Multivariate Time Series Analyses Using Primitive Univariate Algorithms

Maheshkumar Sabhnani, Artur Dubrawski, Jeff Schneider

The Auton Lab, Machine Learning Department, Carnegie Mellon University, 15213

OBJECTIVE

This paper shows how T-Cubes, a data structure that makes tracking millions of disease models simultaneously feasible, can be used to perform multivariate time series analysis using primitive univariate algorithms. Hence, the use of T-Cube in brute-force search helps identify stronger disease outbreak signals currently missed by the surveillance systems.

BACKGROUND

Time series analysis is very popular in syndromic surveillance. Mostly, public health officials track in the order of hundreds of disease models or univariate time series daily looking for signals of disease outbreaks. These time series can be aggregated counts of various syndromes, possibly different genders and age-groups. Recently, spatial scan algorithms find anomalous regions by aggregating zipcode level counts [1]. Usually, public health officials have a set of disease models (for e.g. fever or headache symptom in male adults is indicative of a particular disease). Based on the past experience public health officials track these disease models daily to find anomalies that might be indicative of disease outbreaks. A typical syndromic surveillance system these days will track in the order of 100-200 time series on daily basis using different univariate algorithms like CUSUM, moving average, EWMA, etc.

Let us consider a *representative dataset* of a state which has 100 zipcodes that monitors 10 syndromes among 3 age groups and 2 genders in emergency rooms. There are a total of 6,000 (100 x 10 x 3 x 2) distinct time series for a particular zipcode, syndrome, age-group and gender. This number already seems too high to monitor daily. Hence most syndromic systems only monitor state level aggregates for all syndromes or a few combinations of syndromes, gender and age-groups.

But most real world disease models are more complex and affect multiple syndromes, or multiple age-groups. We need to analyze more complex streams that aggregate multiple values in the attributes to mine more interesting patterns not seen otherwise. As an example, a massive search could reveal that recently senior female patients having fever and nausea have increased in the north eastern part of the state.

TECHNOLOGY

T-Cubes [2],[3] have been recently developed in order to quickly retrieve time series in response to ad-hoc queries. It uses AD-Trees [4] to pre-aggregate counts of simple queries where each attribute (zipcode, gender, etc.) takes a single value (15213, Male, etc.). T-Cubes respond to time series queries in milliseconds [2] as compared to seconds using commercial data cubes. Hence a system that wants to monitor 10,000 time series will now require only approximately 10 secs as compared to 3 hrs. This massive improvement makes it feasible to daily monitor millions of time series looking for new emerging disease patterns or models.

RESULTS

Imagine each attribute in the time series query can take upto ‘k’ values in an attribute. Hence we can only combine ‘k’ syndromes or ‘k’ nearby zipcodes. The number of distinct queries grows exponentially with increasing values of ‘k’. Note that higher ‘k’ implies more complex disease models. We ran experiments on both real-world Emergency room data and over-the-counter sales datasets for varying values of ‘k’ and found very interesting disease outbreak patterns.

CONCLUSIONS

T-Cubes can be used to perform multivariate analysis. It can track millions of disease models using vanilla univariate time series algorithms already being used in surveillance systems. The health officials can now enhance surveillance greatly by incorporating complex disease searches in their current statistical framework that was not practical before. Auton Lab is willing to provide this powerful technology for surveillance purposes.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Neill D, Detection of spatial and station-temporal clusters, Ph.D. Thesis, Carnegie Mellon University, Technical Report, CMU-CS-06-142, 2006.

[2] Sabhnani M, Moore A, Dubrawski A, T-Cube: A Data Structure for Fast Extraction of Time Series from Large Datasets. Technical Report, Machine Learning Department, Carnegie Mellon University, Technical Report, CMU-ML-07-114, 2007.

[3] Sabhnani M, Moore A, Dubrawski A, Rapid Processing of Ad-Hoc Queries Against Large Sets of Time Series, Syndromic Surveillance Conference, 2006.

[4] Moore A, Lee M, Cached sufficient statistics for efficient machine learning with large datasets. Journal of Artificial Intelligence research, 8:67-91, 1998.