Statistical Modeling of Sensor Data and its Application to Outlier Detection

Christoph Heinz Bernhard Seeger

Department of Mathematics and Computer Science
Philipps University Marburg, Germany
{heinzch,seeger}@mathematik.uni-marburg.de

Abstract. Various applications rely on a continuous processing of data streams originating from a network of interconnected and collaborated sensors. The processing of those streams has turned out to be a difficult task as sensors only have limited resources and the data they produce is inherently uncertain and unreliable. In order to bridge the gap from raw, uncertain sensor readings to a meaningful model of the physical phenomenon observed, statistical modeling techniques have proved to be an adequate approach. By means of a statistical model, a wide range of sensor network related topics can be covered. In this work, we present an initial approach to tackle an important problem in sensor processing, namely the detection of outliers, with a statistical model.

1 Introduction

Recent advances in hardware technology combined with decreasing production cost for lightweight devices have facilitated the application of wireless sensor networks for monitoring entities in a plethora of real-world scenarios, e.g. environment monitoring, industry monitoring. In these scenarios, a large number of sensors is deployed and each one continuously monitors physical entities like temperature and air pressure. The raw sensor readings are typically transmitted to a central backend that provides an interface to query the sensor network.

The configuration of a sensor network comprises suitable settings for network topology, routing, and communication protocols [1]. Since the resources of a sensor are limited, the energy efficiency is a crucial factor in this configuration. Besides these low-level problems, a sensor network also has to cope with high-level problems: First, a sensor network produces huge amounts of data in form of transient streams. The time-critical nature of most applications combined with limited resources requires to give up the paradigm of exact answers and to use approximate answers instead [2]. Second, even if we could store all data, we must take the inherent uncertainty and unreliability of the data into account [3]. This uncertainty is due to the facts that a sensor can only provide discrete samples of a (continuous) physical phenomenon and that it is additionally subject to interfering effects like noise, hardware failures, inaccuracies. Hence, the querying of the raw sensor readings may produce misleading or even wrong answers.

Recently, the database research community has addressed those high-level problems. Especially research in data stream processing has come to the fore [2]. The processing of data streams is challenging as their requirements render the application of common database technologies unfeasible [4]. The processing of sensor data streams is even more challenging due to their inherent characteristics: data uncertainty, intra- and inter-stream correlations, sensitivity to energy consumption [5].

A promising approach that takes those characteristics into account is to incorporate statistical modeling techniques into sensor network processing. To convey a notion of the benefits of statistical models in sensor networks, let us sketch some recently proposed approaches: One, for example, is to equip query answers with probabilistic estimates of their validity to cope with imprecise sensor data [6]. In order to clean noisy sensors, [7] combines prior knowledge with noise characteristics of the sensor to obtain more accurate sensor readings. [3] also tackles the cleaning of sensor streams by exploiting temporal and spatial correlations of the sensor readings. The approach presented in [8] provides a statistical model for the complete sensor network that allows to query the network while acquiring new sensor readings only if necessary.

The above approaches share the property that they develop a statistical model for the distribution of the streams generated in a sensor network. The main assumption is that the sensor readings are samples of different physical phenomena under observation. If we describe the entirety of phenomena in terms of random variables, we have a variety of convenient statistical estimation techniques at our disposal. In this context, it is vital for further analysis to reveal the distribution of a random variable, more specifically its probability density function [9]. In [4], [10], we presented solutions to this problem for transient data streams. As these techniques particularly suit for the sensor stream scenario, we propose to use the statistical models they provide as point of origin for further analysis of the sensors. In this work, we present an initial approach to tackle outlier detection, an important task in almost every application on top of sensor networks, with the help of these statistical models. Before going into details, we will give a brief overview of the development of statistical models for sensor networks.

2 Statistical Modeling of Sensor Data

A sensor network acquires samples of physical phenomena with sensors located at the network nodes. We assume that each sensor measures at each time instant a single, real-valued attribute X_i, e.g. temperature. The sensor transmits the raw readings to a central basestation. With respect to the aforementioned unreliability of those readings, the basestation serves as an intermediate preprocessor. In the preprocessing step, the raw sensor readings are transformed into a meaningful statistical model of the complete sensor network. This model is continuously published to the query processing module. This module executes the posed queries with respect to the statistical model. For example, one possible

query is to determine all nodes whose reported temperatures are labeled as outliers with respect to the statistical model. The general architecture for the sensor stream processing is illustrated in Fig. 1. In the following, we will concentrate

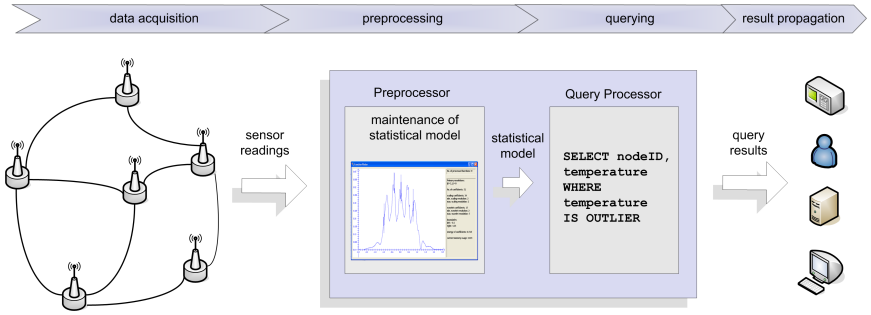

Fig. 1. processing architecture

on the computation of the statistical model within the preprocessor.

Point of origin is to model the set of attributes $X_1, ..., X_n$ as an n-dimensional random variable $X = (X_1, ..., X_n)$, i.e., we assume that the sensor readings are samples of the random variable X. Given the probability density function (pdf) $f(X_1, ..., X_n)$, we can determine the distribution of X in terms of the probabilities of arbitrary attribute constellations:

$$P(X_1 \in [a_1, b_1], ..., X_n \in [a_n, b_n]) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(X_1, ..., X_n) dx_1 ... dx_n. \quad (1)$$

The knowledge of the pdf is crucial as it gives a comprehensive summary of the process described by the random variable. Not only can we determine the above probabilities, but also determine meaningful characteristics like mean, variance, quantiles, correlations.

Let us give an example: For two sensors that measure temperature and air pressure respectively, we could conclude with their pdf that the probability of a temperature above 25 degrees celcius and an air pressure lower than 1000 hPa is extremely low. We also could determine the mean temperature or the correlation, i.e. the degree of linear dependency, between temperature and air pressure.

However, the question remains how we determine the pdf for a given set of measured attributes? In real-world scenarios, we must assume that we have no prior knowledge of the sensor stream distributions; we only have the raw sensor readings. One approach is to assume that the unknown pdf belongs to an a priori known class of densities, e.g. Gaussians. Given a Gaussian distribution, it remains to estimate its mean and variance with the help of the sensor readings. Due to its simplicity, this is a practically relevant approach, e.g. [7]. However, if the random variable does not follow the preset distribution, the resulting model is likely to be useless.

On account of this, we have concentrated in our work on so-called assumption-free density estimation techniques provided by mathematical statistics [9]. Those techniques are very appealing as they let the data speak strictly for themselves without any assumptions. As the computational complexity of these techniques prevents their direct application to data streams, we developed adaptations [4], [10] based on kernels and wavelets respectively. Both techniques continuously provide - with computational low cost - suitable density estimates that keep pace with current trends in the stream. With those density estimates as statistical model of the streams in a sensor network, we can gain insight into the physical phenomena observed by this network. In the following, we will illustrate this with their application to outlier detection.

3 Outlier Detection for Sensor Data

The detection of outliers in a timely fashion is an important task for virtually all applications on top of sensor networks. In facility monitoring, for example, an alert shall be triggered in case of exceptionally high temperatures. Another example is the food industry where perishable items shall be timely detected.

Generally, outlier detection has a long history in statistics and database research and there are many ways to define an outlier. Intuitively, we expect an outlier to be unusual or unexpected in comparison to a given data set; its occurrence is 'improbable'. Given the terminology introduced above, we label a point as an outlier if it lies in a region with low probability. This informal definition includes two important aspects: First, we consider the region around the outlier, quasi its neighborhood, and found the decision on the probability of the region. Second, we set a threshold for this probability. This threshold determines the sensitivity of the outlier classification. We incorporate both aspects in the following formal definition of an outlier:

Definition 1 *Let f be a pdf with support $[a, b]$ and $\epsilon, \delta \in (0, 1)$. A point x is an **outlier**, if $P(X \in [x - \frac{\epsilon(b-a)}{2}, x + \frac{\epsilon(b-a)}{2}]) \leq \delta$ holds.*

The parameter ϵ determines the width of the region while δ determines the rate of false positives and false negatives. The higher δ is set, the more 'normal' points will be labeled as outliers. The lower it is set, the more outliers will not be detected. The explicit setting of the parameters depends on the concrete application. For illustrative purposes, Fig. 2 presents an example for the definition of outliers.

With the above definition, we can develop an online algorithm to detect outliers in sensor streams. The chief part of the algorithm is the maintenance of a density estimate with the techniques mentioned above. Based on this estimate, we label a new sensor reading either as outlier or not. Except the reading is definitely not possible, e.g. negative velocities, we incorporate it into the density estimate. This ensures that we do not label values becoming more frequent always as outliers. The more often they appear, the more this will be reflected in the density estimate, i.e., their probability increases and consequently the probability of being labeled as outlier decreases.

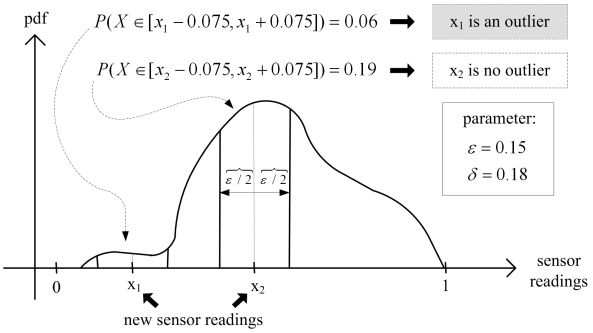

Fig. 2. detection of outliers

4 Conclusions

In this work, we investigated the augmentation of sensor network querying by meaningful statistical models. Instead of exploring the raw sensor readings, a statistical model offers a more reliable way to gain insight into the physical phenomena observed. A key ingredient of statistical models is the probability density function as it provides a comprehensive summary. Based on online computable estimates of the probability density function, we presented an initial approach to detect outliers in streaming sensor data. However, outlier detection is only one of many possibilities to enrich sensor querying with statistical modeling techniques. In fact, we expect this research direction to become highly relevant in future.

References

1. Tubaishat, M., Madria, S.: Sensor networks: an overview. IEEE Potentials **22(2)** (2003)
2. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and Issues in Data Stream Systems. In: PODS. (2002)
3. Jeffery, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J.: A Pipelined Framework for Online Cleaning of Sensor Data Streams. In: Proc. of ICDE. (2006)
4. Blohsfeld, B., Heinz, C., Seeger, B.: Maintaining Nonparametric Estimators over Data Streams. In: Proc. of BTW. (2005)
5. Liu, H., Hwang, S., Srivastava, J.: Probabilistic stream relational algebra: A data model for sensor data streams. Technical report, University of Minnesota (2004)
6. Cheng, R., Kalashnikov, D., Prabhakar, S.: Evaluating Probabilistic Queries over Imprecise Data. In: Proc. of ACM SIGMOD. (2003)
7. Elnahrawy, E., Nath, B.: Cleaning and Querying Noisy Sensors. In: Proc. of WSNA. (2003)
8. Deshpande, A., Guestrin, C., Hellerstein, J., Madden, S., Hong, W.: Model-Driven Data Acquisition in Sensor Networks. In: Proc. of VLDB. (2004)
9. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall (1986)
10. Heinz, C., Seeger, B.: Wavelet Density Estimators over Data Streams. In: Proc. of SAC. (2005)