Adaptive Model Selection for Streaming of Wireless Sensor Data

Y. Le Borgne[1], S. Santini[2], and G. Bontempi[1]

[1]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
[2]Distributed Systems Group, ETH Zurich (Swiss Federal Institute of Technology Zurich), Switzerland

Abstract

In many practical applications of wireless sensor networks, the sensor nodes are required to report approximation of their readings at regular time intervals. For these applications, it has been shown that time series prediction techniques provide an effective way to reduce the communication effort while guaranteeing user-specified accuracy requirements on collected data. Achievable communication savings offered by time series prediction however strongly depend on the type of signal sensed, and a priori choice of a prediction technique is not trivial in practice. We propose an online and lightweight algorithm dubbed adaptive model selection (AMS), that allows sensor nodes to autonomously determine the best performing model among a set of candidate models. Experimental results obtained on the basis of fourteen real-world sensor time series demonstrate the efficiency and versatility of the proposed framework in reducing the communication effort between sensor nodes.

1. Wireless sensors networks (WSN) - Applications

In a WSN, sensor nodes self-organize and wirelessly communicate their readings to recipients such as databases or controllers.

Figure 1: *Tmote Sky sensor node [3].*

Tmote Sky:
Microcontroller 8MHz
10KB RAM, 48KB Code, 512KB Data
Radio 155Kbps @ 2.4GHz

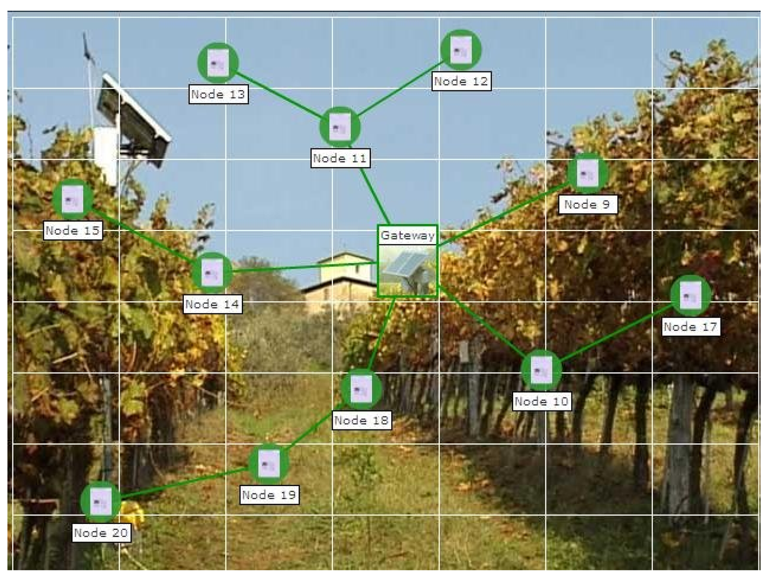

Figure 2: *Goodfood project: Precision agriculture in a vineyard.* Courtesy of the Goodfood project [5].

Examples of application:

- Monitoring: Ecosystems, industry processes, battlefields, ...
- Control: Building automation, precision agriculture, medical healthcare, ...

2. Constraints

- **Energy**: Energy consumption of sensor nodes is a primary concern in WSN as it is directly related to the lifetime of the network. Radio is known to be the main factor of battery depletion on a sensor node.
- **Bandwidth**: Bandwidth capacity does not scale with network size. This directly impacts the resolution at which sensors can be deployed in the environment.

Challenges: Reduce the communication effort among sensor nodes, to both increase network operation lifetime and environment network covering.

3. Dual Prediction Scheme (DPS)

In many applications, only an ϵ approximation to sensor readings is sufficient (e.g. ± 0.5°C, ± 2% humidity, ..). The rationale of the DPS is as follows [2, 4]:

- A sensor node is provided with a time series prediction model $\hat{X}_{t+1} = h(X_h, \theta_h)$ (e.g. autoregressive models) and a learning method for identifying the best set of parameters θ_h (e.g. recursive least squares).
- The sensor node then sends the parameters of the model instead of the actual data to the recipient. The recipient node then runs the models to reconstruct an approximation of the data stream collected on the distant node.
- The sensor node also runs the prediction model. When its predictions diverges by more than $\pm\epsilon$ from the actual reading, a new model is sent to the recipient.

This allows to reduce the communication effort if an appropriate model is run by the sensor node.

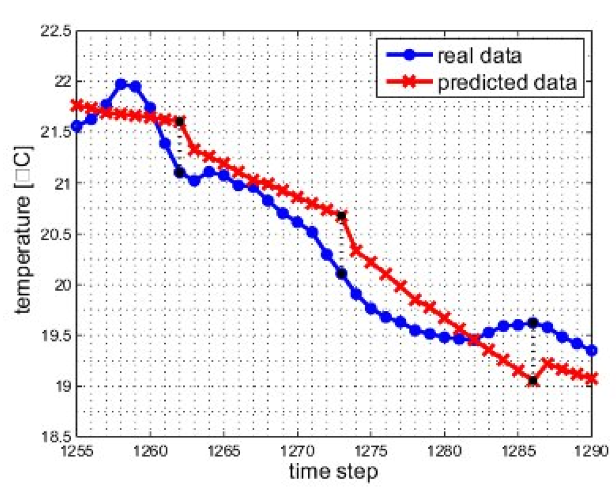

Figure 3: *Prediction based on autoregressive model of order 5.*

- Data stream: Temperature on a grape plant [5].
- Required accuracy: ± 0.5°C.
- At time instants 1262, 1274 and 1286 (vertical dashed black lines), the prediction model diverged by more than ± 0.5°C. A new model (together with the correction for the badly predicted measurement) was therefore sent.
- About 90% communication savings were achievable on that time series.

4. Adaptive Model Selection (AMS)

As the most appropriate model for a sensor data stream cannot be known a priori, the rationale of AMS is to run concurrently a set of different models on the sensor node, and to discard poorly performing ones on the basis of the racing mechanism [1].

RACING

Let U_1^N and U_2^N be the estimated performances of two prediction models h_1 and h_2 after N time instants, and let R be the range of values possibly taken by U_i^N. The racing relies on the Hoeffding bound, a distribution free statistical bound, to assume that if

$$U_1^N - U_2^N > R\sqrt{\frac{ln(1/\delta)}{2N}}$$

Then h_1 is statistically outperformed by h_2, and can be removed from the set of concurrent models.

5. Experimental results

We assessed the performances achievable with AMS in terms of reduction of sent packets on a set of 14 real world time series, using as competing models the constant model and autoregressive models of order 1 to 5.

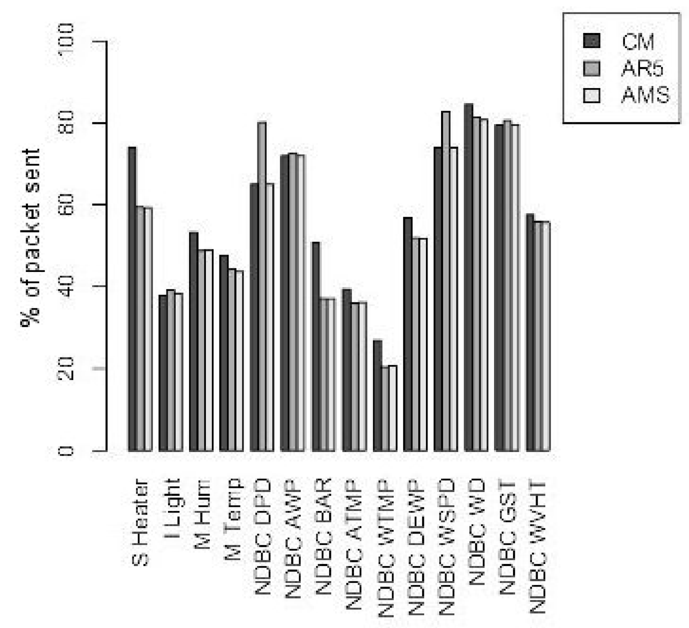

Figure 4: *Comparison of constant model (CM), AR(5) and AMS in terms of percentage of sent packets.*

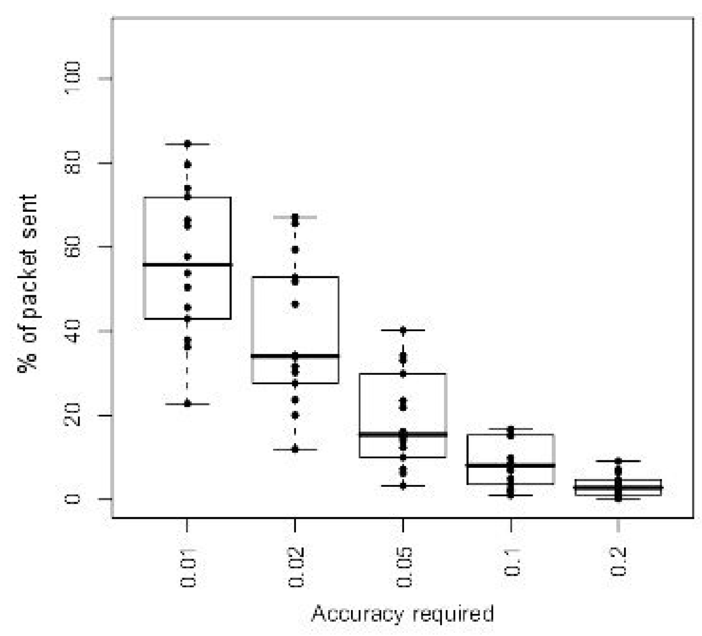

Figure 5: *Percentage of sent packets as a function of error tolerance.*

- Figure 4 reports the achievable reduction in terms of number of sent packets for an accuracy of $0.02 * r$, r being the dynamic of the signal. While the proposed scheme necessarily reduce the number of sent packets, its efficiency depends on the characteristics of the sensor signal. For example, NDBC WSPD (Offshore wind direction, buoy data) is characterized by sharp and sudden changes. The constant model performs best in this case. NDBC WTMP data (Offshore water temperature, buoy data), are however much smoother. Autoregressive model of order 5 was seen to perform best. In all cases, the AMS eventually selected the best model.
- Figure 5 reports achievable reduction in terms of number of sent packets as a function of required accuracy. Even for very high accuracy (error tolerance of $0.01 * r$), almost half of communication could on average be avoided. For rough approximations (error tolerance of $0.2 * r$), only 4% of communication were enough for reconstructing the signal on the recipient side.
- The racing mechanism could in most cases reduce the number of competing models to the best one.

6. Conclusion and future work

Given a set of candidate prediction models, the proposed AMS scheme allows to select in a fully online manner the statistically best performing one for approximating a sensor signal. Future work will consist in

- Assessing achievable performances with exponential smoothing models, which could lead to greater communication savings.
- Adapting the scheme to non-stationary signals.

This research was supported by the European Community **COMP[2]SYS** (MEST-CT-2004-505079) and Embedded WiSeNts (FP6-2003-IST-2, Contract Nr. 04400) projects.

References

[1] O. Maron and A. W. Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1-5):193225, February 1997.

[2] C. Olston, J. Jiang, and J. Widom. Adaptive Filters for Continuous Queries over Distributed Data Streams. In *Proceedings of the 2003 ACM SIGMOD Intl. Conf. on Management of data (SIGMOD '03:)*, pages 563–574, New York, NY, USA, 2003. ACM Press.

[3] J. Polastre, R. Szewczyk, and D. Culler. Telos: Enabling ultra-low power wireless research. In *Proceedings of the 4th Intl. Conf. on Information Processing in Sensor Networks: Special track on Platform Tools and Design Methods for Network Embedded Sensors (IPSN/SPOTS)*, April 2005.

[4] Silvia Santini and Kay Römer. An adaptive strategy for quality-based data reduction in wireless sensor networks. In *Proceedings of the 3rd Intl. Conf. on Networked Sensing Systems (INSS'06)*, Chicago, IL, USA, June 2006.

[5] Sensor Network in a Vineyard. GoodFood EU Integrated Project: Food Safety and Quality Monitoring with Microsystems. Project Website: www3.unifi.it/midra/goodfood/.