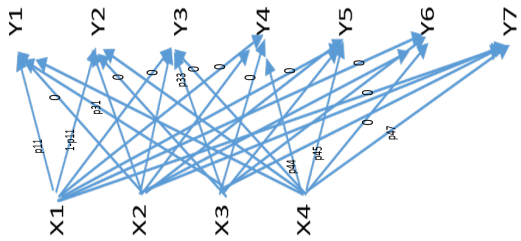

FIG. 1 – Graphe des appariements possibles entre X-la variable avant changement et Y-la variable après changement.

Pour répondre aux contraintes d'identifiablité statistique, nous imposons des paramètres (arcs) à 0. Un modèle est alors un ensemble de paramètres libres, de paramètres non libres et de paramètres forcés à 0. Dans cet exemple, nous estimerions 6 paramètres (arcs) et en aurions 18 à 0, les paramètres restants étant non libres. Comme nous ne savons pas quels sont les paramètres à estimer et ceux à 0, nous comparons les différents modèles possibles de notre famille de modèles à l'aide du critère statistique BIC [2]. Le modèle ayant le plus petit BIC est le modèle le plus en adéquation avec nos données initiales. Nous choisissons d'utiliser un AG stationnaire mono-objectif, afin d'avoir une part d'aléatoire et obtenir une bonne solution rapidement. Notre critère d'évaluation est le critère BIC que nous voulons minimiser. Comme nous souhaitons également estimer des probabilités, nous utilisons un codage réel. Afin d'avoir de bonnes performances rapidement, nous avons comparé différentes métaheuristiques et retenu la méthode la plus robuste et minimisant au mieux le critère BIC. Suite à cette comparaison nous avons choisi d'utiliser un cross-over SBX [3] qui ne s'applique que sur les probabilités estimées. De plus, dans cet opérateur, les enfants gardent les paramètres fixés à 0 des parents. Une mutation polynomiale appliquée sur un seul paramètre est également utilisée.

3 Expériences numériques

Le tableau 1 présente les résultats de notre AG et de la MS. Les 2 premiers jeux de données contiennent des données simulées où la MS retrouve le modèle souhaité et indique le critère BIC à challenger. Le jeu de données suivant contient des données réelles où l'on ne connait pas le modèle à retrouver, on peut juste se fier au critère BIC. Sur les 2 jeux de données simulées,

	Best BIC AG	Best BIC MS
Données simulées 1	$62\,638,0$	$62\,637,1$
Données simulées 2	$688\,351,4$	$688\,226,4$
Données réelles	$58\,287,7$	$58\,306,6$

TAB. 1 – Tableau des différents résultats : AG-Algorithme génétiques, MS-Méthode statistique

on remarque que l'AG trouve un résultat pour le critère BIC très proche de celui donné par la MS. Sur le jeu de données réelles, on constate que le résultat du critère BIC trouvé par l'AG est même meilleur que celui trouvé par la MS. Le fait d'élargir l'espace de modèle nous permet donc de surpasser la MS.

4 Conclusion et perspectives

Pour tester 4095 modèles, la MS met 1h07. Avec le même jeu de données, le temps moyen est de 3.31 minutes avec notre AG. De plus, celui-ci nous permet d'élargir notre espace de modèle et donc d'en trouver de nouveaux. On constate également que les solutions renvoyées par l'AG sont très proches voire supérieures aux solutions renvoyées par la MS. Nous devons désormais valider les résultats d'un point de vue métier et améliorer notre AG avec de nouvelles métaheuristiques prenant en compte de nouveaux opérateurs plus intelligents basés sur les observations statistiques.

Références

[1] Anne-Lise Bedenel, Christophe Biernacki and Laetitia Jourdan. *Appariement de données évoluant en temps.* Société française de statistique, 2016.

[2] Emilie Lebarbier, Tristan Mary-Huard *Le critère BIC : fondements théoriques et interprétation.* 2004

[3] Agrawal, Ram Bhushan and Deb, K and Agrawal, RB *Simulated binary crossover for continuous search space.* Complex systems,9, 115–148,1995