Investigation of Optimal Alarm System Performance for Anomaly Detection

Rodney A. Martin, Ph.D.
NASA Ames Research Center
Intelligent Data Understanding Group
Mail Stop 269-1
Moffett Field, CA 94035-1000
rmartin@email.arc.nasa.gov

Abstract

Design and performance results are presented for a generic example as an application of optimal alarm systems, appealing to its interaction and reliance on data mining and machine learning techniques. By using an optimal alarm system, the fewest false alarms are elicited for a fixed detection probability of a specifically defined level-crossing event. The aim of this paper is to investigate the margin to optimality and subsequent performance when introducing approximations for the design of an alarm system. The optimal alarm system and its approximations use Kalman filtering for univariate linear dynamic systems driven by Gaussian noise, and provide a layer of predictive capability. Other level-crossing based alarm systems are introduced for comparison. These other methods also incorporate auxiliary fixed thresholds or redlines that provide a similar layer of predictive capability, but have no provision for minimizing false alarms.

1. Introduction & Background

This paper explores the development of a novel idea [12] for anomaly detection that is derived from the collusion of decades-old theory [3],[15] with more recent techniques [17],[18]. The idea stems from the design of optimal alarm systems which may enhance reliability and support health management for aerospace applications when monitoring control system error. When unexpected large transients in the control system error occur, this may be indicative of an impending fault or change in the system that may be cause for further diagnostic investigation. This error can be compared against a threshold whose selection is based upon the physics of the system and the margin of safety required. Even though the target application described above is specific to an aerospace platform, the technique is open to a broad range of other applications. These techniques include the prediction of high water levels [17], an application of thermal comfort as studied in [11], and potentially other environmental, science, or financial applications as appealed to in [1].

The idea of an "optimal alarm system" as an anomaly detection algorithm is derived from the prediction of level-crossing events whose optimality lies in the fact that the alarm system will elicit the least false alarms for a fixed detection probability and a given prediction window. The models currently under consideration are restricted to univariate linear time invariant systems driven by Gaussian noise, and hence the relationship to data mining lies in the fact that these models are generated by standard machine learning techniques. Thorough development of the type of model used here can be found in [9]. This model falls within the class of linear dynamic systems whose parameters are learned via the EM algorithm under certain constraints.

The practical applications of Kalman filtering for aerospace systems have largely been relegated to state estimation for guidance, navigation, and control purposes. Although the study of auxiliary failure detection and bad data rejection algorithms have been developed in concert with Kalman filters [8], [15], [19], the main purpose of those Kalman filters was for state estimation in guidance, navigation, and control systems. Kalman filtering has seen limited practical application dedicated to system reliability and health management as related to exceedance of predetermined failure thresholds in aerospace systems or more generally for anomaly detection.

Furthermore, most anomaly detection algorithms that have evolved from the data mining community use only a single threshold for decision or *design* purposes. Inherently, design of an anomaly detection algorithm involves adjusting the threshold in order to achieve an acceptable trade-off between true and false positives or a related performance metric. These thresholds are not traditionally based on physical limits, the physics of the system, or the margin of safety required. The thresholds which do character-

ize these predetermined limits are considered to be *failure*-based. As such, we must make a functional distinction between design-based and failure-based thresholds for anomaly detection based on data mining.

We propose the current optimal alarm system machinery therefore as a means to make the distinction between design-based and failure-based thresholds, in addition to providing a layer of predictive capability. This predictive capability is enabled by the fact that the design based-thresholds incorporate both a design parameter and a failure-based threshold. This allows for decoupling of the alarm system design using relevant performance metrics from the critical event itself, providing a measure of functional distinction.

In the case of using only a design-based single threshold, it is necessary to observe examples of failures in order to generate a metric such as the ROC curve empirically. Such a metric is used for alarm system design, and will sufficiently characterize the alarm system performance. Subject to certain constraints, design of the optimal alarm system can proceed without the need to observe actual examples of failures, and there is no need to estimate the alarm system metrics empirically. This obviates the need to rely upon having actual available examples of failures for alarm system design to generate the ROC curve. That is because they are based on the model and design parameters. However, the hypothesis-based level-crossing event must sufficiently characterize an actual physical failure for the model-based analysis to be of great benefit.

The novelty in the approach that we take with this investigation is that the Kalman filter machinery will be implemented for the express purpose of system reliability and health management, invoking more recently available data mining and machine learning techniques [7], [13], to develop suitable models. In addition, the Kalman filter machinery is more ubiquitously used for aerospace and other applications as distinct from ARMA models. These ARMA models were the original construct in which the practical use of optimal alarm systems was introduced [18]. Using Kalman filtering in tandem with optimal alarm theory will also invoke the predictive and functional strengths of applying both design and failure thresholds.

2. Methodology

2.1. Data-Driven Modeling

We will consider a standard linear dynamic system specified in discrete time by Eqns. 1-2. The state of the system, $\mathbf{x}_k \in \mathbb{R}^n$ evolves according to these equations, and often characterizes some internal physical characteristic of the system, beginning at time $k = 0$, with value $\mathbf{x}_0$ via state matrix $\mathbf{A}$. The scalar output of the system is given by $y_k \in \mathbb{R}$, and evolves through output matrix $\mathbf{C}$.

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{w}_k & (1)\\ y_k &= \mathbf{C}\mathbf{x}_k + v_k & (2) \end{aligned}$$

Both input noise ($\mathbf{w}_k$), which influences the state trajectory, and measurement noise, (v_k) which influences the output are introduced in order to allow for a more realistic model. The noise is modelled stochastically via a standard Gaussian distribution with means and covariances specified below.

$$\begin{aligned} \mathbf{w}_k &\sim \mathcal{N}(0, \mathbf{Q}) \\ v_k &\sim \mathcal{N}(0, R) \end{aligned}$$

Therefore the parameters to be learned are specified below, as the parameter θ.

$$\theta = (\mu_{\mathbf{x}}, \mathbf{P}_0, \mathbf{A}, \mathbf{C}, \mathbf{Q}, R) \quad (3)$$

where

$$\begin{aligned} \mu_{\mathbf{x}} &= E[\mathbf{x}_k] \\ \mathbf{P}_0 &= E[(\mathbf{x}_0 - \mu_{\mathbf{x}})(\mathbf{x}_0 - \mu_{\mathbf{x}})^T] \end{aligned}$$

$\mu_{\mathbf{x}}$ is the mean of the state trajectory, and $\mathbf{P}_0$ is the initial state covariance. These parameters are also shown in Fig. 1, which specify them in relation to the probabilistic graphical modeling paradigm to be used for machine learning purposes.

During the learning procedure for the linear dynamic system, the EM algorithm is used to find the parameters shown in Fig. 1. Details of this procedure are provided in Zoubin and Hinton [5] as well as Digalakis et al. [4], and it is implemented using Murphy's BNT (Bayes' Net Toolbox) [14]. Throughout learning, we attempt to retain the continuous-time analogue of Eqns. 1-2 in controllable canonical structure shown in Eqns. 4-8.

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}_c\mathbf{x}(t) + \mathbf{B}_w w(t) & (4)\\ y(t) &= \mathbf{C}_c\mathbf{x}(t) + v(t) & (5) \end{aligned}$$

where

$$\begin{aligned} w(t) &\sim \mathcal{N}(0, Q_c) \\ v(t) &\sim \mathcal{N}(0, R_c) \end{aligned}$$

$$\begin{aligned} \mathbf{A}_c &= \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\zeta\omega_n \end{bmatrix} & (6)\\ \mathbf{B}_w &= \begin{bmatrix} 0 \\ \omega_n^2 \end{bmatrix} & (7)\\ \mathbf{C}_c &= \begin{bmatrix} 1 & 0 \end{bmatrix} & (8) \end{aligned}$$

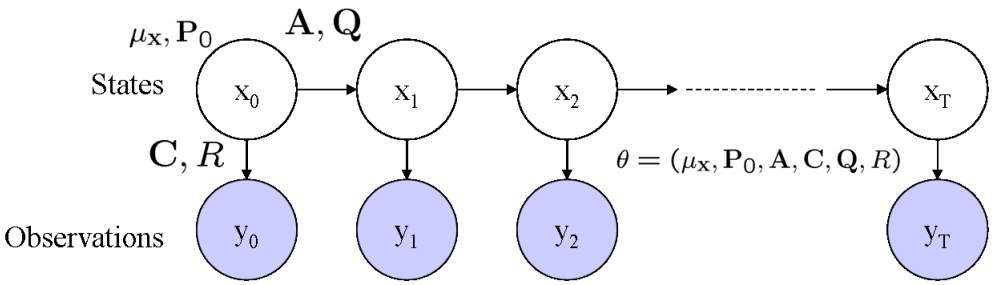

Figure 1. Linear Dynamic System

This is performed in order to allow for a mapping to intuitive canonical parameters: the natural frequency, ω_n, which is clamped during training, and the damping ratio, ζ, whose value is learned during training. Enforcing these constraints is easily performed by slight modification of appropriate open-source routines [14]. Doing so introduces sub-optimality into the learning procedure, which means that the learning curve will not necessarily increase monotonically. However, a reasonable sub-optimal local minimum will be found that best represents the parameter space with enforcement of the controllable canonical form constraint.

Clearly this is an extravagant simplification of the data generating process, however in doing so we allow for arbitrary system dynamics to be represented in an intuitive manner. Furthermore, at the very minimum an allowance for the introduction of serially correlated dynamics is introduced unlike other anomaly detection algorithms such as IMS [6] and Orca [2], [16]. The advantage of using machine learning lies in the fact that only the observations of a system, $\{y_k\}_{k=0}^{T}$ (transformed or raw) are required for determination of θ. In this way the technique is truly data-driven in nature.

2.2. Alarm Systems

The essence of the optimal alarm system is derived from the use of the likelihood ratio resulting in the conditional inequality: $P(C_k|y_0, \ldots, y_k) \geq P_b$. This basically says "give alarm when the conditional probability of the event, C_k, exceeds the level P_b." Here, P_b represents some optimally chosen border or threshold probability with respect to a relevant alarm system metric. It is necessary to find the alarm regions in order to design the alarm system. The event, C_k, can be chosen arbitrarily, and is usually defined with respect to a pre-specified critical threshold, L, as well as a prediction window, d. In this paper, the event of interest is shown in Eqn. 9, and represents at least one exceedance outside of the threshold envelope specified by $[-L, L]$ of the process y_k within the specified look-ahead prediction window, d.

$$C_k \triangleq \{|y_k| > L\} \bigcup \left[\bigcup_{j=1}^{d} \left[\bigcap_{i=0}^{j-1} |y_{k+i}| < L, |y_{k+j}| > L \right] \right] \tag{9}$$

There are three different alarm systems to compare which will all attempt to predict the level-crossing event defined by Eqn. 9, whose probability, $P(C_k)$, can be computed according to formulae presented in [12]. The first alarm system attempts to define an envelope, $[-L_A, L_A]$, outside of which an alarm will activate. In order to provide for a layer of predictive capability, L_A should be chosen such that $L_A < L$. An alarm probability can likewise be computed, $P(A_k) = P(|y_k| > L_A)$ and the details of this formula are also provided in [12]. This "redline" alarm system is termed as such in order to give credence to the fact that a simple level is used, and often the same terminology is used in practice. Even without the benefit of using any predicted future process values, this alarm system would be superior to a true redline system that uses only a single level L. However, in this case two levels are used, L as the failure threshold, and L_A as the design threshold.

The second alarm system incorporates the use of predicted future process values, and is called the "predictive" alarm system. This alarm system also defines an envelope, $[-L_A, L_A]$, outside of which an alarm will sound. Similarly, L_A should be chosen such that $L_A < L$ in order to provide for a layer of predictive capability. However, the alarm probability is defined in a different fashion than the for the redline method, as $P(A_k) = P(|\hat{y}_{k+d|k}| > L_A)$, where the predicted future process value $\hat{y}_{k+d|k}$ is found from standard Kalman filter equations. The final alarm system to be compared to the previous two is the optimal alarm system, and has two approximations, but only the one presented as Eqn. 10 will be used for comparison in this paper. The alarm condition, $P(C_k|y_0, \ldots, y_k) \geq P_b$, can be approximated to form the alarm region specified in Eqn. 10.

$$A_k = \bigcup_{i=0}^{d} |\hat{y}_{k+i|k}| \geq L + \sqrt{V_{k+i|k}}\Phi^{-1}(P_b) \qquad (10)$$

where $\Phi^{-1}(\cdot)$ represents the inverse cumulative normal standard distribution function, and $V_{k+i|k} = \text{Var}(y_{k+i}|y_0, \ldots, y_k)$.

Eqn. 10 plays a pivotal role in enabling the enforcement of the approximation to the alarm region for an optimal alarm system. Using this approximation allows it to outperform the other alarm systems with respect to the minimization of false alarms. All of the three alarm systems described will be compared using the ROC curve. This provides a performance metric with which to assess and compare the performance of each alarm system. The ROC curve parametrically displays the true positive rate against the false positive rate. The parameters of interest are L_A for the redline and predictive methods, and P_b for the approximation to the optimal alarm system. It is possible to generate formulae for the true and false positive rates as a function of these parameters (L_A, P_b) as well as the model parameters (θ) by appealing to Eqns. 11-12. These details for constructing these formulae are provided in [12].

True positive rate:

$$P(C_k|A_k) = \frac{P(C_k, A_k)}{P(A_k)} \qquad (11)$$

False positive rate:

$$P(A_k|C_k^{'}) = \frac{P(C_k^{'}, A_k)}{P(C_k^{'})} \qquad (12)$$

3. Results

The example to be used for the presentation of our results has no specific application, but is generic, and the model parameters are provided in Eqns. 13-16.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.9 & 1.8 \end{bmatrix} \qquad (13)$$

$$\mathbf{C} = \begin{bmatrix} 0.5 & 1 \end{bmatrix} \qquad (14)$$

$$Q \triangleq \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad (15)$$

$$R \triangleq 0.08 \qquad (16)$$

For all cases, the threshold $L = 16$, and the prediction window, $d = 5$. The resulting ROC curve is shown in Fig. 2 for comparison, and qualitative realizations based upon selecting the optimal design point is shown in Fig. 3. The optimal design points for each method have been selected based upon the same (minimax) criterion, indicated on the ROC curve.

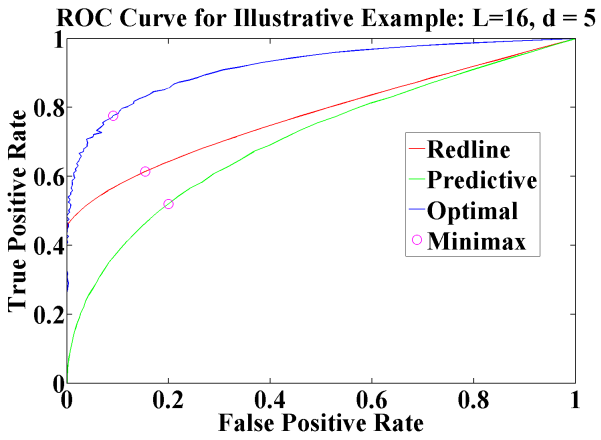

Figure 2. ROC Curve Comparison

4. Conclusion

Clearly, the optimal alarm system outperforms both the redline and predictive methods. This can be ascertained either from the ROC curve in Fig. 2, or by the realizations provided in Fig. 3. In Fig. 2, the optimal alarm approximation is "above" both the predictive and redline curves. This indicates that the true positive rate of the optimal alarm system is higher for any given false positive rate of the redline or predictive alarm systems. In Fig. 3, the "minimax" design criterion is used to select the design point from each respective ROC curve. Recall that for the each alarm system, the event of interest is to predict at least one exceedance out of the envelope $[-L, L]$ in the next d steps.

The critical levels that comprise the envelope in addition to the monitored process are displayed along with the alarm thresholds, false alarms, missed detections, correct detections, and predicted future process values (where applicable) for each alarm system. For easy comparison, the realization is identical for each alarm system. It is clear that on the bottom graph representing the optimal alarm system there are much fewer false alarms (identified by the red crosses), and many more correct detections (identified by black circles). For the other two graphs on top of Fig. 3, many more false alarms appear due to predictions that are sub-optimally based on a fixed L_A rather than time-varying optimal thresholds based on P_b shown on the bottom.

Therefore, using either method of presenting the results, Fig. 2, or Fig. 3, it is apparent that using the best approximation to the optimal alarm system will always outperform the predictive method. In this case, the predictive method actually performs worse than the redline method. In some cases this may be true of the approximations to the optimal alarm system, but this is dependent on the dynamics of the system, and the fidelity of the approximation, both of which may be investigated in future studies.

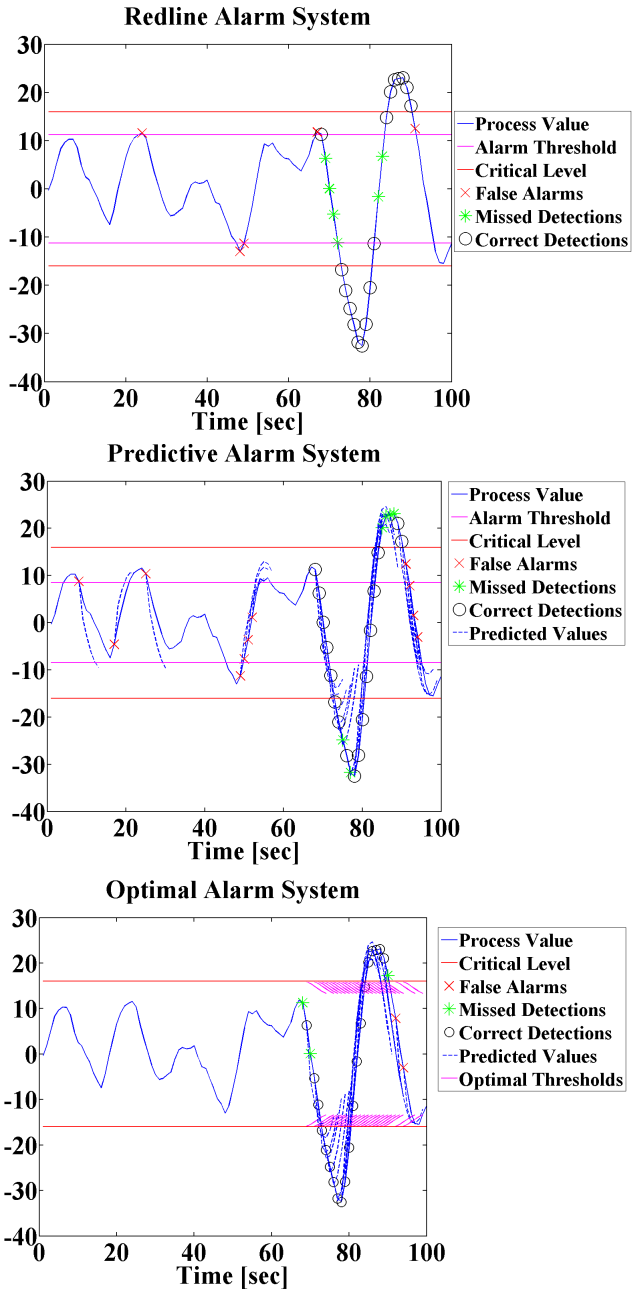

Figure 3. Qualitative Comparison of Realizations

References

[1] M. Antunes, A. A. Turkman, and K. F. Turkman. A Bayesian approach to event prediction. *Journal of Time Series Analysis*, 24(6):631–646, November 2003.

[2] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *KDD '03: Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38, New York, NY, 2003. ACM Press.

[3] H. Cramér and M. Leadbetter. *Stationary and Related Stochastic Processes*. John Wiley and Sons, 1967.

[4] V. V. Digalakis, J. R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):431–442, 1993.

[5] Z. Ghahramani and G. E. Hinton. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto, 1996.

[6] D. L. Iverson. Inductive system health monitoring. In *Proceedings of The 2004 International Conference on Artificial Intelligence (IC-AI04)*, Las Vegas, Nevada, June 2004. CSREA Press.

[7] M. I. Jordan. An introduction to probabilistic graphical models. Manuscript used for Class Notes of CS281A at UC Berkeley, Fall 2002.

[8] T. H. Kerr. False alarm and correct detection probabilities over a time interval for restricted classes of failure detection algorithms. *IEEE Transactions on Information Theory*, IT-28(4):619–631, July 1982.

[9] R. Martin. Unsupervised anomaly detection and diagnosis for liquid rocket engine propulsion. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, March 2007.

[10] R. Martin, M. Schwabacher, N. Oza, and A. Srivastava. Comparison of unsupervised anomaly detection methods for systems health management using space shuttle main engine data. In *Proceedings of the 54th Joint Army-Navy-NASA-Air Force Propulsion Meeting*, Denver, CO, May 2007.

[11] R. A. Martin. *Optimal Prediction, Alarm, and Control in Buildings Using Thermal Sensation Complaints*. PhD thesis, University of California, Berkeley, 2004.

[12] R. A. Martin. Approximations of optimal alarm systems for anomaly detection. *IEEE Transactions on Information Theory (preprint)*, 2007.

[13] K. Murphy. Switching Kalman Filters. Technical report, Department of Computer Science, University of California, Berkeley, 1998.

[14] K. P. Murphy. The Bayes' Net Toolbox for MATLAB. *Computing Science and Statistics*, 33, 2001.

[15] S. F. Schmidt. The Kalman Filter: Its recognition and development for aerospace applications. *Journal of Guidance, Control, and Dynamics*, 4(1):4–7, 1981.

[16] M. Schwabacher. Machine learning for rocket propulsion health monitoring. In *Proceedings of the SAE World Aerospace Congress*, volume 114-1, pages 1192–1197, Dallas, Texas, 2005. Society of Automotive Engineers.

[17] A. Svensson. *Event Prediction and Bootstrap in Time Series*. PhD thesis, Lund Institute of Technology, September 1998.

[18] A. Svensson, J. Holst, R. Lindquist, and G. Lindgren. Optimal prediction of catastrophes in autoregressive moving-average processes. *Journal of Time Series Analysis*, 17(5):511–531, 1996.

[19] A. S. Willsky and H. L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21(1):108–112, 1976.