Statistical learning and random forests for spatio-temporal data

Application to wireless sensor networks data to predict emergency situations

Institution	:	Univ. Bretagne Sud, Univ. Bretagne Occidentale, Lab-STICC UMR CNRS 6285.
Location	:	Campus Tohannic, Vannes, France.
Supervisors	:	Audrey Poterie & Ahcène Bounceur.
Contact	:	*audrey.poterie@univ-ubs.fr*

Keywords: random forests, spatio-temporal data, machine learning, statistical learning, wireless sensor networks.

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 (Hart and Martinez, 2006), weather forecasts (Rajasegarar et al., 2013), health care (Lorincz et al., 2004) and fire detection (Saoudi et al., 2017). 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 communication bandwidth. These sensors monitor a spatio-temporal phenomenon of interest containing some desired attributes (e.g. wind speed, seismic activity, temperature, concentrations of substance, etc.).

In a centralized setting, 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 (Nevat et al., 2015), 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. Futhermore, in very large WSNs, using centralized sensor communication is often not possible. Many algorithms have been developed in this context to improve the accuracy of such a constrained network to solve the challenging task of interest. During the past decades, WSNs have seen increasingly intensive adoption of advanced machine learning (ML) techniques such as neural networks or decision trees, see (Kumar et al., 2019) 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. Random forest (RF), originally proposed by Breiman (2001), is part of the most successful

statistical methods currently used to handle prediction problems. 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 medecine (Díaz-Uriarte and De Andres, 2006) and ecology (Cutler et al., 2007).

Although some applications on times series (Fischer et al., 2017) and spatio-temporal data (Hengl et al., 2018; Georganos et al., 2019) could be found and a variant of RF have been recently proposed for time series (Goehry, 2019), 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 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. However these methods generally require quite large datasets which could be complicated to obtain in real-life scenarios.
(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.

The aim of the PhD is therefore to propose rigorous and efficient statistical learning algorithms based on random forests to handle the complex space-time depedent structure of WSN data.

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 (Zhang et al., 2018), 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 decentralised 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. The simulation process will use CupCarbon (Bounceur et al., 2018), a powerful simulator of smart city and internet of things wireless sensor networks. This simulation tool generates realistic WSN scenarios in few steps and so it allows to clarify the implementation of a WSN before its real deployment. Another aspect of this PhD project could be the integration of all the proposed algorithms in CupCarbon, allowing thereby this simulator to be used not only as a visualization tool but also as a performance assessment tool.

Details

This PhD position (three-year contract) is available from September/October 2020 at the Université de Bretagne Sud located at Campus Tohannic in Vannes and in the DECIDE team of the CNRS laboratory Lab-

STICC. The student will also sometimes work at the Université de Bretagne Occidentale located in Brest.

The PhD is a partnership between the Université de Bretagne Sud and the Université de Bretagne Occidentale. During the PhD thesis, the student will also collaborate with François Septier, Marc Sevaux and other researchers at both the Université de Bretagne Sud and the Université de Bretagne Occidentale.

Candidate profile and application process

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).

To apply for this position, the candidate is requested to send us the following documents in English or in French:

- a CV,
- a covering letter,
- a proof of Master's Degree (if already obtained) and BsC (*licence* in French),
- the marks obtained in Master (the two years),
- transcripts of both your Bachelor and your Master,
- two contacts references.

These documents must be emailled to the address *audrey.poterie@univ-ubs.fr*.

References

Bounceur, A., Marc, O., Lounis, M., Soler, J., Clavier, L., Combeau, P., Vauzelle, R., Lagadec, L., Euler, R., Bezoui, M., et al. (2018). Cupcarbon-lab: An iot emulator. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–2. IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Fischer, A., Montuelle, L., Mougeot, M., and Picard, D. (2017). Statistical learning for wind power: A modeling and stability study towards forecasting. *Wind Energy*, 20(12):2037–2047.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, pages 1–16.

Goehry, B. (2019). Random forests for time-dependent processes. working paper or preprint.

Hart, J. K. and Martinez, K. (2006). Environmental sensor networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3-4):177–191.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518.

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

Lorincz, K., Malan, D. J., Fulford-Jones, T. R., Nawoj, A., Clavel, A., Shnayder, V., Mainland, G., Welsh, M., and Moulton, S. (2004). Sensor networks for emergency response: challenges and opportunities. *IEEE pervasive Computing*, 3(4):16–23.

Nevat, I., Peters, G. W., Septier, F., and Matsui, T. (2015). Estimation of spatially correlated random fields in heterogeneous wireless sensor networks. *IEEE Transactions on Signal Processing*, 63(10):2597–2609.

Rajasegarar, S., Havens, T. C., Karunasekera, S., Leckie, C., Bezdek, J. C., Jamriska, M., Gunatilaka, A., Skvortsov, A., and Palaniswami, M. (2013). High-resolution monitoring of atmospheric pollutants using a system of low-cost sensors. *IEEE transactions on geoscience and remote sensing*, 52(7):3823–3832.

Saoudi, M., Lalem, F., Bounceur, A., Euler, R., Kechadi, M.-T., Laouid, A., Bezoui, M., and Sevaux, M. (2017). D-lpcn: A distributed least polar-angle connected node algorithm for finding the boundary of a wireless sensor network. *Ad Hoc Networks*, 56:56–71.

Zhang, P., Nevat, I., Peters, G. W., Septier, F., and Osborne, M. A. (2018). Spatial field reconstruction and sensor selection in heterogeneous sensor networks with stochastic energy harvesting. *IEEE Transactions on Signal Processing*, 66(9):2245–2257.