ISDS 2012 Conference Abstracts

Searching for Complex Patterns Using Disjunctive Anomaly Detection

Maheshkumar Sabhnani*, Artur Dubrawski and Jeff Schneider

Carnegie Mellon University, Pittsburgh, PA, USA

Objective

Disjunctive anomaly detection (DAD) algorithm [1] can efficiently search across multidimensional biosurveillance data to find multiple simultaneously occurring (in time) and overlapping (across different data dimensions) anomalous clusters. We introduce extensions of DAD to handle rich cluster interactions and diverse data distributions.

Introduction

Modern biosurveillance data contains thousands of unique time series defined across various categorical dimensions (zipcode, age groups, hospitals). Many algorithms are overly specific (tracking each time series independently would often miss early signs of outbreaks), or too general (detections at state level may lack specificity reflective of the actual process at hand). Disease outbreaks often impact multiple values (disjunctive sets of zipcodes, hospitals, multiple age groups) along subsets of multiple dimensions of data. It is not uncommon to see outbreaks of different diseases occurring simultaneously (e.g. food poisoning and flu) making it hard to detect and characterize the individual events.

We proposed Disjunctive Anomaly Detection (DAD) algorithm [1] to efficiently search across millions of potential clusters defined as conjunctions over dimensions and disjunctions over values along each dimension. An example anomalous cluster detectable by DAD may identify zipcode = {z1 or z2 or z3 or z5} and age_group = {child or senior} to show unusual activity in the aggregate. Such conjunctive-disjunctive language of cluster definitions enables finding real-world outbreaks that are often missed by other state-of-art algorithms like What's Strange About Recent Events (WSARE) [3] or Large Average Submatrix (LAS) [2]. DAD is able to identify multiple interesting clusters simultaneously and better explain complex anomalies in data than those alternatives.

Methods

We define the observed counts of patients reporting on a given day as a random variable for each unique combination of values along all dimensions. DAD iteratively identifies K subsets of these variables along with corresponding ranges of their values and time intervals that show increased activity that cannot be explained by random fluctuations (K is generally unknown and could be 0). The resulting set of clusters maximizes data likelihood while controlling for overall complexity. We have successfully derived a versatile set of scoring functions that allow Normal, Poisson, Exponential or Non-parametric assumptions about the underlying data distributions, and accommodate additive-scaled, additive-unscaled or multiplicative-scaled models for the clusters.

Results

We present results of testing DAD on two real-world datasets. One of them contains daily outpatient visit counts from 26 regions in Sri Lanka involving 9 common diseases. The other data contains semi-synthetically generated terrorist activities throughout regions of Afghanistan (Sigacts). Both span multiple years and are representative of data seen in biosurveillance applications.

Figure 1 shows DAD systematically outperforming WSARE and LAS. Each algorithm's parameters were tuned to generate one false positive per month in baseline data. The graphs represent average days-to-detect performance of 100 sets with synthetically injected clusters using additive-scaled (AS), additive-unscaled (AU), and multiplicative-scaled (MS) models of cluster interactions.

Conclusions

We extend applicability of DAD algorithm to handle wide variety of input data distributions and various outbreak models. DAD efficiently scans over millions of potential outbreak patterns and accurately and timely reports complex outbreak interactions with speed that meets requirements of practical applications.

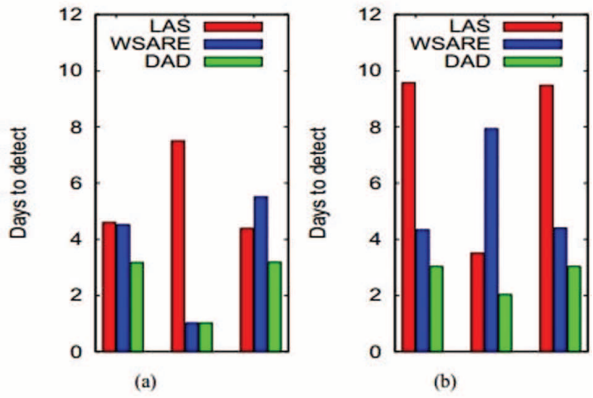

Figure 1: Alg. performance (a) Srilanka, (b) Sigacts

Keywords

outbreak detection; anomalous clusters; disjunctive anomaly detection; prospective surveillance

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0911032.

References

1. Sabhnani M., Dubrawski A., Schneider J. Detection of Multiple Overlapping Anomalous Clusters in Categorical Data. Advances in Disease Surveillance, 2010.
2. Shabalin A., Weigman V., Perou C., Nobel A. Finding Large Average Submatrices in high dimensional data. Annals of Statistics 3(3):985-1012, 2009.
3. Wong W., Moore A., Cooper G., Wagner M. What's Strange About Recent Events (WSARE). J. of Machine Learning Research, 6:1961-1998, 2005.

*Robin Sabhnani
E-mail: sabhnani@cs.cmu.edu