Company or University + lab: IRISA-Inria Rennes, équipe-projet LACODAM

Address: 263 Avenue Général Leclerc, 35000 Rennes

Supervisors (to be contacted for applying):

- Tassadit BOUADI, Inria/IRISA (tassadit.bouadi@irisa.fr)
- Luis GALÁRRAGA, Inria/IRISA (luis.galarraga@inria.fr)
- Anne-Isabelle GRAUX, INRA (anne-isabelle.graux@inra.fr)

Internship title: Discriminant pattern analysis on multi-class and multi-dimensional simulated data.

Keywords: crop simulation model, multidimensional data warehouse, discriminant patterns

Motivation

STICS (Brisson et al., 2003, https://www6.paca.inra.fr/stics_eng/) is a dynamic and mainly mechanistic crop model that simulates the effects of climate, soil conditions and agricultural practices on crop production and environmental fluxes. Grassland yield, water and nitrogen fluxes were simulated at the scale of France with soil, climate, vegetation and management conditions defined on a high-resolution grid (Graux et al., 2017, 2019). This work produced a data repository of 238 GBs and 173260 simulations, where a simulation combines one grid cell, one grassland type of vegetation, one type of management within a period of 30 years. The simulated data produced by STICS can be used to answer multiple research questions in agronomy. Many of those questions ask for characteristics that "discriminate" parcels, territories or regions that exhibits a particularity of interest. This particularity can be, for instance an optimal trade-off between a maximum yield and a minimum environmental risk of N-leaching.

Discriminant pattern mining offers a way to answer such questions. A **discriminant or contrast pattern** (Li et al., 2005) is a motif (set of conditions) whose occurrence correlates with a particular class of scenarios. For example, a study (Le Bars, 2019) conducted on the STICS simulations reveals that parcels located in regions with *an oceanic climate, an average draining, and an average mineralisation* are 15 times more prone to high N-leaching, i.e., the growth ratio of this pattern is 15. Such an analysis is possible given that someone has defined the classes high and low N-leaching. In the presence of multiple classes, e.g., high, medium, and low N-leaching, discriminant pattern mining methods require a reformulation of the problem into a two-class setting. This re-discretization is usually possible, but it comes at the expense of accuracy. Depending on the applied re-discretization, e.g., a pattern that occurs frequently in the class of medium N-leaching could still rank as discriminant for the class of high N-leaching. To alleviate this problem we need to re-define the concept of discriminant patterns for multiple classes. In this line of thought we could answer more fine-grained questions such as "which conditions are strongly observed in the scenarios of high N-leaching, but almost never observed in each of the other scenarios (medium and low N-leaching)?" To answer this particular question we would need to run a discriminative pattern mining twice, which can be time consuming for large datasets such as the collection of STICS simulations.

The multi-class setting shows one limitation of discriminant patterns, however multi-dimensional data poses additional challenges as well. In the example of STICS, one may want to investigate whether a pattern that discriminates the scenarios of low leaching *persists* across multiple levels of aggregation, e.g., parcel, department, or administrative region. Again, this scenario requires us to run a discriminative pattern mining algorithm for each level and see which are the common patterns. The natural research question is therefore, can we do better than running the algorithm multiple times?

Internship description

The internship will therefore study ways to extend the notion of discriminant patterns to multi-class and multi-dimensional settings as described in the introduction. We count on the data produced by STICS, which has been modelled as a data cube (Le Bars, 2019). The methods will be tested on this repository. Furthermore, we will rely on the domain expertise of Anne-Isabelle Graux from INRA to validate the utility of the patterns produced by the developed algorithms.

References

Brisson N., et al., 2003. An overview of the crop model STICS. European Journal of Agronomy, 18, 309-332.

Graux A.-I., et al., 2017. Les prairies françaises: production, exportation d'azote et risques de lessivage. Rapport d'étude, INRA (France), 74 p.

Graux, A.-I., Resmond, R., Casellas, E., Delaby, L., Faverdin, P., Le Bas, C., Ripoche, D., Ruget, F., Therond, O., Vertès, F., Peyraud, J.-L. (2019). High-resolution assessment of French grassland dry matter and nitrogen yields. *European Journal of Agronomy* (under press)..

Haiquan Li, Jinyan Li, Limsoon Wong, Mengling Feng, and Yap-Peng Tan, 2005. Relative risk and odds ratio: a data mining perspective. PODS. DOI=http://dx.doi.org/10.1145/1065167.1065215

Nedjar, Sébastien, et al. "Emerging cubes for trends analysis in OLAP databases." *International Conference on Data Warehousing and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2007

Sophie Le Bars. Conception d'un entrepôt de données pour une visualisation et interrogation des données de simulation en vue de répondre à des questions agronomiques. 2019. Rapport de stage M2

Ugarte, Willy, et al. "Compressing and Querying Skypattern Cubes." *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, Cham, 2019.