Statistical learning and random forests for spatio-temporal data: application to wireless sensor networks data.

Fully funded PhD Position

Institutions:	Université Bretagne Sud, LMBA CNRS 6205
Location:	Campus Tohannic, Vannes, France
Advisors:	Audrey Poterie & François Septier
Contact:	`audrey.poterie@univ-ubs.fr`, `francois.septier@univ-ubs.fr`

Subject

During the past decades, wireless sensor networks (WSN) have attracted considerable attention due to the large number of applications in various fields, such as environmental monitoring, weather, health care and fire detection. In addition, WSN technology has been identified as one of the key components in designing future Internet of things (IoT) platforms. A WSN typically consists of a set of spatially distributed sensors that have generally limited resources, such as energy and memory. These sensors monitor a spatio-temporal phenomenon of interest that contains some desired attributes (e.g. wind speed, seismic activity, temperature, concentrations of substance, etc.).

In a centralized setting (the "ideal" situation), the sensors are assumed to be able to communicate regularly their observations to a base station (BS). The BS collects all these observations and fuses them in order to detect, predict or reconstruct the signal of interest, based on which effective management actions are made. Unfortunately, in practice, owing to the inherently resource constraints of the sensors (e.g. power, connectivity), the inference task has to be performed in a decentralized manner which requires sensor nodes to communicate only with their one-hop neighbors. Furthermore, in very large WSNs, using centralized sensor communication is often not possible. Since the rise of WSNs, many algorithms have been developed to improve the accuracy of such a constrained network to solve the challenging task of interest. Nowadays, these algorithms have seen increasingly intensive adoption of advanced machine learning (ML) techniques such as neural networks or decision trees, see [1] for a survey.

In this project, we will focus more especially on the study of the random forest algorithm in the context of WSN data. The aim of the PhD is therefore to propose rigorous and efficient random forests methods for spatio-temporal data. These new algorithms will be more especially developed to handle WSN data.

Random forest (RF), originally proposed by [2], is part of the most successful statistical methods currently used to handle problems in supervised statistical learning. The popularity of RF can be mainly explained by the fact that it is easy to implement and the method can be applied to a wide range of applications in various fields such as for example medicine [3, 4] and ecology [5]. Although some applications on times series [6] and spatio-temporal data [7] could be found and a variant of RF have been recently proposed for time series [8], RF does not in essence take account of the space-time dependent structure of the data.

So using RF to deal with WSN data remains quite challenging and some of the main issues are:

1. As mentioned previously, by assuming that data are independent and identically distributed, RF does not integrate the space-time dependent structure of the data.

2. RF, as most of the ML models, does not need rigid statistical assumptions about the data contrary to parametric models. However, compared with a parametric approach, these methods generally require larger datasets which could be complicated to obtain in real-life scenarios, especially in the decentralized setting when we only observations from a very small number of sensors.

3. The resource constraints of each sensor imply a trade-off between the model accuracy and its computational cost.

4. RF fails to make prediction beyond the range in the training data (extrapolation). When dealing with WSN data, extrapolation methods are frequently used to address lots of problems such as for instance the search for the optimal position of a new sensor or the efficient prediction of a phenomenon of interest not only at the locations of the actual sensors but at all locations.

We propose firstly to explore the current state-of-art work of ML methods, especially RF, in the context of data with a space-time dependent structure, and next to develop new RF approaches for WSN data. Methods commonly used to make inference with WSN data, as for instance the methods involving gaussian processes [9], will be also studied. Then novel techniques integrating both these methods and RF could be also proposed in order to overcome some limitations of the gaussian process methods when dealing with WSN data. First of all, the PhD thesis will be focused on centralized WSN. Next, the context of networks with sensors that communicate in a decentralized way will be addressed and the methods introduced previously for centralized WSN could be extended to this more challenging situation. Extensive simulation studies and applications on real WSN data will be performed in order to assess the performances of each proposed approach.

Candidate profile

We are looking for a motivated and talented student who should:

- Hold a master's degree in applied mathematics: probability/statistics, machine learning, data science or signal processing,
- Have a strong backgroung in scientific programming, preferably in R and/or Python.
- Have English skills allowing scientific communication (oral/reading/writing).

Details

A fully funded PhD position (three-year contract) is available from September/October 2020 at the Université Bretagne Sud located at Campus Tohannic in Vannes [link]. The student will enjoy an international and creative environment where research seminars and reading groups take place very often. This project will also benefit from strong collaborations with Dr. Ido Nevat [link], senior researcher at TUMCreate (Singapour). Indeed, the methods developed in this project could be assessed and validated on some data from the project Cooling Singapore [link] whose Dr. Ido NEVAT is one of the leaders.

The student will be supervised by:

- Audrey Poterie [link]: `audrey.poterie@univ-ubs.fr`
- François Septier [link]: `francois.septier@univ-ubs.fr`

The candidate is requested to firstly send us a CV and a motivation letter to apply for this position.

References

[1] D. P. Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: A survey," *Information Fusion*, vol. 49, pp. 1–25, 2019.

[2] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] A. Poterie, J. Dupuy, V. Monbet, and L. Rouvière, “Classification tree algorithm for grouped variables,” *Computational Statistics*, vol. 34, p. 1613–1648, 2019.

[4] R. Diaz-Uriarte and S. A. De Andres, “Gene selection and classification of microarray data using random forest,” *BMC bioinformatics*, vol. 7, no. 1, pp. 1–3, 2006.

[5] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, “Random forests for classification in ecology,” *Ecology*, vol. 88, no. 11, p. 2783–2792, 2007.

[6] A. Fischer, L. Montuelle, M. Mougeot, and D. Picard, “Statistical learning for wind power: A modeling and stability study towards forecasting,” *Wind Energy*, vol. 20, no. 12, p. 2037–2047, 2017.

[7] S. Georganos, T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, , and S. Kalogirou, “Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling,” *Geocarto International*, vol. 7, no. 1, pp. 1–16, 2019.

[8] P. Joslin, “Prévision multi-échelle par agrégation de forêts aléatoires. application à la consommation électrique.” Ph.D. dissertation, Thèse de doctorat de Mathématiques appliquées, Université Paris Saclay, 2019.

[9] P. Zhang, I. Nevat, G. W. Peters, F. Septier, and M. A. Osborne, “Spatial field reconstruction and sensor selection in heterogeneous sensor networks with stochastic energy harvesting,” *Geocarto International*, vol. 66, no. 9, p. 2245–2257, 2018.