

USAGE OF THE KALMAN FILTER FOR DATA CLEANING OF SENSOR DATA

Klemen Kenda, Jasna Škrbec, Maja Škrjanc
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773528; fax: +386 1 4773851
e-mail: klemen.kenda@ijs.si

ABSTRACT

This paper presents a methodology for data cleaning of sensor data using the Kalman filter. The Kalman filter is an on-line algorithm and as such is ideal for usage on the sensor data streams. The Kalman filter learns parameters of a user-specified underlying model which models the phenomena the sensor is measuring. Usage of the Kalman filter is proposed to predict the expected values of the measuring process in the near future and to detect the anomalies in the data stream. Furthermore the Kalman filter prediction can be used to replace missing or invalid values in the data stream. Algorithm only requires sensor measurements as an input, which makes it ideal to be placed as near to the resource tier in the N-tier architecture as possible.

1 INTRODUCTION

A sensor measurement today is cheap. One single sensor can produce thousands of measurements per day, per hour or even per minute. However, such measurements have often proven to be unreliable (not being delivered), a subject to environmental or system noise or even invalid, which quite often happens due to network failures or software errors along the process pipeline.

Due to the high volume of sensor measurements data cleaning should be performed in a computationally efficient way.

In this paper we propose the usage of the Kalman filter [2] to assist the data cleaning process. We propose to use prediction part of the Kalman filter for filling in the missing values in the sensor data stream. We propose to take advantage of prediction also with the detection of invalid sensor measurements. It is possible to perform a simple anomaly detection, based on comparison of new measurements and predictions and then according to the previously learned thresholds decide whether the new measurement is valid or not.

For the learning of thresholds we propose a semi-supervised method, where the user follows the evaluation of sensor measurements on a sample dataset and decides whether a measurement is an outlier or not.

Section 2 presents basics of the Kalman filter, relevant for understanding its role in the data cleaning process, which is described in Section 3. In Section 4 we propose the placement of data cleaning component within the N-tier architecture. We proceed with a presentation of prototype results on a sample sensor dataset where we also discuss a question and solution for the instability of the method. We finish with conclusion and ideas for the future work.

2 THE KALMAN FILTER

The Kalman filter is a method for solving the discrete-data linear problem. The filter consists of a set of mathematical equations that can estimate the underlying (hidden) state of a process in a way that the mean of the squared error is minimized. The filter supports estimation of past, present and even future states [5].

Underlying process to be modeled is a Gauss-Markov process (see Figure 1). This means that any subsequent state is only dependent on the previous state of the system. Figure depicts observations (in our case sensor measurements) x_j and underlying hidden states θ_j (vectors of a real value of the measured phenomena and its first and second temporal derivative). The arrows in the figure depict the Gaussian processes and point from underlying state θ_j to the next state θ_{j+1} (transitional equation) and also from the underlying state θ_j to the observable state of the system x_j (observational equation).

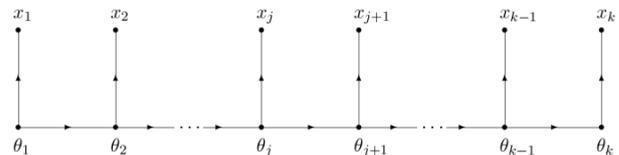


Figure 1: *Diagram of the Gauss-Markov process.*

The result of solving such a problem is a set of equations that include prediction and correction phase depicted in Figure 2. Comprehensive explanation and derivation of the filter equations can be found in the literature [3][4].

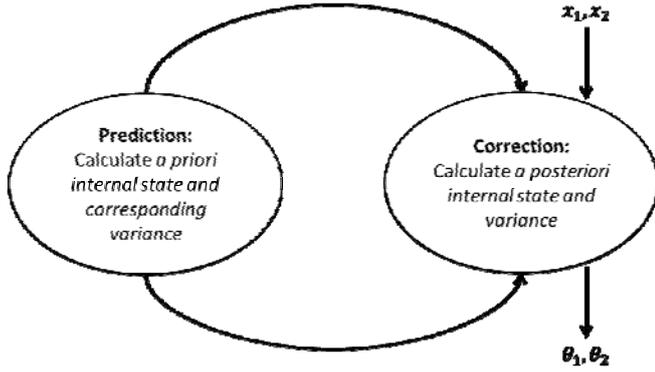


Figure 2: Kalman filter schema – prediction and correction phase.

The Kalman filter loop consists of two phases. Correction and prediction phase. The correction phase takes a measurement (x_j) and corrects the prediction (or the initial state) of the system. The correction phase also updates the underlying model parameters of the filter, which makes the filter adaptive. This means that the model is changing according to the measurements in an on-line manner. The prediction phase relies on the model-nature of the algorithm and is therefore able to project the current underlying state of the system into the future.

To illustrate the prediction phase of the Kalman filter only one equation from the whole set needs to be understood – the transitional equation below.

$$\theta_{k+1}^- = \Phi_k \theta_k$$

The new *a priori* state (prediction) θ_{k+1}^- is obtained by simply multiplying the transitional matrix Φ_k with the *a posteriori* state vector (previous correction). In our case, each row in the transitional matrix describes how each element of the state vector is transformed.

3 DATA CLEANING

Data cleaning is the first step to a proper preparation of input data stream for the research. The process of data cleaning detects and corrects data, which is corrupt, inaccurate, incorrect, incomplete, irrelevant, duplicated or missing. The data is corrected with a replacement or modification. The input of valid and correct data is very important for modeling, detection and prediction methods. Data cleaning includes data transformation, elimination of duplicated values, detection of missing data, statistical methods, error correction and detection of lost information. Data obtained from sensors can contain many errors, which may happen at the sensor level (the power supply is cut off, a sensors environment changes, mechanical failure, or a sensor measures imprecisely) or during the transfer of the measurement to the data acquisition system (a network problem, errors in gathering software or device, communication with a sensor is lost) [1].

In the process, decisions need to be made how to handle missing values, how big will be threshold and how data cleaning will be achieved with a live data stream. All these

decisions have to be based on a profound analysis and knowledge of the raw data.

3.1 Kalman Filter for Data Cleaning of Sensor Data

In our experience, the characteristics of the sensor data are as follows:

- streaming (on-line)
- high frequency (i.e. sensor readings are much more frequent than big changes of the property they are measuring)
- measured property is continuous and is changing smoothly (no big sudden jumps are expected with most of the properties; exceptions should be handled in the phase of semi-supervised initialization of the data cleaning filter)
- there are either only vague or too complex models for modeling the physical phenomena being measured

Considering the features above, we conclude that Kalman filtering is a suitable method for detecting outliers in sensor data. With its prediction features, one can also be able to replace the incorrect or missing data in the sensor data stream.

We propose a second-degree model (also suggested in [5]) that takes into account first and second order temporal derivatives of the measured properties. The equalities in Figure 3 define the state vector and the model to be used with a Kalman filter for data cleaning. A denotes the physical phenomena the sensor is measuring, and t the time. The state vector θ_k is 3-dimensional and includes the actual value of the physical phenomena in the first component, its first temporal derivative in the second component and its second temporal derivative in the third component.

$$\theta_k = (A, dA/dt, d^2A/dt^2)$$

$$\Phi_k = \begin{pmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{pmatrix}$$

Figure 3: The state vector and the transitional matrix in a dynamic linear model.

The concept of usage is straightforward. We exploit the prediction phase of the Kalman filter to predict the value of the measured phenomena for the timestamp of the measurement. Based on the comparison of the prediction with the actual value, the system decides whether the new measurement is correct or if it should be classified as an outlier. The difference between prediction and measurement is interpreted in terms of variance.

Figure 4 shows two examples for assessing whether a new measurement is an outlier or not. The principle can be generalized to any method using the prediction, not only Kalman filtering. In the first case prediction lies within the defined gap and in the second case the measurement lies outside the gap and is therefore discarded. The gap could be learned with a semi-supervised method, where the user would assist the algorithm by manually annotating the *good* measurements and outliers in the training set. The set

defines a hard border for the gap: every measurement that falls outside the gap is considered as an outlier to the algorithm.

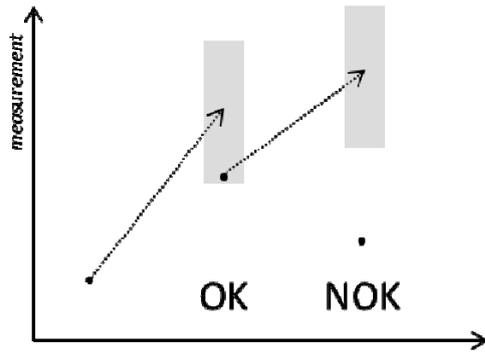


Figure 4: Evaluation of new measurements with Kalman filter prediction and learned threshold.

Stochastic invalid or missing values can be replaced with prediction values of the Kalman filter. These errors occur at random and usually represent single, isolated events. The second class of invalid/missing values represents those which are a consequence of a system failure (device or network), last longer and can not be compensated.

4 ARCHITECTURE

Many sensor network systems architecture is based on the N-Tier scheme. We propose to keep the data cleaning as close to the resource tier as possible. This enables all the upper layers to use cleaned data. Kalman filtering approach that we propose is relatively independent. It only needs sensor measurements to function properly and can be therefore implemented even at the sensor itself. However, we propose implementation just above the Data Access Tier (see Figure 5), which enables uniform access to all the data sources and optimizes the implementation.

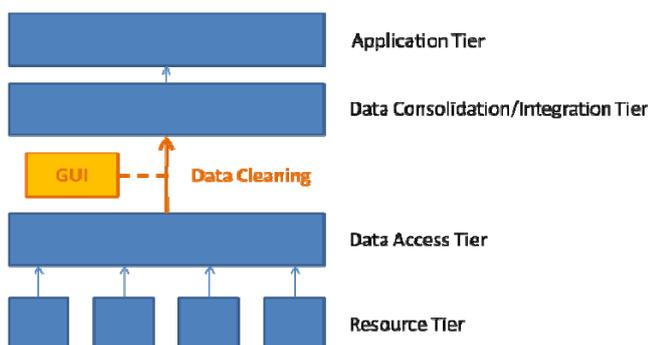


Figure 5: Position of data cleaning within the system architecture.

It needs to be pointed out that with such an approach we lose the possibility of data fusion aiding the data cleaning methods (e.g. using spatially or semantically correlated sensors).

The data cleaning process is semi-supervised, which in the context of on-line data streams means that the supervised part should be executed off-line. Supervised process includes data analysis and fine-tuning of parameters. Fine-tuning of parameters is done via a dedicated data cleaning GUI, where the expert user is able to supervise the process of data cleaning. After this process is finished, data cleaning can run automatically.

5 EVALUATION ON THE USE-CASE

The proposed methodology has been implemented in an early prototype for Data Cleaning in the NRG4Cast project. Experiments have been performed on the outside temperature dataset with one sensor reading per 15 minutes. The dataset included measurements from July to August 2013 with occasional stochastic failures in the form of 0.0°C readings.

Figure 6 shows basic principle of the algorithm. The Kalman filter in its prediction phase returns two relevant values: prediction for the value of the temperature and its variance. After semi-supervised stage, an expert user has determined proper upper and lower bound interval, which was 5σ (only observing the first dimension of state vector).

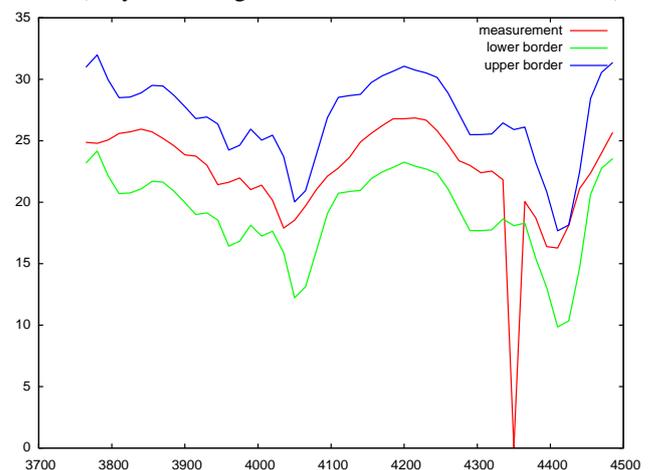


Figure 6: Identifying an outlier with the Kalman filter.

Figure shows the upper and lower boundary calculated from prediction and its variance and the actual measurements. Where a measurement lies outside the band delimited by the lower and upper boundary an outlier is detected.

With a good choice of model parameters, we have been able to achieve the results that included all the true positives and none of the false negatives. However, errors in the dataset have been specific and easily identifiable in the summer time (with high temperatures). The most difficult problem for the algorithm was a sudden change in temperature – as expected.

The Kalman filter is not a complex algorithm, but it can be difficult to adjust all the required initial conditions and parameters, which demand either an expert user or a statistical method to adjust them. With optimal tuning of the parameters very good results can be achieved.

6.1 Instability

The principle had proven to be unstable and the Kalman filter prediction had diverged in some cases. If the filter encounters a false negative it relies on the prediction model, which can then move the lower/upper boundary so that no measurement ever again fits the criteria (see Figure 7).

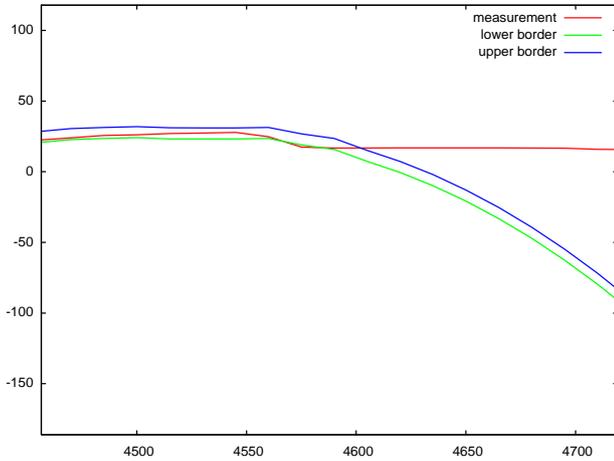


Figure 7: Instability of the algorithm when detecting a false negative.

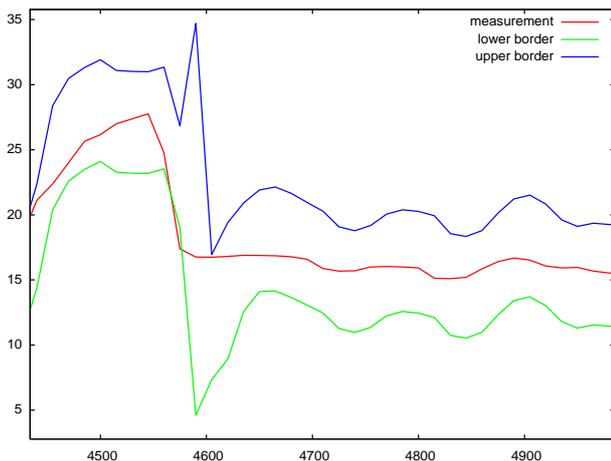


Figure 8: Instability workaround.

A possible workaround includes an artificial increase of *a posteriori* variance (see band enlargement after false negative in the Figure 8). Theoretically, changing variance is mathematically incorrect, but in practice it has proven to be efficient. Besides, variance in the case of slowly changing values of sensor data soon converges to the vicinity of the correct values. Further research on the proposed approach is needed.

6 CONCLUSION AND FUTURE WORK

We have proposed a methodology for data cleaning of the sensor data that requires no complex data processing. It relies solely on sensor measurements and initialization data. The use of this approach is possible already at the data source, but we propose to use it just above the Data Tier in

the N-Tier system architecture, which still enables upper layers (Integration and Application Tier) to work with clean data. There is however a trade-off between complexity and efficiency, which has not been examined.

We have tested some basic linear models with the Kalman filter equations and proposed to use a general second-degree model with sensor measurements. Kalman filtering results (its prediction and corresponding variance) have been proposed to be used for detection of outliers in the data stream. Method has been successfully tested on an environmental dataset.

As the experiments revealed an instability of the approach we have successfully addressed the issue with a modification of the algorithm.

Many interesting ideas have been identified for the future work. A very big problem when using the Kalman filter is initialization of the filter. One needs to optimize the behavior of the filter to a large number of parameters (approx. 20 for a 3-dimensional model). It would be interesting to investigate an optimization with gradient descent or other efficient methods (Levenberg-Marquardt), where the measure to minimize would be the χ^2 measure.

Another idea would include a more explicit definition of the instability workaround with increasing *a posteriori* variance. An algorithm should be used, which reduces number of initial parameters (initial variance and internal state) [6] and which would replace the classical Kalman filter.

Acknowledgements

This work was supported by the Slovenian Research Agency, by the Ministry of Education, Science and Sport within the Competence Center Open Communications Platform and the ICT Programme of the EC under PlanetData (ICT-NoE-257641), ENVISION (IST-2009-249120) and NRG4Cast (ICT-EeB- 600074).

References

- [1] S. Jeffery, G. Alonso, M. Franklin, W. Hong, and J. Widom. A pipelined framework for online cleaning of sensor data streams. In ICDE, page 140, 2006.
- [2] R. E. Kalman. A new approach to linear filtering and prediction problem. Journal of basic Engineering, 82(1):35-45, 1960.
- [3] R. G. Brown, P. Y. C. Hwang. Introduction to random signal and applied Kalman filtering. John Wiley, New York, 1996.
- [4] P. S. Maybeck. Stochastic models, estimation, and control, volume 141 of Mathematics in Science and Engineering. Academic Press, 1979.
- [5] N. N. Vijayakumar, B. Plale. Missing Event Prediction in Sensor Data Streams Using Kalman Filters, Knowledge Discovery from Sensor Data, 2009.
- [6] T. Podobnik, T. Živko. On Probabilistic Parametric Inference. Journal of Statistical Planning and Inference vol. 142 issue 12. December, 2012. p. 3152-3166.