Modélisation interactive en agro-alimentaire : l'outil LiDeoGraM

Interactive Modelling in the Agrifood Domain: the LiDeoGraM tool

E. Lutton, N. Boukehlifa, A. Tonda, T. Chabin, N. Méjean-Perrot, J-D. Fekete

English Abstract – *Modelling complex agrifood processes like for instance cheese ecosystems is challenging due to the complex multi-scale nature of interactions among and between the constituent components. Furthermore, producing experimental data in this context is both tedious and expensive, resulting in scarce datasets, where the number of dimensions far exceeds the number of data samples. We present here LiDeoGraM, a visual analytics tool that combines data-driven machine learning with domain experts' knowledge to produce multi-scale models of living ecosystems. This tool is inspired from the idea of model stacking where an ensemble of simple local models are generated for each component of the studied system. The modelling process is carried out in three iterative steps: (a) using regression, LiDeoGraM proposes a set of local models for each component; (b) via a graphical interface, domain experts evaluate those local models; (c) an interactive evolutionary algorithm builds a global model while mediating between expert's subjective assessment, and an automatic evaluation based on fitting error and complexity. A first validation of our approach has been performed at three levels, combining computational and human-centered evaluations: (i) automatic tests to assess the robustness of the local model generation; (ii) a toy model test to evaluate how domain experts use our tool to discover a 'known' model; and (iii) a use case study to examine how domain experts use LiDeoGraM to model real-life multi-scale biological systems. Our results show that domain experts are able to operate our tool to discover a known model, and are able to generate new hypotheses when exploring their own datasets.*

1 Introduction : modeliser dans le domaine agro-alimentaire

Dans un cadre où la digitalisation de l'agriculture et de l'industrie agro-alimentaire produit de plus en plus de données, et sous la pression demandes politiques et sociétales stratégiques (durabilité des filières, réduction d'intrants, sécurité alimentaire et nutritionnelle, réduction de la consommation de protéines animales, santé et souffrance animale), il devient de plus en plus nécessaire de savoir exploiter efficacement ces données pour comprendre, prédire et prendre des décisions [5].

Les questions sont multiples : comment combiner techniquement les impératifs de traitement d'énormes volumes de données (big-data), de données hétérogènes, sensibles, rares (sparse-data), et les besoins exprimés par les différents acteurs (optimisation, besoins de transparence, cadres politiques et juridiques, etc.) ? Comment renouveler les approches de modélisation en employant différentes formalisations (mécanistes, statistiques, symboliques, etc.) et différents types de connaissances (expertise, savoirs profanes) ? Comment comprendre et représenter des phénomènes qui interagissent à des échelles qui peuvent être très différentes ?

Enfin, l'on doit aujourd'hui considérer incertitudes et multi-objectif (voire le *many-objectifs*), pour prendre en compte, et parfois optimiser conjointement différents critères antagonistes (environnementaux, économiques, sanitaires, etc.).

Les méthodes automatiques actuelles reposent sur l'intelligence artificielle, sous sa forme traditionnelle (traitement de signaux, d'images, fouille de données, modélisation, simulation, optimisation), bien-sûr, mais aussi sur des approches qui mettent l'humain et les collectifs au cœur des traitements (visualisation, interactions homme-machine, formalisation et intégration de l'expertise, aide à la décision).

L'enjeu actuel est de savoir articuler de façon fluide les algorithmes autonomes (apprentissage, optimisation) et les interactions homme-machine [4]. La visualisation y joue un rôle essentiel, car pour mettre l'expert au centre d'un processus de modélisation et de décision, il faut qu'il puisse comprendre, explorer, corriger, décider en connaissance de cause, avoir accès à la complexité des espaces de recherche sans être noyé, être guidé sans contrainte, remettre en question différentes hypothèses, mobiliser son expertise, exercer sa curiosité et sa créativité.

De nombreuses techniques automatiques reposent sur des algorithmes d'optimisation stochastiques puissants, auxquels est accordé une grande confiance (deep learning, par exemple). Cependant, la puissance et la validité des modélisations, et des décisions humaines qui en découlent, dépendent de nombreux facteurs relativement peu contrôlables, subjectifs, variables selon le contexte et les applications visées : quantité (du big au sparse data), qualité des données et des expertises sur lesquels sont construits ces modèles d'une part, définition et quantification des objectifs d'autre part.

Ce contexte particulier nous amène à considérer l'étape d'optimisation sous un angle différent : tout d'abord, la façon dont le problème à résoudre est formulé se révèle curieusement plus déterminante

que l'algorithme de résolution lui-même (« problem finding », plus que « problem solving »), et ensuite, il s'avère que l'efficacité des méthodes interactives repose plus sur des capacités d'exploration que sur de l'optimisation pure et dure. L'approche présentée ici repose sur une technique évolutionnaire [3] pilotée par une interface de visualisation interactive.

2 LA MODELISATION INTERACTIVE : CO-APPRENTISSAGE, CO-ADAPTATION, CO-EVOLUTION ?

La modélisation interactive et plus largement l'apprentissage interactif sont des stratégies intéressantes à plusieurs titres : elles permettent (i) d'intégrer des connaissances expertes précieuses, parfois impossibles à intégrer directement dans un modèle mathématique ou informatique, (ii) de mieux gérer certaines incertitudes, (iii) de construire une confiance, grâce à l'implication l'humain au sein du processus de modélisation, (iv) de faciliter l'émergence de nouvelles solutions, parfois inattendues (créativité, exploration).

Dans de tels systèmes, l'interaction entre homme et machine se fait via des mécanismes complexes : le système apprend les préférences de l'utilisateur et en même temps, l'utilisateur s'adapte lui-même au système. Ces mécanismes sont subtils [8], en particulier si l'on s'intéresse à des systèmes eux-mêmes complexes, appris sur des données bruitées, incomplètes ou incertaines.

Mesurer l'efficacité, comprendre les mécanismes est une tâche ardue, en particulier au regard de la mobilisation de l'expertise implicite (savoir-faire, connaissances comparatives, etc.), par opposition à l'expertise explicite, plus simple à gérer, où l'expert sait exprimer ses connaissances sous forme de contraintes ou de fonctions-objectifs.

3 LIDEOGRAM

LiDeoGraM est un outil de visualisation interactive qui combine apprentissage automatique piloté par les données et expertise humaine pour produire un modèle multi-échelle. Cette combinaison se révèle essentielle pour modéliser dans le cas où les données expérimentales sont rares et coûteuses, en les complétant par des informations expertes. LiDeoGraM est fondé sur un empilement de modèles locaux simples, qui sont générés par apprentissage automatique pour chaque variable du système considéré. Le processus de modélisation consiste à itérer trois étapes (a) une régression pour générer des modèles locaux (sous formes d'équations) ; (b) une manipulation interactive via une interface graphique pour que l'expert puisse sélectionner les modèles locaux pertinents ou en éliminer certains ; (c) un algorithme évolutionnaire pour construire ensuite un modèle global en sélectionnant un modèle local par variable, sur la base d'un compromis entre précision et complexité (optimisation multi-objectif). Nous proposons une première validation de ce système : (i) des tests automatiques pour évaluer la robustesse de la régression automatique ; (ii) un test sur un « toy model » simple et connu, afin d'évaluer le comportement du système quand il est manipulé par des utilisateurs ; (iii) un cas réel biologique (un écosystème fromager) manipulé par un expert du domaine. Nos premiers résultats sont satisfaisants, et montrent que des experts sont capables de manipuler LiDeoGraM pour reconstruire un modèle connu, et, sur leurs propres données expérimentales, de retrouver des faits connus et de générer de nouvelles hypothèses originales et pertinentes.

4 CONCLUSIONS

L'apprentissage interactif pose différentes questions [7]. L'élaboration d'algorithmes robustes de traitement de données capables d'intégrer des savoirs subjectifs nécessite de savoir où placer le curseur entre apprentissage et exploration interactive : il ne faut ni brider l'expert ni le perdre dans un océan de possibilités.

Se pose d'autre part le problème délicat de l'évaluation de l'utilité et de l'efficacité de ces outils [1,2]. Nous avons ici couplé deux types de validation (multifaceted validation) : une validation algorithmique (algorithm-centered) pour vérifier le comportement calculatoire, et une validation centrée sur l'humain, pour évaluer l'efficacité de l'outil du point de vue de l'utilisateur.

5 REFERENCES

1. Boukhelifa, N., Bezerianos, A., Cancino, W., Lutton, E. (2017). Evolutionary visual Exploration: evaluation of an IEC framework for guided visual search. Evolutionary Computation, 25(1), 55-86.
2. N. Boukhelifa, A. Bezerianos, A. Tonda, and E. Lutton. Research prospects in the design and evaluation of interactive evolutionary systems for art and science. In ACM CHI Workshop on Human Centered Machine Learning, San Jose, CA, United States, 2016
3. Chabin, T., Barnabé, M., Boukhelifa, N., Fonseca, F., Tonda, A., Velly, H., Lemaître,,B., Perrot, N., Lutton, E. (2017). LIDeOGraM: An interactive evolutionary modelling tool. International Conference on Artificial Evolution (Evolution Artificielle), 25-27 October 2017.
4. Lutton, E., Tonda, A., Boukhelifa N., Perrot, N. (2016). Complex Systems in Food Science: Human Factor Issues. FOODSIM, Catholic University Leuven, Ghent, Belgium.
5. Perrot, N., De Vries, H., Lutton, E., Van Mil, H.G.J., Donner, M., Tonda, A., Martin, S., alvarez, A., Bourgine, P., van der Linden, E. Axelos, M. (2016). Some remarks on computational approaches towards sustainable complex agri-food systems. Trends in Food Science and Technology. 48, 88-101.
6. Perrot, N., Baudrit, C., Trelea, I.C., Trystram, G., Bourgine, P. (2011). Modelling and analysis of complex food systems: state of the art and new trends. Trends in Food Science and Technology, 22(6), 304-314.
7. Sacha, D., Sedlmair, M., Zhang, L., Lee, J.A., Weiskopf, D., North, S., Keim, D. (2016). "Human-centered machine learning through interactive visualization." ESANN.
8. W. Mackay. Responding to cognitive overhead: co-adaptation between users and technology. *Intellectica*, 30(1):177–193, 2000.

- *Evelyne Lutton, Nadia Boukehlifa, Alberto Tonda, Thomas Chabin, Nathalie Méjean-Perrot : INRA-GMPA, Versailles-Grignon.*
  E-mail : Prenom.Nom@inra.fr.
- *Jean-Daniel Fekete: INRIA Saclay, AVIZ*
  E-mail : Jean-Daniel.Fekete@inria.fr

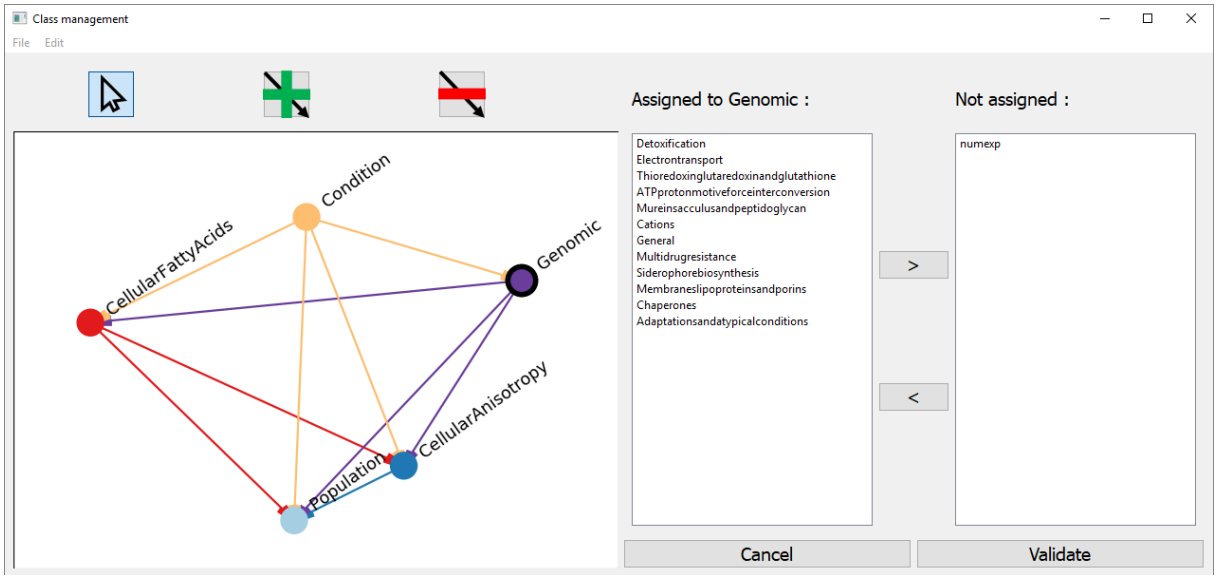

Figure 1 : Interface de classification interactive préalable des données : l'expert intègre ses connaissances explicites qui sont exploitées comme des contraintes pour la construction du graphe des dépendances (figure 2). Un lien entre deux classes signifie que toutes les variables de la classe parent peuvent être utilisées dans les équations de toutes les variables de la classe enfant. Dans l'exemple ci-dessus la classe "genomic" est composée de la liste des variables donnée dans la fenêtre du milieu (« assigned to Genomic »). Les boutons « < » et « > » permettent de faire sortir ou entrer certaines variables dans cette classe.

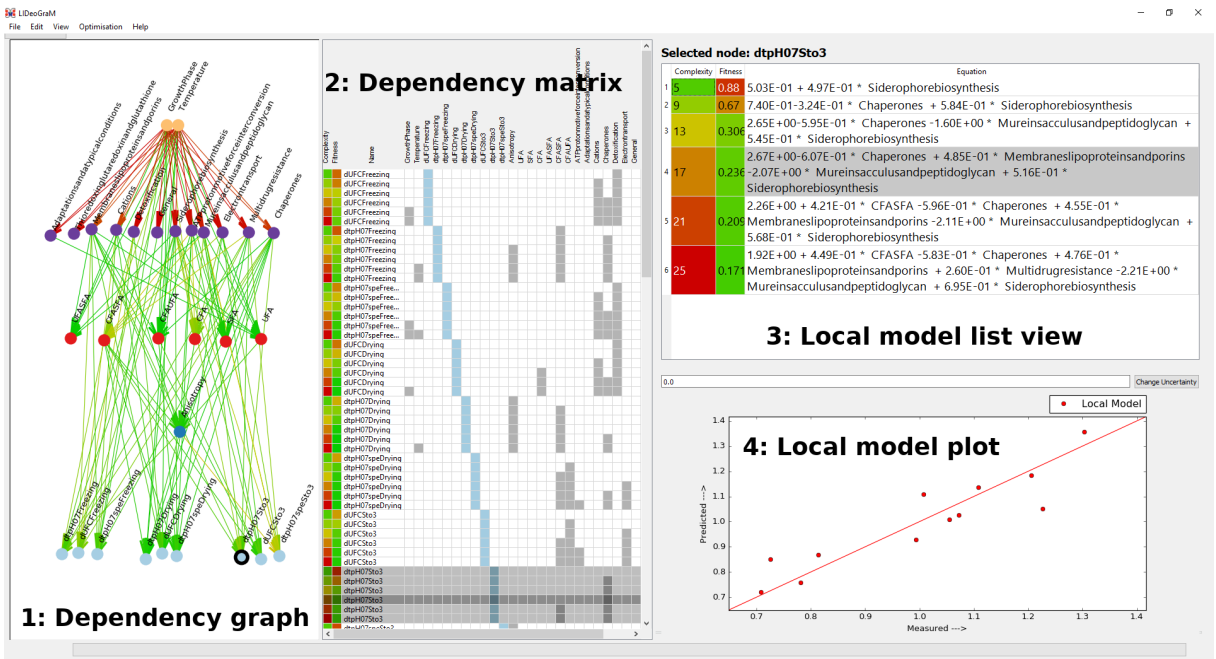

Figure 2 : Exploration d'un jeu de données sur l'affinage de fromage à l'aide de LiDeoGraM. Les experts créent de façon interactive un graphe de dépendances sur leur données (1), chaque nœud représente une variable et les liens des dépendances possibles (c'est-à-dire une équation donnant la variable du nœud en fonction d'un ensemble de variables « parents »). Ces « modèles locaux » sont générés automatiquement à partir des données disponibles. La représentation matricielle (2) donne une vue condensée de ces modèles "locaux" : chaque ligne correspond à une équation, et chaque colonne à une variable du système. Lorsqu'un nœud est sélectionné dans le graphe (ici le nœud « anisotropy »), la liste des équations possibles est présentée sous forme explicite en une liste ordonnée (3) allant de l'équation la plus simple à la plus compliquée. Deux valeurs sont données : complexité et précision. Lorsque l'une de ces équations est sélectionnée, un graphique (4) donne la répartition des erreurs sur les points de données (la valeur prédite en fonction de la valeur mesurée).