Using more realistic data models to evaluate sensor network data processing algorithms

Yan Yu, Deborah Estrin, Mohammad Rahimi
UCLA/CENS
yanyu, destrin@cs.ucla.edu, mhr@cens.ucla.edu

Ramesh Govindan
USC/ISI
ramesh@usc.edu

Abstract

Due to lack of experimental data and sophisticated models derived from such data, most data processing algorithms from the sensor network literature are evaluated with data generated from simple parametric models. Unfortunately, the type of data input used in the evaluation often significantly affect the algorithm performance. Our case studies of a few widely-studied sensor networks data processing algorithms demonstrated the need to evaluate algorithms with data across a range of parameters. In the end, we propose our synthetic data generation framework.

1. Introduction

Sensor network research is still in its infancy. Due to lack of experimental data from deployed systems and sophisticated models derived from such data, most data processing algorithms from the sensor network literature are evaluated with data generated from simple parametric models (*i.e.*, models defined by a set of parameters). For example, uniform or Gaussian data input has been commonly used to evaluate data collection and estimation algorithms. Unfortunately, the type of data input used in the evaluation often significantly affect the algorithm performance.

We identify a few widely-studied classes of problems that are potentially sensitive to data input: Statistics estimation of the field data; Data compression; and Field estimation. We use them as examples to investigate the dependency of algorithm performance on data. Due to the space limitations, we present results on instances of the first and third class of problems, namely, a median computation and an adaptive sampling algorithm. Using experimental data sets, we demonstrate how different data input can change the algorithm performance dramatically, the performance comparison between two algorithms may even change depending on the different data inputs. Further, we propose an experimental-oriented synthetic data generation framework to generate realistic data sets with a wide range of parameters.

2. Case studies of the algorithm's performance dependency on data input

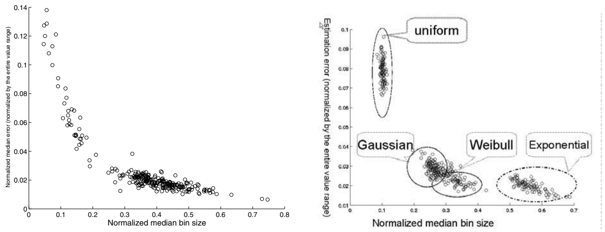
Median Computation: We evaluate a uniform sampling based median computation algorithm against 4 data sets: data generated from uniform, Gaussian, and Bimodal distribution; and the S-Pol radar data set[1].

Our statistical analysis consists of three key steps. First, we define our performance metric to be *normalized estimated median error*, defined as *the normalized difference between the estimated median and the real median*. Second, we identify the relevant data characteristic to be *normalized median bin size*, which is defined as the ratio of the size of the median bin relative to the size of the entire data set. In an equally spaced histogram, *median bin* is the bin that includes the median. As a final step, we statistically study how the algorithm performance varies with changing data characteristics. Both scatter plots (Figure 1) and correlation coefficients suggest that the algorithm performance is well correlated with our defined data characteristic. Further, in terms of the normalized median bin size, the experimental data sets cover a wide range of parameter space not covered by any single distribution (for each distribution family, we vary the parameter across a wide range). We believe this result strongly suggest the importance of experimental data in algorithm evaluations.

Adaptive Sampling: As an alternative to raster scan, Fidelity Driven Sampling [1] is an efficient way to sample the environmental field. Following the Fidelity Driven Sampling operation (or raster scanning data acquisition), the returned sample points are used to reconstruct the environmental field. We define the performance evaluation met-

[1] S-Pol (S band polar metric radar) data were collected during the International H 2O Project. The S-Pol radar data provided by NCAR records the intensity of reflectivity in dBZ. We acknowledge NCAR and its sponsor, the National Science Foundation, for provision of the S-Pol data set.

IEEE COMPUTER SOCIETY



(a) Results on experimental data, correlation coefficient= -0.8239

(b) Results on Gaussian, uniform, Weibull and exponential distributed data, correlation coefficient=-0.7818

Figure 1. Scatter plot of normalized estimated median error vs. normalized median bin size: with increasing normalized median bin size, the estimation error decreases. Further, the experimental data covers a super set of all 4 families of data distributions and more.

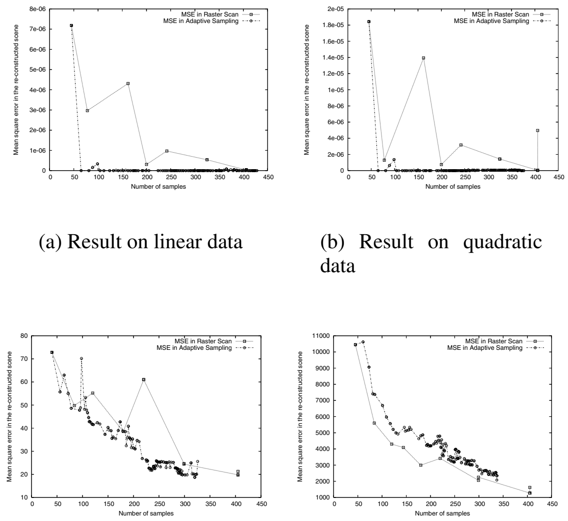
ric as the Mean Squared Error (MSE) between this reconstructed field map and the ground truth.

When evaluated with data simulated from linear and quadratic models, the Adaptive Sampling performs several magnitudes better than Raster Scan in terms of MSE (Figure 2(a)- 2(b)). However, when evaluated with the experimental data, the MSE obtained from Adaptive Sampling is very close to or worse than the MSE obtained from Raster Scan (Figure 2(c), 2(d)). This may suggest that the evaluation results from data input solely based on simple parametric models may be misleading. Evaluating algorithms using experimental data with various features helps identify the regime of the parameter space where the algorithm may perform well compared to other alternatives.

3. Algorithm evaluation with realistic data input across a wide range of parameters

The huge parameter space of data input makes exhaustive exploration of parametric models impractical. By driving simulations from previously collected experimental data, we focus our testing on the part of parameter space that matters in reality.

Existing experimental data is often collected from regular grids, whereas real deployments may have an irregular topology. Leveraging the existing experimental data we propose to generate synthetic data of irregular topology from modeling the experimental data. Our proposed synthetic data generation techniques attempt to approximate the experimental data in terms of distribution, spatial correlation, or other features of interest. Our synthetic data generation tool-box introduced in [2] includes eight spatial interpolation algorithms, which in turn will generate eight different data sets based on one single experimental data set. Which synthetic data set is more desirable depends on the specific application and algorithm in the study.



(a) Result on linear data

(b) Result on quadratic data

(c) Result on data with smooth curvature

(d) Result on data with rough curvature

Figure 2. comparison of Adaptive Sampling vs. Raster Scan: for data generated from linear and quadratic models, AS performs several magnitudes better than RS; however, for experimental data collected from a lab environment, AS is very close to or worse than RS.

Since it is difficult to design an algorithm that is provably insensitive to data input and it is hard to predict the statistics of the data that is going to be sampled, we recommend evaluating algorithms with data across a range of parameters, and investigate how the algorithm's performance changes with different data characteristics. Our proposed synthetic data generation approach can be used to generate realistic data sets with a wide range of parameters.

References

[1] M. A. Batalin, M. Rahimi, Y.Yu, D.Liu, A.Kansal, G. Sukhatme, W. Kaiser, M.Hansen, G. J. Pottie, M. Srivastava, and D. Estrin. Towards event-aware adaptive sampling using static and mobile nodes. In *Sensys*, 2004.

[2] Y. Yu, D. Ganesan, L. Girod, D. Estrin, and R. Govindan. Synthetic data generation to support irregular sampling in sensor networks. In *Geo Sensor Networks*. Taylor and Francis Publishers, Oct 2003.