Poster Abstract: Online Data Cleaning in Wireless Sensor Networks

Eiman Elnahrawy
Department of Computer Science
Rutgers University, USA
eiman@paul.rutgers.edu

Badri Nath
Department of Computer Science
Rutgers University, USA
badri@cs.rutgers.edu

ABSTRACT

We present our ongoing work on data quality problems in sensor networks. Specifically, we deal with the problems of outliers, missing information, and noise. We propose an approach for modeling and online learning of spatio-temporal correlations in sensor networks. We utilize the learned correlations to discover outliers and recover missing information. We also propose a Bayesian approach for reducing the effect of noise on sensor data online.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]

General Terms

Algorithms, Design

Keywords

Noisy Sensors, Outliers, Fault-Tolerance, Bayesian Theory

1. INTRODUCTION

With the wide deployment of wireless sensor networks it is not only possible to obtain a fine grain real-time information about the physical world but also to act upon this information. Unfortunately, data obtained from sensor networks usually suffers from two problems: missing information and outliers from malicious sensors. There are several factors that contribute to the existence of these problems such as packet loss and collisions, low battery levels and nonsensical readings arising from sensors when they are about to run out of power, potential harsh environmental conditions, and the technology and the quality of the sensors.

In addition, wireless sensors are subject to numerous sources of errors that can be classified broadly as either systematic (bias) or random (noise). Systematic errors arise due to changes in the operating conditions, or other factors such as ageing of the sensor. They can be corrected by calibration [1]. The sources of random errors, on the other hand, are numerous and cannot be recovered by calibration. They include, but are not limited to, noise from external sources, random hardware noise, inaccuracies and imprecision, and various random environmental effects. Several examples from the current technology reveal that sensors vary significantly in their precision and accuracy, tolerance to noise, etc., based on their type, cost and application. The aim of the industry however is to manufacture tiny cheap sensors that can be scattered everywhere and disposed when they run out of their batteries. Therefore, intolerance to noise, imprecision and inaccuracies are inevitable and highly expected among those cheap sensors.

The quality, reliability, and timeliness of data obtained from sensor networks are extremely important since detrimental actions are usually taken based upon these sensed values. Specifically, the cost of any "unclean" data can be very significant since it is usually used for critical decisions (possibly immediate) or activation of actuators. All the above problems however seriously impact the data obtained from such networks. They imply that they must operate with imprecise, imperfect or incomplete information. Hence, our current work is focused on online energy-efficient techniques for handling these various issues in sensor data. The rest of this report is organized as follows. Section 2 presents our work on (1) detecting outliers in sensor data and discovering malicious sensors, and (2) handling missing sensor data. Section 3 introduces our approach to reducing the effect of noise on sensor data. Finally, Section 4 concludes this report.

2. CONTEXT-AWARENESS

This section introduces our approach to handling outliers and missing information in sensor networks. It also presents our preliminary results and future work directions.

2.1 Approach

Our approach is based on exploiting the spatio-temporal relationships that exist among sensors, e.g., in monitoring or tracking applications [6, 5]. These networks are generally dense with redundant and correlated readings for coverage and connectivity purposes and for tolerating network failures [6, 4, 5]. We present an approach for modeling such correlations and we introduce a scalable energy-efficient technique for learning them online. We also present procedures for utilizing these correlations in detecting malicious sensors and discovering outliers, predicting any missing value, and effectively reducing energy consumption in the network by sampling. We refer to such spatio-temporal correlations as the contextual information of the network. Contextual information encodes the spatial dependencies between spatially adjacent nodes in the network as well as the temporal dependencies between history readings of the same sensor node. It therefore enables sensors to locally predict their current readings knowing both their own past readings and the current readings of their neighbors. In other words, sensors are aware of their context (the neighborhood and history).

Our approach is based on statistical Bayesian classifiers; we map the problem of learning and utilizing contextual information to the problem of learning the parameters of a Bayes classifier, and then making inferences, respectively [3]. The different values of sensor readings represent the different classes. We quantize the range of

SenSys'03, November 5–7, 2003, Los Angeles, California, USA..
ACM 1-58113-707-9/03/0011.

possible sensor values into a finite set of non-overlapping subintervals (classes), not necessarily of equal length. The features of classification are the current readings of some immediate neighbors of the sensor (spatial information), and the last reading of the sensor (temporal information), based on a Markovian assumption. To cut down the number of training instances needed for parameter-learning, we utilize the "Naive Bayes" assumption that states that the features are conditionally independent given the target class. The parameters of our classifier reduce to a set of probabilities, while the inference problem is to calculate the most likely class of the current sensor reading given the parameters and the observed spatio-temporal information. The inference is done probabilistically using maximum a posteriori. We propose an energy-efficient procedure for online learning of these parameters in-network, in a distributed fashion, where sensors collect training data and estimate the parameters locally.

2.2 Preliminary Evaluations

We evaluated our approach on phenomena with sharp boundaries such as those used in tracking [5], and on monitoring data collected from the Great Duck Island [6]. Our preliminary evaluations showed the applicability and a good performance of our approach. We achieved 90% accuracy in prediction in the two applications. Our major future work directions include building a prototype for our system and using it for more experimentations, applying our approach to different sensor applications and investigating several design and deployment issues such as working with heterogeneous sensors, the optimal number of neighbors, selecting the neighbors intelligently versus randomly, etc., and extending our approach to the multi-dimensional case.

3. SENSOR FUSION

This section introduces our approach to reducing the effect of noise on sensor data. It also presents our preliminary results and future work directions.

3.1 Approach

We are working on a framework for cleaning and querying noisy sensors [2]. Our overall framework consists of two major modules, a cleaning module and a query processing module. We focus on the cleaning part in this report. Our work on querying generally involves several statistical algorithms for answering traditional database queries over uncertain sensor readings.

We use a Bayesian approach for reducing the uncertainty associated with sensor data, that arise due to random noise, in an online fashion. Our approach yields more accurate estimates of the true "unknown" readings of noisy sensors. We consider both single-attribute and multi-attribute sensors. Our cleaning functionality is applied to every sensor. In general, even if the readings of a set of sensors are combined into a single more robust reading to reduce the effect of noise (i.e., fusion over multiple sensors), our approach is applied to the resultant single reading. Thus, yielding more accurate results. One of the applications that we are currently focusing on is monitoring the quality of food (perishable items) for inventory management.

There are three inputs to our cleaning module: the noisy sensor observations, metadata about the noise characteristics of every sensor (error model), and information about the distribution of the true reading at each sensor (prior knowledge). The error model of each sensor is basically the distribution of noise that affects it. We compute the parameters of the error model based on the specification of each sensor (accuracy, precision, etc.), and on testing calibrated sensors under normal deployment conditions. The error models may change over the time. They are stored as a metadata at the cleaning module. We are considering several sources of the prior knowledge and studying the quality of the approach in each case, e.g., using facts about the sensed phenomenon, learning over time (history), using less noisy readings as priors for the noisier ones, or even expert knowledge or subjective conjectures. We are also considering dynamic priors that changes at each time instance, if the sensed phenomena is known to follow a specific parametric model. Our approach in this case of dynamic priors resembles Kalman filters. The output of the cleaning module is probabilistic uncertainty models of the reading of each sensor (posterior), i.e., a probability density function of the true unknown sensor reading taking on different values. This output is computed using Bayes' rule.

The cleaning module can be placed at the sensor level or the base-station. We are studying the tradeoffs and explicit cost models for each option. This involves the communication and processing costs, and consequently energy consumption, and the storage cost. Bayesian approaches for reducing uncertainty have been used in literature in several fields. In general, it is a fundamental fact among the estimation theory community that the use of prior knowledge leads to more accurate estimators which motivated our approach. The novelty of our work however lies in designing an overall framework that utilizes Bayes' rule for online cleaning of noisy sensors, and in introducing and quantifying the different tradeoffs.

3.2 Preliminary Evaluations

Our preliminary results using synthetic data showed that this approach is efficient in reducing the uncertainty associated with noisy sensors. Our major future work direction is to integrate this framework with our approach for context-aware sensors in order to further improve the accuracy of the latter. Specifically, we will place a cleaning module at every sensor and compute a single estimate of its true reading from the computed posteriors online. We will then apply our approach for context-aware sensors on the resultant readings. We will build our overall prototype on calibrated sensors to achieve further improvement.

4. CONCLUSION

We reported on our ongoing work for online cleaning of sensor data online. Specifically, we discussed the following data quality problems: outliers, missing information, and noise. We briefly presented our preliminary evaluations. Finally, we introduced our major future research directions.

5. REFERENCES

[1] BYCHKOVSKIY, V., MEGERIAN, S., ESTRIN, D., AND POTKONJAK, M.A collaborative approach to in-place sensor calibration.In *Proceedings of IPSN'03* (2003).

[2] ELNAHRAWY, E., AND NATH, B.Cleaning and querying noisy sensors.In *Proceedings of ACM WSNA'03* (2003).

[3] ELNAHRAWY, E., AND NATH, B.Context-aware sensors.In *Proceedings of 1st European Workshop on Wireless Sensor Networks (EWSN)* (To appear, 2004).

[4] GANESAN, D., AND ESTRIN, D.DIMENSIONS: Why do we need a new data handling architecture for sensor networks?In *Proceedings of Hotnets-I* (2002).

[5] LIU, J., CHEUNG, P., GUIBAS, L., AND ZHAO, F.A dual-space approach to tracking and sensor management in wireless sensor networks.In *Proceedings of ACM WSNA'02*.

[6] MAINWARING, A., POLASTRE, J., SZEWCZYK, R., CULLER, D., AND ANDERSON, J.Wireless sensor networks for habitat monitoring.In *Proceedings of ACM WSNA'02*.