Dynamique des populations : pallier le manque de données

Etienne Auclair, Régis Sabbadin, Nathalie Peyrard

tangente HORS SÉRIE

tangente

l'aventure mathématique

n° HS 68

Intelligence artificielle

L'alliance des mathématiques et de la technologie pour transformer le monde

Notre quotidien bouleversé

Traduction automatique
Voitures autonomes
Régulation et prédiction en finance

Des machines qui apprennent

Quand l'ordinateur explique ses choix
Go, bridge : deux approches de l'IA

DOM - LUX - BELG : 7,30 € CANADA : 11,99 $ can SUISSE : 12.20 CHF
TUNISIE : 7,20 DTU MAROC : 70 DH ISSN 1294-9949 Octobre 2018

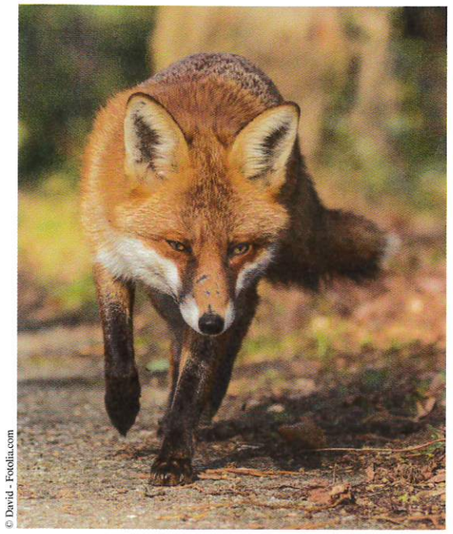
M 05446 - 68 - F: 6,80 € - RD

par É. Auclair, N. Peyrard & R. Sabbadin

Dynamique des populations : pallier le manque de données

Dans certains domaines, les données sont peu nombreuses. Pas question d'utiliser la statistique descriptive ! Les mathématiques proposent alors d'autres outils : la topologie et les probabilités aident l'écologue à prévoir l'évolution de groupes d'espèces.

Étienne Auclair est doctorant en intelligence artificielle ; Nathalie Peyrard est directrice de recherche en statistiques computationnelles et appliquées ; Régis Sabbadin est directeur de recherche en intelligence artificielle. Ils font partie de l'unité de « Mathématiques et informatique appliquées » du centre INRA de Toulouse (Haute-Garonne).


© David - Fotolia.com

Afin de comprendre le fonctionnement d'un écosystème, les écologues observent les espèces qui y cohabitent. Ils collectent ainsi des données de présence / absence des espèces sur plusieurs années. Comment extraire de l'information de ces suites de 0 et de 1, en particulier pour apprendre quelles espèces influencent la survie d'autres (par prédation, parasitisme, mutualisme…) ? Ici, on est loin du « Big Data », car obtenir ces données est coûteux… Elles sont donc rares ! Il faut surveiller la zone considérée suffisamment longtemps pour être (à peu près) sûr que les espèces sont présentes ou absentes.
L'idée est donc d'aider l'apprentissage des interactions au sein d'une communauté d'espèces grâce au formalisme des graphes et à l'incorporation de connaissances « expertes » dans un modèle de la dynamique des espèces.

+ Graphes et probabilités

Un *graphe d'interactions écologiques* est un outil topologique de modélisation qui recense, pour chaque espèce, l'ensemble des autres espèces dont elle influence la survie, positivement ou négativement, d'un instant donné à l'instant suivant. Informellement, les sommets de ce graphe représentent les espèces et une arête dirigée de ce graphe, l'influence d'une espèce sur une autre. Ce graphe d'interaction peut être utilisé pour modéliser la dynamique de la communauté des espèces au cours du temps. On se sert de sa structure pour construire un réseau bayésien particulier, dit *dynamique* (voir *Tangente* 182), décrivant les probabilités de présence de toutes les espèces à l'instant t en fonction de la liste des espèces présentes à l'instant $t-1$. Ces probabilités sont représentées de manière concise, en exploitant des propriétés d'indépendance entre certaines espèces, la présence d'une espèce à l'instant t étant supposée ne dépendre que de la présence de quelques autres seulement (dénommées du terme générique de *parents*) à l'instant précédent.
Cette hypothèse que chaque espèce n'est influencé que par peu d'autres permet une grande économie de moyens, mais l'apprentissage d'un tel modèle peut rester difficile, en particulier lorsque l'on dispose de peu de données réelles. Si chacune des n espèces possède k parents, le nombre de paramètres à considérer est de $n \times 2^k$, et il n'est hélas pas imaginable que le nombre d'années d'observation soit de cet ordre de grandeur (sans compter que le nombre maximum de parents, k, n'est pas connu *a priori*).

Le modèle RBDE (pour *réseau bayésien dynamique étiqueté*) permet de réduire le nombre de paramètres du modèle. Mais en contrepartie, les arêtes du graphe sont étiquetées par un + ou un –, pour modéliser une influence positive ou négative. On définit par ailleurs les probabilités de présence de chaque espèce à l'instant t en fonction de la communauté présente à l'instant $t-1$ à partir d'un paramètre par étiquette et d'un paramètre de réintroduction.

+ Modéliser les influences entre espèces

Dans la réalité, on ne connaît pas le graphe d'interaction (en tout cas, pas bien), ni la valeur des paramètres. En revanche, on peut espérer avoir des observations de la présence / absence des différentes espèces, pour un petit nombre de pas de temps. À partir de ces quelques données, on va tâcher d'« apprendre » le graphe étiqueté, c'est-à-dire les influences positives et négatives entre les espèces, en cherchant le modèle qui rend compte des données le plus fidèlement possible. En pratique, on le fait en optimisant une *fonction de vraisemblance*.

Une fois cela fait, on peut alors aller voir notre écologue préféré pour lui soumettre les interactions entre espèces que l'algorithme a calculées. Sa réaction typique est alors : *« Il est bizarre ton réseau, tout le monde sait que les mulots ne mangent pas les renards. Et puis on sait d'ailleurs que les petites bêtes ne mangent pas les grosses (en général) ! »*
Tout le monde le sait, mais pas notre algorithme… enfin, pas quand on ne le lui dit pas ! Un avantage de la méthode d'apprentissage du modèle RBDE est qu'elle permet d'incorporer de la connaissance *a priori* sur les relations entre les espèces. En effet, puisque la phase d'apprentissage des ensembles de parents consiste à résoudre un programme d'optimisation discrète dont certaines des variables représentent la présence ou l'absence de relation positive ou négative entre paires d'espèces, rien de plus facile que d'intégrer des connaissances du type « telle espèce a / n'a pas un effet positif sur telle autre ».
Comment, cependant, incorporer des connaissances du type « *en général*, les petites bêtes ne mangent pas les grosses » ? On peut considérer que la taille des espèces est connue, on peut faire confiance à notre écologue préféré pour les avoir pesées et mesurées (ou avoir consulté la littérature à ce sujet). Mais n'oublions pas que dans toutes les sciences du vivant, l'exception est la norme. *« En général »* signifie seulement qu'il est plus probable qu'une relation + existe d'une espèce plus petite vers une espèce plus grande, car « en général », on s'attaque à des proies plus petites que soi. De même, une relation – existe, le plus souvent, d'une espèce plus grande vers une espèce plus petite. Mais on ne peut pas éliminer la possibilité de relations inversées.
Heureusement, le révérend Bayes veille sur nous et nous fournit les outils mathématiques permettant d'incorporer cette nouvelle connaissance incertaine. Cela repose sur une loi de probabilité *a priori* sur l'objet « graphe ».


© paolofrisacchia - Fotolia.com

+ Les maths soulèvent de nouvelles questions

C'est ainsi que les statistiques et l'algorithmique offrent des outils qui permettent d'exploiter des observations écologiques (présence / absence d'espèces), pour permettre de proposer un modèle de leur structure. Dans le cadre RBDE, le graphe appris a une interprétation qui peut être confrontée aux expériences et aux connaissances de l'écologue. Mais le mathématicien doit rester humble ! Il sait (voir à nouveau *Tangente* 182) que les réseaux bayésiens ne font que représenter des indépendances conditionnelles entre des variables observées, et non des liens de causalité ou d'influence entre espèces. De plus, les observations sont elles-mêmes entachées d'incertitude. Comment être certain qu'une espèce non observée est bien absente ? Ou que l'écologue n'a pas confondu deux espèces ? Des facteurs externes peuvent également influencer la présence ou l'absence d'espèces (la météo par exemple)… Enfin, même dans les situations les plus favorables, les données resteront limitées en quantité, ce qui laisse peser une incertitude sur le réseau écologique appris.

□— É.A, N.P. & R.S.

Références

- *Mathématiques et géographie*. Bibliothèque Tangente 40, 2010.
- *Mathématiques et biologie*. Bibliothèque Tangente 42, 2011.