Markov Chain-based Models for Missing and Faulty Data in MICA2 Sensor Motes

Farinaz Koushanfar
Electrical Engineering and
Computer Science Department
University of California, Berkeley
Berkeley, California 94720
Email: farinaz@eecs.berkeley.edu

Miodrag Potkonjak
Computer Science Department
University of California, Los Angeles
Los Angeles, California 90095
Email: miodrag@cs.ucla.edu

***Abstract*— We have developed Markov chain-based techniques for infield modeling the missing and faulty data for the widely used MICA2 sensor motes. These models help designers of sensor nodes and sensor networks to gain insights into the behavior of any particular sensor platform. The models also enable users of sensor networks to collect high integrity data from the deployed networks in a more efficient and reliable way. The new approach for development and validation of faults and missing data has two phases. In the first phase, we conduct exploratory analysis of data traces collected from the deployed sensor networks. In the second phase, we use the density estimation-based procedure to derive semi Markov models that best capture the patterns and statistics of missing and faulty data in the analyzed sensor data streams. We have applied the fault detection and missing data modeling procedure on light, temperature and humidity sensors on MICA2 motes in sensor networks deployed in office space and natural habitats. The technical highlight of the research presented in this paper include: (i) exploratory data analysis and studying the properties of the sensor data streams, (ii) adoption of a new class of semi Markov-chain models for capturing and predicting missing and faulty data in actual data trace streams.**

I. Introduction

The state-of-the-art sensor nodes and sensor networks are being built using inexpensive and ultra low power components, communication and operation protocols. A mote is a wireless node with computing and sensing capabilities and is the most commonly used node in sensor network applications [13]. Since sensor networks are often deployed unattended and in harsh environmental conditions, faulty measurements and missing data are unavoidable. For example, our analysis from sensor data traces collected at Intel Berkeley research lab [14] over a 3 week interval shows that almost 40% of the data was missing and about 8% of the data were faulty. Similar characteristics have been found in other collected measurements in sensor networks with MICA2 motes built by the Crossbow Technology [13]. Figure 1, shows a picture of a typical MICA2 sensor mote.

The key assumption for infield fault modeling for sensors is that in a sensor network settings where there are correlated measurements of a stimulus, the majority of sensors exposed to the stimulus are non-faulty. In general, if we can control the stimuli input to a sensing device, we can model and characterize its response and its fault models. However, during infield testing of sensors, we have to use fault models that do not assume the concept of controllable stimuli. Another difficulty in establishing the infield sensor fault models is the inherent uncertainty in sensor readings. Almost all of the sensor measurements are noisy. Therefore, temporary presence of small measurement errors is inevitable. We often have to accept a measurement with a relatively small error as a correct measurement.

Our approach to infield fault identification is data-driven and follows two principles: (i) predictability and (ii) consistency. We define predictability in the following way: consider a sensor network consisting of sensor set $S = \{s_1, s_2, s_3, ...s_N\}$. We say that sensor s_Y is predictable if one can calculate the values and measurements at sensor s_Y, using the measurements of one or more sensors from the set $S \backslash s_Y$. Consistency principle captures the assumption that at a given moment in time, the majority of sensors in a sensor network are observing the physical phenomena correctly (i.e. without faults). We also assume that deployment of the nodes in sensor networks is such that, under the assumptions of no faults, for each node we have at least one node from which we can predict its value accurately. Using predictability and consistency principles, we specify a faulty measurement as a measurement that is not predictable from any other sensors in the infield sensor network settings within a maximal target error bound of ϵ_t.

Fig. 1. Picture of a MICA2 mote. The MICA2 mote contains: (1) An Atmel ATmega128L low-power microcontroller, (2) A multi-Channel Radio Transceiver supporting 433, 868/916, or 310 MHz, and (3) A 51-pin expansion connector supports Analog Inputs, Digital I/O, I2C, SPI, and UART interfaces. The mote can have a variety of sensor types including light, temperature, and humidity sensors [13].

We have developed a family of Markov chain-based models for accurate capturing of time variability of the collected data. The Markov property is essentially a conditional independence of the future evolution on the past. Markov chains have the Markov property and consist of a set of time dependent random variables. The time dependent random variables are the states of the Markov chain and can assume values in a finite (or countably infinite) discrete set. The discrete set of the states is also referred to as the state space. More formally, we use a Markov process that is characterized as follows:

The state q_t at time t is one of a finite number of states in the range $\{s_1, ..., s_M\}$. Assuming that the process runs only from time 0 to time N and that the initial and final states are known, the state sequence could be presented by a finite vector $Q = (q_0, ..., q_N)$. If $P(q_t = s_i | q_0 = s_{j0}, q_1 = s_{j1}, ..., q_{t-1} = s_{j(t-1)})$ denotes the probability of the state q_t at time t conditioned on all states up to $t-1$. The process is called a first order Markov chain, since the probability of being in state q_t at time t given all the states up to the time $t-1$ depends only on the previous state q_{t-1}. The first order Markov process is more formally described in Equation 1.

$$P(q_t = s_i | q_0 = s_{j0}, q_1 = s_{j1}, ..., q_{t-1} = s_{j(t-1)}) \quad (1)$$
$$= P(q_t = s_i | q_{t-1} = s_{j(t-1)})$$

In the n-th order Markov process, the probability of being in state q_t at time t given all the states up to the time $t-1$ depends on the previous states up to the state q_{t-n}:

$$P(q_t = s_i | q_0 = s_{j0}, q_1 = s_{j1}, ..., q_{t-1} = s_{j(t-1)}) \quad (2)$$
$$= P(q_t = s_i | q_{t-n} = s_{j(t-n)}, ..., q_{t-1} = s_{j(t-1)})$$

We observe that even a simple Markov Chain models (e.g. a first order Markov chain) provides a significantly more accurate prediction of patterns for missing and faulty data than the current practice. It is possible to develop a Markov chain model that captures not only frequency, but also the lagged autocorrelation of the actual streams that contains correct, missing and faulty data. We have adopted a class of Markov chain models called *semi-Markov chain models* that ensure the correct lagged autocorrelation statistical properties, while keeping size of the models very compact. The accuracy of the models is improved by employing resubstitution and nonparametric smoothing for derivation of the density functions.

Applications of such semi Markov-chain models enabled us to develop better protocols for collecting data in the presence of faulty and missing samples. As an example, we show that by judicious use of redundant sampling and data retransmission, we are able to increase the percentage of collected data from less than 43% to more than 96% with only a 40% increase in power consumption.

The remainder of the paper is organized in the following way. First, we briefly survey the related literature in Markov-chain modeling and model-based fault and missing data recovery in sensor networks. Next, we present exploratory data analysis of correct, missing and faulty readings in sensor data streams. Our exploratory data analysis suggest that semi-Markov chain models are strong candidates for compact capturing of the relevant statistical properties of such streams. After that, we describe the details of the semi-Markov chain models. Due to the space limitations, we do not present the results for experimentation and evaluation of the semi-Markov chain modeling techniques. At last, we briefly state a number of future research directions and conclude the paper.

II. Related Work

Although Markov processes have been introduced about a century ago, their flexibility and wide range of applications is still deriving a variety of new extensions and applications [9], [5]. Bharucha-Reid provides a good overview of the discrete Markov processes [1].

The concepts of fault detection and identification are being widely studied in complex automatic control systems. Isermann and Balle [3] prepared a summary for a number of fault-detection and diagnosis methods and have shown strong trends towards applying the model-based fault-detection. Also, a number of discrete models have been proposed for characterizing the faults [8]. Examples of such models include fault trees [6] and structural graphs [12]. Lunze [7] presented the Markov properties of a state measurement sequence. Lunze has also presented a timed discrete-event abstraction for continuous-variable systems [8].

There are a number of conceptual, statistical, and optimization differences between the previous missing and faulty data modeling methods and the semi-Markov chain models presented in this paper. First, we use nonparametric data-driven statistical modeling methods to develop probability density functions utilized in semi-Markov models. Second, all of our developed density functions are validated using the techniques for statistical validation of random numbers with uniform distribution [4]. Finally, the consistent following of non-parametric data-driven paradigms during model development of validation facilitates incorporation of designer insights into the fault models in such a way to both increase the model accuracy as well as to make the models more amenable for consequent use in optimization procedures [4].

In this paper, we only describe modeling of the faulty and missing data. A related topic is replacing the missing data, based on the statistical patterns in the data streams. There are a variety of methods proposed for his task, including the expectation maximization (EM) algorithm [2] and the multiple imputation (MI) algorithm [10].

III. Exploratory Data Analysis

We performed several exploratory data analysis methods for the three states of interest in sensor data collection: (a) correct measurements, (b) missing measurements, and (c) faulty measurements. These states form the state-space for our Markov chain models. In Figure 2, the histograms show the number of consecutive measurements in the same state on a temperature sensor for the three states (a), (b) and (c). From these histograms, we learned that the number of consecutive

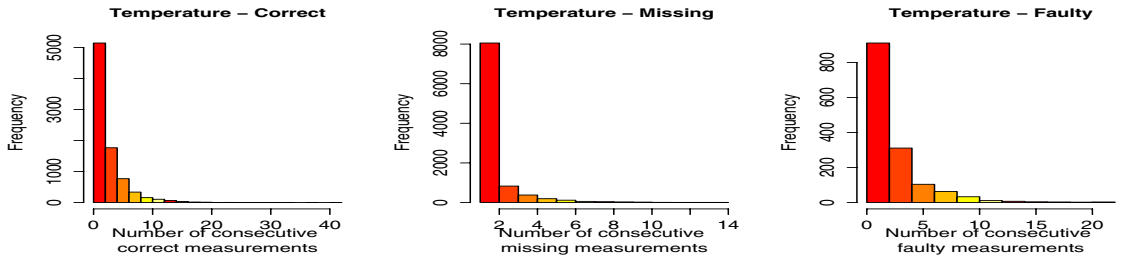


Fig. 2. Histograms for the number of consecutive correct measurements (left), the number of consecutive missing measurements (middle), and the number of consecutive faulty measurements (right) for the temperature sensors.

measurements in the same state is often more than one, and can easily go up to 10, especially for correct and faulty states. The missing state is more transient and often lasts less than 5 readings. In Figure 3, we take this analysis one step further by illustrating the factor level plot of the second states, vs. the probabilistic plot of the first states. The results are shown for light, temperature and humidity sensors on 3 different motes. We clearly see the low probability of having a correct measurement after a faulty one, showing a sustainable trend in faulty measurements. According to our experiments, the results are consistent for correct, missing, and faulty measurements across three different sensors and across different motes.

The accuracy of Markov chain models increases as we increase the number of previous data samples that are used to predict the probability of the current sample. However, as can be seen in Equation 2, adding more previous samples dramatically increases the size of our models. As we show on the example in Figure 2, often sequences of consecutive correct, or consecutive missing, or consecutive faulty recordings are rather long. We would like to have methods that capture autocorrelations between current data sample and lagged (previous) data samples without dramatic increase in the size of our model. In order to overcome the size limitations of the ordinary Markov chain models, we have developed a special class of semi Markov chain models. Semi-Markov models ensure that lagged autocorrelation statistical properties are properly captured without a significant overhead on size and complexity of the models.

IV. Special Semi-Markov Chain models

For the sake of clarity and due to space limitations, we will describe only the simplest and the most compact class of semi-Markov chain model for data streams with missing or faulty data. In this case, we consider only two states for the data: (1) correct, and (2) faulty. Note that, once the procedure for detection of faulty data is available, we can similarly treat missing and faulty data assuming that the missing points were faulty. The corresponding semi-Markov chain model will also have two states. The first state is indicated by zero and denotes the correct recordings. The second state is denoted by 1 and corresponds to missing or faulty data. As shown in Figure 4 (left diagram), the semi-Markov chain models for this case has only two deterministic transitions. The key property of the semi-Markov chain model is that it spends in each state as many time sampling periods as indicated by smoothed probability density functions.

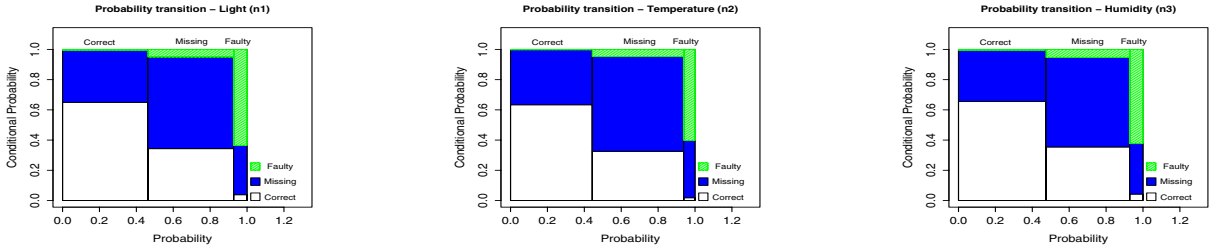
The procedure for derivation of the new class of semi-Markov chain model is very simple. We first build histograms that estimate the nonparametric probability density function (PDF) of the consecutive correct and consecutive faulty measurements. Examples of such histograms were shown in Figure 4. Consequently, we apply kernel smoothing techniques [11] to improve statistical robustness of the estimated density functions. Probability of transition between the two states now depends on the probability of staying in the same state. The probability of staying in the same state is already quantified by the probability density function.

Note that, if we want to have a model that independently captures correct, missing, and faulty data instead of using the PDFs, we use the conditional probability for the transition from one state to another. The accuracy of the semi-Markov chain models can be easily further improved by expanding the set of conditional probabilities over multiple sequences of consecutive readings where the different sequences have identical density functions and transition behaviors.

One of the most important issues in extracting and using semi-Markov models is to determine the length of times spent in each state. The measured samples are usually available at periodic time moments (in our case every 30 seconds). The sampling period does not necessarily have any correlations with the switching times between faulty and correct states. Until now, our presentation was assuming that the switching times are equivalent to the original sampling period. In order to resolve this limitation, we have developed the following procedure for calculating the switching time for the models.

First, we derive a semi-Markov model assuming that the model's switching time denoted as τ_{smc} is equal to the original sampling period denoted as τ_s. After that, we devise all of our available samples in two disjoint sets: odd and even sample sets. For each of the sets, we derive a semi-Markov model. The semi-Markov models are denoted as $smc_{2,o}$ and $smc_{2,e}$, for the odd and even sets respectively. The procedure of arranging the samples into sets can be further generalized to deduce semi-Markov models $smc_{p,q}$, where the subsampling is conducted with a rate p and we only consider samples that have an index K with the property that K is equal to $q(mod)p$.

Next, we study the properties of the subsampled models to deduce the best switching times for a given error range δ. As



(a) Light Sensor (a) Temperature Sensor (a) Humidity Sensor

Fig. 3. Factor level plots for the probability of different states for the second measurement, given the probability of the first state of the measurement shown on the x-axis. The plots show the light sensor (left), temperature sensor (middle), motes.

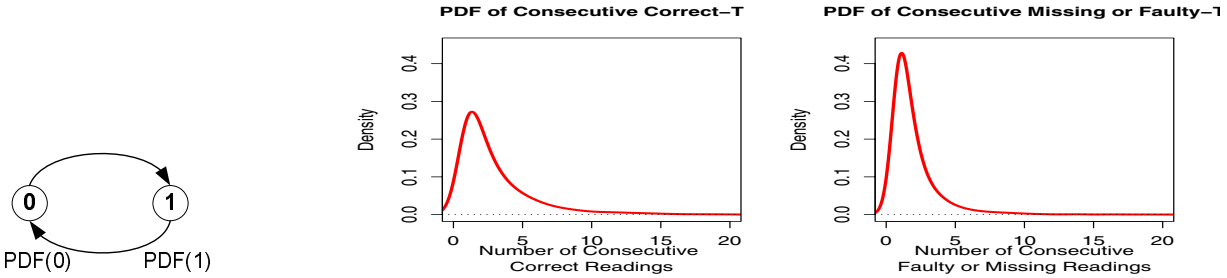


Fig. 4. Density transition between correct and incorrect (missing or faulty) measurements (left). The density function for the number of consecutive correct measurements (middle), and the number of consecutive incorrect measurements (right) for the temperature sensors.

we decrease number of samples, we obtain PDF's that have less precision than the original PDF derived using all samples. We accept the semi-Markov model with a PDF within the δ error range of the original PDF, such that it has the least number of samples. We also generalize the switching times to cases where the sampling rate is not periodic. For the sake of brevity and due to space limitations, the procedures are only presented in the journal version of this paper. Note that, the PDF's obtained in our measurements are such that one can assume any switching time and still use the obtained PDF's for the analysis and simulation purposes.

V. Limitations, Future Work and Conclusion

There are several limitations for the presented semi-Markov chain models for faulty and missing data that will be the target of our future work. There are two major limitations: First, there is a need to evaluate the procedure for developing sensor fault models for sensors deployed in different environments and to deduce which readings are faulty due to the limitations of the sensors, and which are faulty due to the impact of the instrumented environment. Second, we currently consider a single data stream at a single sensor for development of models for missing and faulty data. The usefulness of the developed models will improve by creating models that capture the correlations between faults at different sensors and different nodes distributed in the environment.

In summary, we have developed semi-Markov chain models for identification of faulty readings in sensor networks. The core of the paper is dedicated to a new class of semi-Markov chain models for capturing frequency and timing properties of false and missing data in real-life data streams. We have demonstrated that the models are statistically sound and can significantly contribute to the development of operational protocols for sensor networks.

References

[1] A.T. Bharucha-Reid. *Elements of the Theory of Markov Processes and Their Applications*. New York: McGraw-Hill, 1960.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likilihood from incomplete data via the em algorithm. *Royal Statistical Society Series*, pages 1–38, 1977.

[3] R. Isermann and P. Balle. Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice*, 5(5):709–719, 1997.

[4] F. Koushanfar. *Ensuring Data Integrity in Sensor-based Networked Systems*. Phd dissertation, Electrical Engineering and Computer Science Department, University of California, Berkeley, 2005.

[5] D. Kulp, D. Haussler, M.G. Reese, and F.H Eeckman. A generalized hidden markov model for the recognition of human genes in dna. In *Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 134–142, 1996.

[6] W. S. Lee, D. L. Grosh, F.A. Tillman, and C.H Lie. Fault tree analysis, methods, and applications - a review. *IEEE Transactions on Reliability*, R-34(3):194–203, 1985.

[7] J. Lunze. On the markov property of quantised state measuirement sequences. *Automatica*, 34(11):1439–1444, 1998.

[8] J. Lunze. A timed discrete-event abstraction of continuous-variable systems. *International Journal of Control*, 72(13):1147–1164, 1999.

[9] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257–286, 1989.

[10] D.B. Rubin. Multiple imputation after 18+ years. *Journal of the American Satisttical Association*, (91):473–489, 1996.

[11] B.W. Silverman and P.J. Green. *Density estimation for statistics and data analysis*. London: Chapman and Hall, 1986.

[12] M. Stroswiecki and M. Attouche, S. nd Assas. A graphic approach for reconfigurability analysis. In *Workshop on Principle of Diagnosis*, pages 250–256, 1999.

[13] *CrossBow Products:* http://www.xbow.com/products.

[14] *Intel Research at Berkeley:* http://www.intel-research.net/berkeley/.