Learning Stable Multivariate Baseline Models for Outbreak Detection

Sajid M. Siddiqi, Byron Boots, Geoffrey J. Gordon, Artur W. Dubrawski

The Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213

OBJECTIVE

We propose a novel technique for building generative models of real-valued multivariate time series data streams. Such models are of considerable utility as baseline simulators in anomaly detection systems. The proposed algorithm, based on Linear Dynamical Systems (LDS) [1], learns stable parameters efficiently while yielding more accurate results than previously known methods. The resulting model can be used to generate infinitely long sequences of realistic baselines using small samples of training data.

BACKGROUND

Time series analysis is an essential component of syndromic surveillance. Any outbreak detection technique, however, must rely on a stable and robust baseline model. Moreover, with the increasing utilization of multi-channel bio-surveillance data streams, it is important that any baseline model scales to multidimensional observations. In many cases, different channels in such data are correlated, and their temporal evolution can be explained by the dynamics of an underlying low-dimensional process. This is precisely the LDS model assumption.

METHOD

Our LDS model and its learning algorithms are explained in more detail in [2]. It assumes that a latent variable x is evolving under Markov assumptions over a sequence of discrete time steps based on linear transformation with a dynamics matrix A. At every time step, a higher-dimensional observation y is emitted according to another linear transformation. Reliable simulation of long sequences requires stability of the dynamics matrix, i.e. that the magnitude of its highest eigenvalue be at most 1. Most algorithms for learning the dynamics matrix from data ignore stability [1]; others are inefficient and less accurate than our approach, sacrificing too much accuracy for the sake of stability. Our method is based on constraint generation for a Quadratic Program optimization. The set of stable matrices is non-convex, making direct optimization difficult, but we utilize the convexity of a closely related set of matrices instead.

RESULTS

We examine daily counts of Over-The-Counter drug sales, a subset of National Data Retail Monitor (NDRM) collection [3]. We first isolate a 60-day subsequence where the dynamics are stationary, and learn stable LDS parameters from the first 15 days. Then we simulate sequences of future baseline values. Figure 1(A) plots counts of 22 different drug categories aggregated over all 29 zip codes in Pittsburgh, PA, and Figure 1(B) plots a single drug category *(cough/cold)* for individual zip codes. In both cases, our method is able to use very little training data to learn a stable model that captures the periodicity in the data, while the naively learned model is unstable and its observations diverge over time. LB-1 learns a model that is stable but is over-constrained, so that the simulated observations dampen quickly.

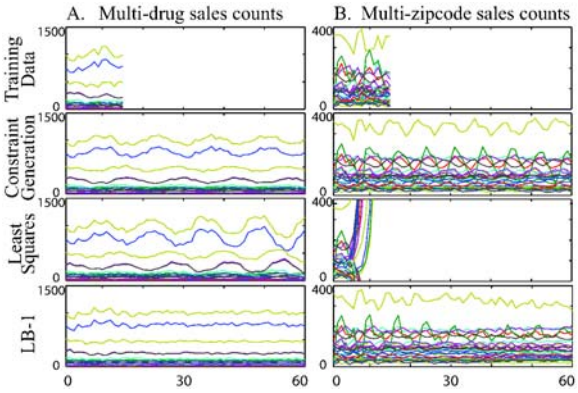

Figure 1. (A) 22-dimensional data corresponding to distinct drug categories aggregated over all zip codes in Pittsburgh. (B) 29-dimensional data for the cough/cold category for individual Pittsburgh zip codes. Top to bottom: the training data, simulated output from our method, output from the naively learned unstable model, and output from the over-dampened LB-1 model.

CONCLUSION

We have proposed a novel technique for learning stable baseline models for multivariate real-valued data streams. The results are promising and indicate that our baseline learning technique can play a valuable role in systems for anomaly detection in multivariate time series data.

ACKNOWLEDGEMENTS

This work was supported by the Centers of Disease Control (award R01-PH000028). This material is based upon work that was supported by the National Science Foundation under grant IIS-0325581.

REFERENCES

[1] Ghahramani Z and Hinton GE. Parameter Estimation for Linear Dynamical Systems. *Tech Report, U. Toronto, CRG-TR-96-2,* 1996

[2] Siddiqi S, Boots B and Gordon G. A Constraint Generation Approach to Learning Stable Linear Dynamical Systems. In *Proceedings of NIPS, 2007*

[3] Wagner M. et al. A National Retail Data Monitor for Public Health Surveillance. Morbidity and Mortality Weekly Report, 2004. 53 (Supplement): 40-42.