**Université de technologie de Compiègne – Thesis proposal**

| Part 1: Scientific sheet | |
|---|---|
| Thesis proposal title | **Probabilistic Graphical Models for Complex Learning Tasks** |
| PhD grant | PhD contract based on a Ministry of Research Grant |
| Host laboratory | Research team: CID, Heudiasyc UMR 7253 |
| Thesis supervisors | Vu-Linh NGUYEN (Junior/Assistant professor)<br>Yves GRANDVALET (CNRS senior researcher) |
| Scientific domain(s) | Computer science |
| Research work | In the context of machine learning, much work has been devoted to probabilistic classifiers that predict the probability distribution of a single class variable. These models allow to compute the optimal predictions under different loss functions as well as information about the uncertainty associated with their predictions.<br>Multidimensional classification (MDC) is a generalization of multi-class classification in which several non-exclusive class variables are predicted, in a model that accounts for dependencies between these class variables [3, 4]. In this framework, multidimensional Bayesian networks (MBNs) [3] is one of the few families of predictive models that can be interpreted, by providing a possibly compact representation of the relationships between different variables (features and classes). However, interpretable and reliable MBNs still fail to solve difficult real-world problems, whose descriptors include continuous and discrete features as well as high-dimensional objects such as images.<br>To the best of our knowledge, the ongoing project "Multidimensional Probabilistic Classification" (initiated by Linh at Eindhoven University of Technology) is the first attempt to address this problem in a generic and rigorous way. Our ambitions are to implement the developed methods in high-stakes applications, such as health data analysis or autonomous vehicles. We will build on advances made on several fronts: in probabilistic graphical model learning [1], in optimization [2, 8], and in deep learning [5] to develop efficient learning and inference algorithms generalizing those that exist for multi-label classification (see [6, 7] and references therein).<br>There are many specific questions that deserve to be addressed in the MDC. Below we mention three possible starting points for the thesis work. The choice of one of them can be left to the student according to his or her tastes and skills.<br>*Regularization of the learning problem.* A multidimensional classifier is defined by the structure of the network modeling the dependencies between variables and its parameters. Learning can return a multidimensional classifier that is too complex. In this situation, a general approach is to develop regularization functions that favor simple models. We have developed a family of regularized conditional likelihood functions. Their improvement and implementation in a Python package could be a first step of the thesis.<br>*Incorporation of logical rules in inference.* Reliable and interpretable logical rules (AND, OR, XOR, ...) can sometimes be extracted from the data or provided by domain experts. However, inference techniques for MDC are not able to integrate this information to make their predictions. We have developed new Bayesian inference techniques in which the estimation of the most probable predictions is formalized as an integer programming problem, where logical rules are encoded as numerical constraints. The efficient encoding of composite logical rules and their combinations requires new developments that demand strong computer skills.<br>*Active learning.* The modeling of the MDC problem by a Bayesian network offers a potential for the development of scalable approaches for recommending the acquisition of the most influential and relevant non-observed variables (according to a performance criterion). These approaches correspond to active learning strategies requiring efficient search algorithms to compute optimal queries. |
| Keywords | Multi-dimensional classification, Multimodal learning, Interpretability, Computational complexity, Scalability |

| Part 2: Job description | |
|---|---|
| Starting time | 01/10/2023 |
| Duration | 36 months |
| Research laboratory | Heudiasyc UMR 7253, Université de Technologie de Compiègne |
| Requirements | Master 2 or engineer in computer science, good programming skills (Python, PyTorch, TensorFlow, ...) and/or a strong background in mathematics. |
| Additional missions | Teaching is possible, but not mandatory |
| Material resources | Shared office, laptop, access to the laboratory's GPU servers and the Jean Zay supercomputer installed at IDRIS, as well as to the laboratory's platforms, ... |
| Human resources | Internal and external collaborations |
| Working conditions | The supervision team proposes a two-stage pedagogical project. During the first stage, the student will be guided in the choice of results, algorithmic solutions, existing software packages, on which to base their own results, software, experimental protocols, and to communicate them through scientific articles.<br>Once the necessary knowledge and skills have been acquired, the student will be invited to tackle more difficult problems in collaboration with supervisors and collaborators, to develop the research skills to work both independently and collaboratively.<br>The candidate will be funded by Allocation MESR, and will get financial support for travel (conferences, workshops, summer schools, ...). |
| Collaborations | We collaborate with Cassio de Campos, UAI research group, Eindhoven University of Technology, The Netherlands. Cassio is a specialist in probabilistic graphical models. He is willing to host the student for a 3-6 month internship when the student is ready and if (s)he wishes. |
| Contact | Applications should include a letter of motivation, a curriculum vitae and contact information of at least one reference.<br>Applications and questions should be sent to:<br>• Vu-Linh Nguyen (vu-linh.nguyen@hds.utc.fr)<br>• Yves Grandvalet (yves.grandvalet@hds.utc.fr) |

# References

[1] J. Cussens, M. Järvisalo, J. H. Korhonen, and M. Bartlett. Bayesian network structure learning with integer programming: Polytopes, facets and complexity. *Journal of Artificial Intelligence Research*, 58: 185–229, 2017.

[2] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2): 131–163, 1997.

[3] S. Gil-Begue, C. Bielza, and P. Larrañaga. Multi-dimensional bayesian network classifiers: A survey. *Artificial Intelligence Review*, 54(1):519–559, 2021.

[4] B.-B. Jia and M.-L. Zhang. Decomposition-based classifier chains for multi-dimensional classification. *IEEE Transactions on Artificial Intelligence*, 3(2):176–191, 2021.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[6] V.-L. Nguyen and E. Hüllermeier. Multilabel classification with partial abstention: Bayes-optimal prediction under label independence. *Journal of Artificial Intelligence Research*, 72:613–665, 2021.

[7] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains: a review and perspectives. *Journal of Artificial Intelligence Research*, 70:683–718, 2021.

[8] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296, 2005.