Real-time Bayesian Anomaly Detection for Environmental Sensor Data

David J. Hill
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign
Phone: (217) 333-1657, Fax: (217) 333-6968, E-Mail: djhill1@uiuc.edu

Barbara S. Minsker
Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign
Phone: (217) 333-9017, Fax: (217) 333-6968, E-Mail: minsker@uiuc.edu

Eyal Amir
Department of Computer Science, University of Illinois at Urbana-Champaign
Phone: (217) 333-8756, Fax: (217) 265-6591, E-Mail: eyal@cs.uiuc.edu

Recent advances in sensor technology are facilitating the deployment of sensors into the environment that can produce measurements at high spatial and/or temporal resolutions. Not only can these data be used to better characterize systems for improved modeling, but they can also be used to produce better understandings of the mechanisms of environmental processes. One such use of these data is anomaly detection to identify data that deviate from historical patterns. These anomalous data can be caused by sensor or data transmission errors or by infrequent system behaviors that are often of interest to the scientific or public safety communities. Thus, anomaly detection has many practical applications, such as data quality assurance and control (QA/QC), where anomalous data are treated as data errors; focused data collection, where anomalous data indicate segments of data that are of interest to researchers; and event detection, where anomalous data signal system behaviors that could result in a natural disaster. This study develops two automated anomaly detection methods that employ Dynamic Bayesian Networks (DBNs). These machine learning methods can operate on a single sensor data stream, or they can consider several data streams at once, using all of the streams concurrently to perform coupled anomaly detection. This study investigates these methods' abilities, using both coupled and uncoupled detection, to perform QA/QC on two windspeed data streams from Corpus Christi, Texas; false positive and false negative rates serve as the basis for comparison of the methods. The results indicate that a coupled DBN anomaly detector, tracking the actual windspeeds, their measurements, and the status of these measurements, performs well at identifying erroneous data in these data streams.

Keywords: coastal environment; Bayesian networks; anomaly detection; sensor networks; machine learning

Introduction

In-situ environmental sensors (sometimes called "embedded" sensors) are sensors that are physically located in the environment they are monitoring. They collect time series data that flow continuously to a repository, creating a data stream. Recently, there have been efforts to make use of streaming data for real-time applications. For example, draft plans for the Water and Environmental Research Systems (WATERS) Network, a proposed U.S. environmental observatory network, have identified real-time analysis and modeling as significant priorities (NRC 2006). The value of streaming data for real-time forecasting and decision making has been demonstrated using a simulated oil spill (Bonner *et al.*, 2002), and continuing efforts are being directed towards facilitating near-real-time hydrodynamic forecasting using these data (Shah *et al.*, 2005).

Because *in-situ* sensors operate under harsh conditions, and because the data they collect must be transmitted across communication networks, the data can easily become corrupted. Undetected errors can significantly affect the data's value for real-time applications. Thus, the National Science Foundation (2006) has indicated a need for automated data quality assurance and control (QA/QC). This can be accomplished via anomaly detection, which is the process of identifying data that deviate markedly from historical patterns (Hodge & Austin 2004). Anomalous data can be caused by sensor or data transmission errors or by infrequent system behaviors that are often of interest to scientific and regulatory communities. In addition to data QA/QC, where anomalous data are treated as erroneous, anomaly detection has many other practical applications, such as adaptive monitoring, where anomalous data indicate phenomena that researchers may wish to investigate further through increased sampling; and anomalous event detection, where anomalous data signal system behaviors that could result in a natural disaster. These applications require real-time detection of anomalous data, so the anomaly detection method must be rapid and must be performed incrementally, to ensure that detection keeps up with the rate of data collection.

Traditionally, anomaly detection has been carried out manually with the assistance of data visualization tools (Mourad & Bertrand-Krajewski 2002), but these approaches are too time consuming for real-time detection in streaming data. More recently, researchers have suggested automated statistical and machine learning approaches, such as minimum volume ellipsoid (Rousseeuw & Leroy 1996), convex pealing (Rousseeuw & Leroy 1996), nearest neighbor (Tang 2002; Ramaswamy *et al.* 2000), clustering (Bolton & Hand 2001), neural network classifier (Kozma *et al.* 1994), support vector machine classifier (Bulut *et al.* 2005), and decision tree (John 1995). These methods are faster than manual methods, but they have drawbacks that make them unsuitable for real-time anomaly detection in streaming data; for example, some require that all the data to have accumulated before anomalies can be identified; some are computationally intractable for large quantities of data; some require pre-classified anomalous data, which characterize all anomalies that may be encountered; and some require pre-classified non-anomalous data, which characterize the range of possible non-anomalous data.

Several researchers have suggested anomaly detection methods specifically designed for real-time detection in streaming data. These methods are often referred to as analytical redundancy methods because they employ a model of the sensor data stream as a simulated redundant sensor whose measurements can be compared with those of the actual sensor. The classification of a measurement as anomalous is based on the difference between the model prediction and the sensor measurement. Hill and Minsker (2006) present an analytical redundancy method for detecting anomalies in environmental sensor data and compare its performance using several data-driven modeling approaches, including nearest neighbor, clustering, perceptron, and artificial neural networks. This method, however, is limited, because it cannot consider several data streams at once and because missing values in the data stream render it incapable of classifying measurements that immediately follow the missing values.

To address these limitations, this study develops two real-time anomaly detection methods that employ dynamic Bayesian networks (DBNs) to identify anomalies in environmental streaming data. DBNs are artificial intelligence techniques that model the evolution of discrete and/or continuous valued states of a dynamic system by tracking changes in the system states over time. The following section describes these methods in detail. Next, nine instantiations of two DBN-based methods are tested through a case study in which they are used to identify

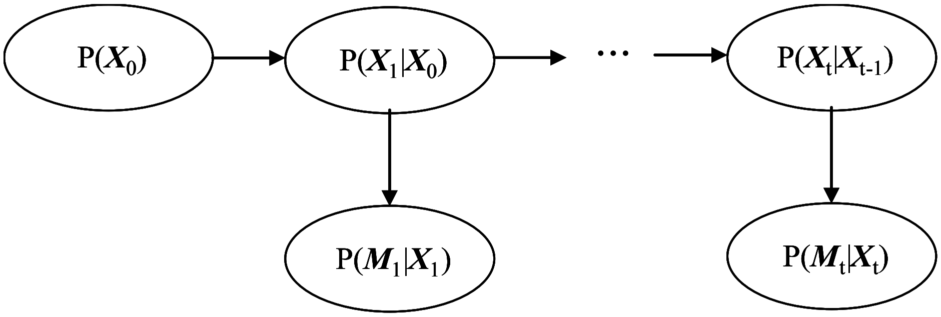
P($\boldsymbol{X}_0$)
P($\boldsymbol{X}_1$|$\boldsymbol{X}_0$)
…
P($\boldsymbol{X}_t$|$\boldsymbol{X}_{t-1}$)
P($\boldsymbol{M}_1$|$\boldsymbol{X}_1$)
P($\boldsymbol{M}_t$|$\boldsymbol{X}_t$)

Figure 1: Graphical structure of DBN-1. Vector $\boldsymbol{X}$ represents the continuous valued, hidden system variables and vector $\boldsymbol{M}$ represents the continuous valued, observed system variables. Subscripts indicate time.

erroneous measurements in two windspeed data streams from the WATERS Network Corpus Christi Bay testbed, provided by the Shoreline Environmental Research Facility (SERF). Finally, implications of these modeling methods are discussed.

METHODS

This study investigates the use of dynamic Bayesian networks (DBNs) for detecting anomalies in environmental sensor data streams. Bayesian networks are directed, acyclic graphs, in which each node contains probabilistic information regarding all the possible values of a state variable (Russell & Norvig 2003). This information, combined with the network topology, specifies the full joint distribution of the state variables, which, given a set of known variable values, can be used to infer the most likely value of the unknown variables. Dynamic Bayesian networks are Bayesian networks with network topology that evolves over time, adding new state variables to represent the system state at the current time *t*. State variables can be categorized as either unknown (hidden) state variables that represent the true states of the system or measured (observed) state variables that represent imperfect measurements of one or more of the true states; state variables can be either discrete or continuous valued.

Because the network size increases over time, performing inference using the entire network would be intractable for all but trivial time durations. Luckily, efficient recursive algorithms have been developed that perform exact inference on specific types of DBNs (Maybeck 1979) or approximate inference on more general types of DBNs (Doucet *et al.* 2000a). Two of these algorithms are Kalman filtering and Rao-Blackwellized particle filtering. Both algorithms perform filtering, or inference of the current hidden system states given all of the observed states to date. Kalman filtering employs the assumption that all state variables are linear Gaussian random processes to perform exact inference, while Rao-Blackwellized particle filtering uses a sample of the state distributions (the particles) to perform approximate inference and thus does not limit the type of state variables (Doucet *et al.* 2000b).

Two strategies for detecting anomalous data were considered in this study: Bayesian credible interval (BCI) and maximum *a posteriori* measurement status (MAP-ms). The BCI method uses the simple DBN shown in Figure 1, hereafter referred to as DBN-1, which tracks the multivariate distributions of linear Gaussian state variables corresponding to the hidden system states and their observed counterparts that are measured by the environmental sensors. The hidden states are assumed to be first-order Markov processes, so the state at time *t* only depends

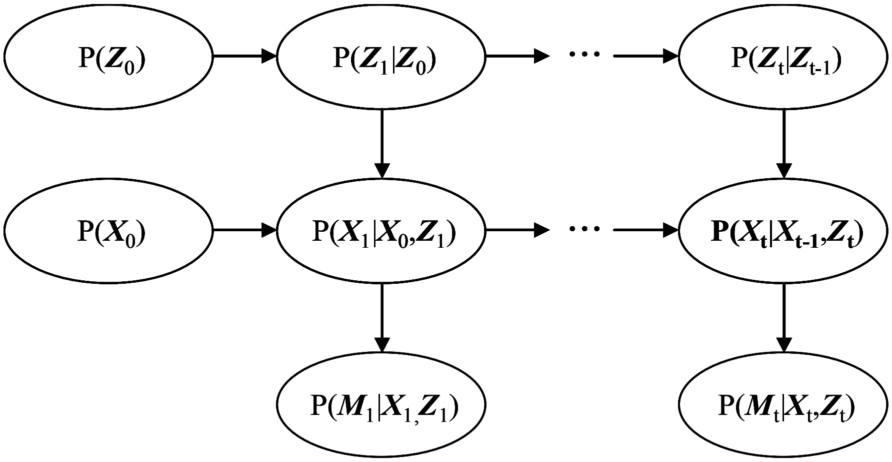
$P(\boldsymbol{Z}_0)$
$P(\boldsymbol{Z}_1|\boldsymbol{Z}_0)$
…
$P(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1})$
$P(\boldsymbol{X}_0)$
$P(\boldsymbol{X}_1|\boldsymbol{X}_0,\boldsymbol{Z}_1)$
…
$P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t)$
$P(\boldsymbol{M}_1|\boldsymbol{X}_1,\boldsymbol{Z}_1)$
$P(\boldsymbol{M}_t|\boldsymbol{X}_t,\boldsymbol{Z}_t)$

Figure 2: Graphical structure of DBN-2. Vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent the continuous valued and discrete valued hidden system variables, respectively, and vector $\boldsymbol{M}$ represents the continuously valued observed system variables. Subscripts indicate time.

on the state at time t-1. Kalman filtering is used to sequentially infer the posterior distributions of hidden and observed states as new measurements become available from the sensors. The posterior distribution of the observed state variables can then be used to construct a Bayesian credible interval for the most recent set of measurements. The p% credible interval indicates that the posterior probability of the observed state variables falling within the interval is p; thus, the Bayesian credible interval delineates the range of plausible values for sensor measurements. For this reason, any measurements that fall outside of the p% Bayesian credible interval can be classified as anomalous. The network parameters (i.e. the probability distributions $P(\boldsymbol{X}_0)$, $P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})$, $P(\boldsymbol{M}_t|\boldsymbol{X}_t)$) for DBN-1 were learned from sensor data using the expectation-maximization algorithm (Digalakis *et al.* 1993).

The MAP-ms method uses the more complex DBN shown in Figure 2, hereafter referred to as DBN-2, which tracks the multivariate distributions of linear Gaussian state variables corresponding to hidden system states and their observed counterparts, which are measured by the environmental sensors, as well as the distribution of a discrete hidden state variable which indicates the status (e.g. normal/anomalous) of each sensor measurement. For example, if there are two measured states, then the measurement status variable will have four values: (normal, normal), (anomalous, normal), (normal, anomalous), and (anomalous, anomalous). Rao-Blackwellized particle filtering is used to sequentially infer the posterior distributions of the hidden and observed states as new measurements become available from the sensors. The maximum *a posteriori* estimate, (e.g. the most likely value given the posterior distribution) of the hidden state variable indicating the measurement status can then be used to classify the sensor measurements as normal or anomalous. DBN-2 requires (1) network parameters describing the time-evolution of the linear Gaussian states conditioned on each value of the discrete state and (2) parameters describing the time-evolution of the discrete state. For the case in which all sensor measurements were normal, the parameters of the linear Gaussian states were specified to be the same as those learned for DBN-1. For the cases in which one or more measurements was anomalous, the parameter specifying the measurement variance of the anomalous measurement was set to be a large number (e.g. 10,000), indicating that regardless of the true state of the system, the measurement could take any real value with approximately equal probability. This description of anomalous measurements was used because it indicates that an anomalous measurement is more likely to fall outside the range of plausible measurements than a non-anomalous measurement yet it does not require *a priori* knowledge of the types of anomalies that can occur. The discrete state distributions (i.e. $P(\boldsymbol{Z}_0)$, $P(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1})$) were set manually, using domain knowledge/intuition. Manually setting the parameters for the cases in which one or more measurements was anomalous was necessary because anomalous measurements are, by definition, infrequent; as such, insufficient information is available for learning these parameters from the data. Furthermore, learned parameters may define anomalies too narrowly to identify the range of anomalies that may be encountered.

CASE STUDY

To demonstrate and compare the efficacy of the anomaly detection methods developed in this study for data QA/QC, they were applied to two SERF windspeed sensor data streams (CC003 and CC009) from different locations within Corpus Christi Bay. The sensors are R. M. Young model 05106 marine wind monitors, which collect windspeed and direction at a frequency of 1/120 hertz (i.e. one measurement every two minutes). The BCI and MAP-ms strategies were tested using each data stream individually to perform uncoupled anomaly detection and using both data streams concurrently to perform coupled anomaly detection. Thus, there are two instances each of DBN-1 and DBN-2 for uncoupled anomaly detection (one for CC003 and one for CC009) and one instance each of DBN-1 and DBN-2 for coupled anomaly detection. Two BCI's were compared for the instances of DBN-1: 95% and 99%. These nine combinations will hereafter be referred to as anomaly detectors.

Parameters for the DBN-1 detectors were learned from the approximately 22,300 windspeed measurements collected during the month of October 2006 by each of the two sensors, and parameters for DBN-2 detectors were selected based on the appropriate DBN-1 parameters as described above. Data from October was used for learning rather than data randomly sampled from several months, because the EM algorithm requires contiguous measurements for learning. Furthermore, 1000 particles were used for Rao-Blackwellized particle filtering on the DBN-2 based detectors. Testing of the anomaly detectors was performed using the over 21,600 measurements collected during November 2006. Learning the model parameters required approximately 60 seconds regardless of whether the data streams were uncoupled or coupled. Classification of a new measurement by the resulting DBN-1 based detectors required 0.0028 seconds, while the resulting DBN-2 based detectors required 0.33 seconds.

The testing data indicate that between consecutive measurements, the largest windspeed decrease, average windspeed change, and largest windspeed increase for CC003 are -7.08m/s, 0m/s, and 4.51 m/s, respectively. These statistics for CC009 are -10.13m/s, 0m/s, and 2.76 m/s, respectively. Since no erroneous measurements were known to exist within the November windspeed data, synthetic errors were injected into the data stream. These errors were modeled as transient (i.e. short duration) faults that affected either sensor independently (i.e. both sensors could fail at the same time) with equal probability (6%) and increased or decreased the actual measurement by an offset selected uniformly from a range of positive numbers. To investigate the effect of the offset range, two ranges were compared in this study: R1, [1.03 – 11.3]m/s, and R2, [2.57 – 12.9]m/s. These ranges were selected arbitrarily, such that when subtracted from (or added to) the actual measurement, the minimum values of R1 and R2 fall roughly into the 2nd

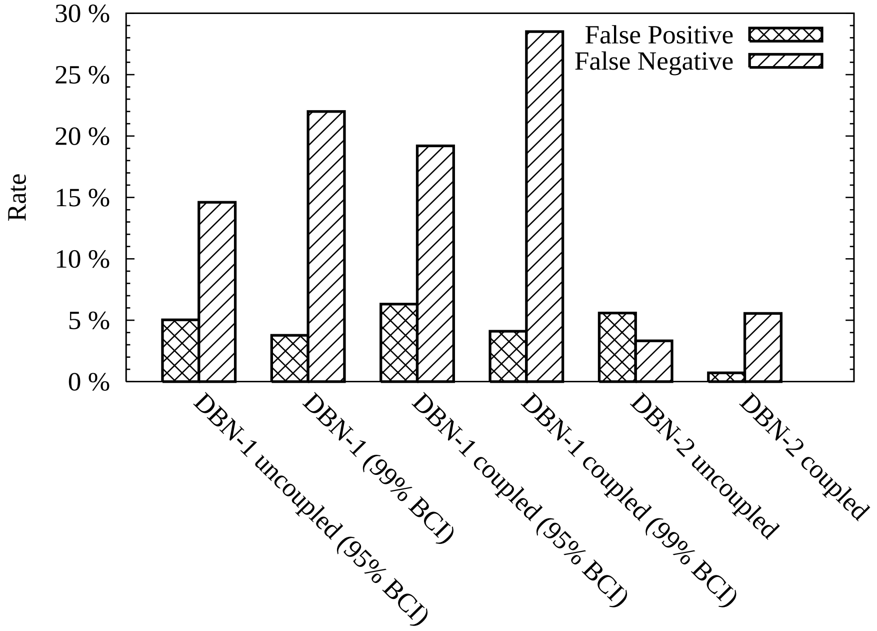

Figure 3: Performance of the anomaly detectors operating on the CC003 windspeed data stream with synthetic errors drawn from R1.

(98^{th}) and 0.05^{th} (99.95^{th}) percentile of windspeed change observed during the month of November, respectively, and such that the ranges do not allow excessively large offsets, which would be easy to classify. Both ranges provide challenging errors for the detectors because there is some overlap between the distribution of valid measurements and erroneous measurements. It is important to note that as the overlap is greater for errors drawn from R1, a greater number of classification errors should be expected using this range. False positive/negative rates were used to compare the performance of the nine fault detectors. The false positive/negative rates indicate the ratio of erroneous/valid data that are misclassified. These rates were quantified based on the assumption that the only errors in the data were the injected synthetic errors.

Figures 3 and 4 show the performance of the nine detectors for identifying errors sampled from R1 and injected into the data streams CC003 and CC009, respectively. From these figures, it can be clearly seen that the DBN-1 based detectors produce many more false negatives than the DBN-2 based detectors, and that coupling appears to degrade the performance of the DBN-1 detectors. This behavior can be attributed to the effects of erroneous measurements on the DBN-1 prediction of the posterior probability distribution of the system states, which are exacerbated when both streams are used for inference.

The coupled version of DBN-2 performs significantly better on errors drawn from R1 than any of the other detectors, with false positive and false negative rates of 0.76% and 5.6% for CC003, and 1.0% and 3.5% for CC009. This result indicates that including a discrete state tracking the status of the measurement and coupling the detection process significantly enhance the ability of the DBN to model the system sufficiently to identify anomalies. However, for these detectors, coupling appears to have the effect of increasing the false negative rate. This result appears to be caused by the number of discrete states that the coupled DBN must use to account for the measurement status. The uncoupled DBN-2 needs only two discrete values to represent

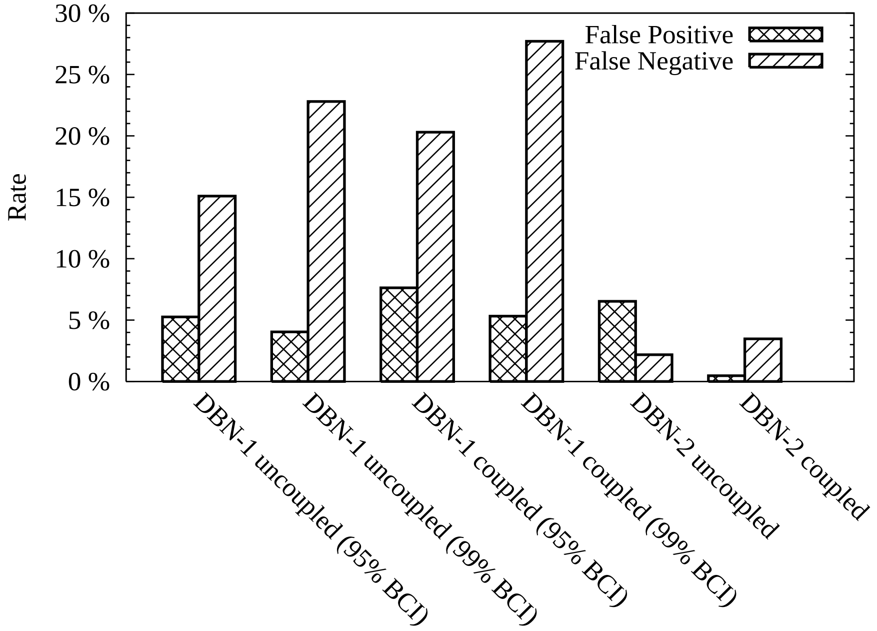

Figure 4: Performance of the anomaly detectors operating on the CC009 windspeed data stream with synthetic errors drawn from R1.

the cases of normal measurement or anomalous measurement, whereas the coupled DBN-2 must use four states to account for the possible combinations of normal measurements and anomalous measurements on both data streams. Thus, in some cases of coupled anomaly detection, where one or both of the measurements are erroneous, the MAP estimate of the measurement status may indicate that neither of the measurements are anomalous, even though fewer than half of the particles support this estimate, because the majority of particles suggest the presence of an error, but are split regarding the type of error they indicate.

Figures 5 and 6 show the performance of the nine detectors for identifying errors sampled from R2 and injected into data streams CC003 and CC009, respectively. From these figures, it can again be seen that the DBN-1 detectors have a higher false negative rate than the DBN-2 based detectors. However, in general, the detectors' false negative rates have decreased from the R1 case, due to the fact that errors drawn from R2 vary more significantly than those drawn from R1 and, thus, are easier to identify. The false positive rates for the DBN-1 detectors appear to have increased slightly from the R1 case, while the false positive rates for DBN-2 detectors appear to have decreased slightly. This trend is caused by the way erroneous measurements are processed by the different DBN's. DBN-1 gives all measurements equal weight in the inference of the posterior state distributions, whereas DBN-2 largely ignores measurements classified as anomalous. Thus, the larger average magnitude of the errors drawn from R2 exacerbate the effect of anomalous measurements on inference with DBN-1, while inference with DBN-2 is improved because the errors are easier to identify and, thus, fewer anomalous data will be misclassified and given the same weight as valid measurements during inference. Finally, it is clearly visible that for detecting errors drawn from R2, the coupled version of DBN-2 performs significantly better than any of the other detectors, with false positive and false negative rates of 0.80% and 0.16% for CC003, and 1.0% and 0.00% for CC009. Again, this result indicates that including a discrete

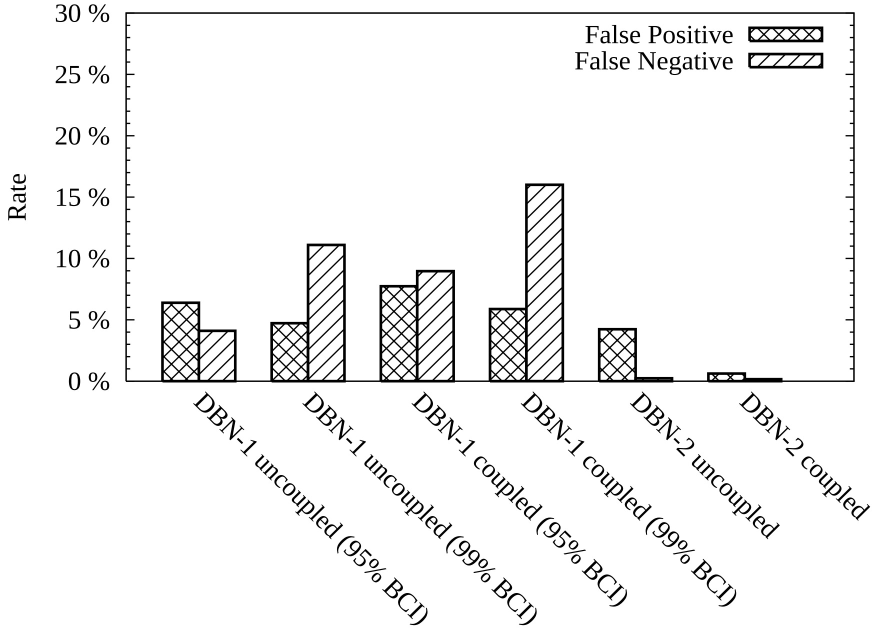

Figure 5: Performance of the anomaly detectors operating on the CC003 windspeed data stream with synthetic errors drawn from R2.

state tracking the status of the measurement and coupling the detection process significantly enhance the ability of the DBN to model the system sufficiently to identify anomalies.

CONCLUSIONS

Real-time detection of anomalies in environmental streaming data has many practical applications, such as data QA/QC, adaptive data collection, and anomalous event detection. This research developed two anomaly detection methods based on DBNs. These methods perform fast, incremental evaluation of data as it becomes available, can scale up to large quantities of data, and require no *a priori* information regarding process variables or the types of anomalies that may be encountered. Furthermore, because the Bayesian framework does not require a specific set of measurements to be available for classification, it is easy to apply to a network of heterogeneous sensors, in which one or more sensors can be expected to fail to report a measurement.

This case study demonstrates the value and efficacy of the proposed anomaly detection methods for data QA/QC. Anomaly detectors developed to process the CC003 and CC009 windspeed data streams from Corpus Christi Bay, using either a DBN that tracked only the windspeed and its measurement at one or both locations (DBN-1) or a DBN that tracked the windspeed, its measurement, and the status of the measurement at one or both locations (DBN-2), performed well at identifying synthetic transient errors injected into the data streams. In particular, the findings show that using DBN-2 to perform coupled anomaly detection on both data streams concurrently produces the best results. This result indicates that including a discrete state tracking the status of the measurement and coupling the detection process both significantly enhance the DBN's ability to model the system. On errors drawn from R1, the coupled DBN-2 had false positive and false negative rates of 0.76% and 5.6% for CC003, and 1.0% and 3.5% for

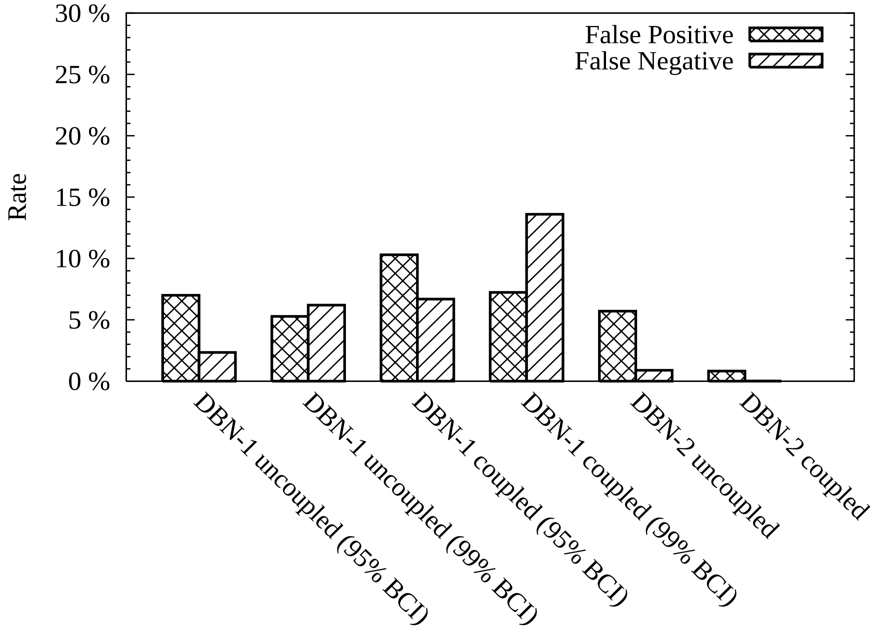

Figure 6: Performance of the anomaly detectors operating on the CC009 windspeed data stream with synthetic errors drawn from R2.

CC009. For errors drawn from R2, the coupled DBN-2 had false positive and false negative rates of 0.80% and 0.16% for CC003, and 1.0% and 0.00% for CC009. Additionally, the results indicate that the false negative rates may be reduced if a larger number of particles are used for Rao-Blackwellized particle filtering, because the accuracy with which the discrete state is modeled improves as the number of particles is increased. Furthermore, because DBN-2 indicates the measurement status through the discrete variable, it could also be useful for modeling other specific types of sensor failures for which the behavior of the failing sensor can be described; thus, not only would the anomaly detector be able to indicate that a measurement was erroneous and identify which sensor reported it, but it would also be able to indicate the most likely cause of the faulty measurement. This information could then be used for remedial action.

This case study only considers two sensor data streams (CC003 and CC009 windspeed). There are many more sensors in the Corpus Christi Bay testbed observatory, which could be added to the coupled DBN-2 anomaly detector. Because these sensors operate at different sampling frequencies from each other and the windspeed sensors, their inclusion in the DBN is non-trivial. We are currently addressing this issue and plan to present additional results quantifying the performance of anomaly detectors that couple such sensors at the conference.

Acknowledgements

The authors would like to thank Desiree Trujillo, John Perez, Terry Riggs, and Jim Bonner of the Shoreline Environmental Research Facility of Corpus Christi, Texas, for providing data and knowledge regarding the sensors used in this study. This study was supported by the Office of Naval Research under grant N00014-04-1-0437.

REFERENCES

Digalakis, V., Rohlicek, J. R., and Ostendorf, M.1993. ML Estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. IEEE Trans. Speech and Audio Proc., 1(4): 431 – 442.

Doucet, A., de Freitas, N., Murphy, K., and Russell, S. 2000a. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In Proc. of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI2000), Stanford, pp. 176-183.

Doucet, A., Godsill, S., and Andrieu, C. 2000b. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing, 10(3): 197-208.

Russell, S. and Norvig, P. 2003. Artificial Intelligence: A Modern Approach, Second Edition. Prentice Hall, Saddle River, New Jersey.

Hill, D. J. and Minsker, B. S. 2006. Automated Fault Detection for *In-Situ* Environmental Sensors. Proc. 7th International Conference on Hydroinformatics, HIC 2006, Nice, France.

Maybeck, P. 1979. Stochastic Models, Estimation, and Control, Volume 1. Academic Press, Inc., Burlington, MA.

Bolton R. J. and Hand, D. J. 2001. Unsupervised profiling methods for fraud detection. Proc. Credit Scoring and Credit Control VII, Edinburgh, UK, pp. 5-7.

Bonner, J.S., F.J. Kelly, P.R. Michaud, C.A. Page, J. Perez, C. Fuller, T. Ojo, and M. Sterling, 2002. Sensing the Coastal Environment. Proc. Third International Conference on EuroGOOS; Building the European Capacity in Operational Oceanography, pp. 167-173.

Box, G. and C. Jenkins, 1970. Time Series Analysis: Forecasting and Control, Holden-Day Inc., San Francisco.

Bulut, A., Singh, A. K., Shin, P., Fountain, T., Jasso, H., Yan, L., and Elgamal, A. 2005. Real-time nondestructive structural health monitoring using support vector machines and wavelets. Proc. SPIE Advanced Sensor Technologies for Nondestructive Evaluation and Structural Health Monitoring (eds. N. Meyendorf, G. Y. Baakline, and B. Michel), Vol. 5770, pp. 180-189.

Hodge, V. J. and Austin, J. 2004. A survey of outlier detection methodologies. Artificial Intelligence Review 22, 85-126.

John, G. H. 1995. Robust decision trees: Removing outliers from databases. Proc. 1st International Conference on Knowledge Discovery and Data Mining, pp. 174-179.

Mourad, M. and Bertrand-Krajewski, J.-L.. 2002. A method for automatic validation of long time series of data in urban hydrology. Water Science & Technology 45(4-5): 263-270

NRC (National Research Council) 2006. CLEANER and NSF's Environmental Observatories. Washington, D.C., National Academy Press.

NSF (National Science Foundation) 2005. Sensors for Environmental Observatories Report: of the NSF Sponsored Workshop December 2004. Arlington, VA, NSF.

Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms of mining outliers from large data sets. Proc. ACM SIGMOD Conference on Management of Data, Dallas TX, 427-438.

Shah, K., J. Bonner, D. Trujillo, C. Page and F. Kelly, 2005. Development of Real-time Data Monitoring System for Coastal Margin Research, Proc. 2005 International Oil Spill Conference, Miami.

Tang, J., Chen, Z., Fu, A. and Cheung, D. 2002. A robust outlier detection scheme in large data sets, Proc. 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taipei, Taiwan, May 2002.