Support Vector Machines for Syndromic Surveillance

Anna L. Buczak, PhD, Linda J. Moniz, PhD, Joseph Lombardo, MS

The Johns Hopkins University - Applied Physics Laboratory, Laurel, MD

Objective

This paper depicts a novel method for reliable detection of disease outbreaks. The methodology and initial results obtained on ESSENCE data are presented.

Background

Early and reliable detection of anomalies is a critical challenge in disease surveillance. Most surveillance systems collect data from multiple data streams but the majority of monitoring is performed at univariate time series level. Purely statistical methods used in disease surveillance look at each time series separately and tend to generate a large number of false alarms. Support Vector Machines (SVM) can be used to develop rich multivariate models that allow detecting abnormal relationships between different time series leading to greatly reduced number of false alarms.

Method

At a high level our anomaly detection approach consists of: 1) Learning the model of normalcy; 2) Detecting anomalies based on their dissimilarity from regular behavior.

For learning the normal behavior we use the SVM algorithm based on statistical learning theory [1]. We use the one-class SVM extension to the algorithm [2] that requires only positive training examples, in our case - the normal (no outbreak) data.

The method (Figure 1) has the following steps:

- Exponential Weighted Moving Average (EWMA) smoothing of each of the syndrome/ subsyndrome/ gender/ age time series with counts.
- Computing the descriptor: each of 3021 numbers is a value of test statistic [3] regularly used in biosurveillance for time series of counts. Value of 3 or more signifies an alert in statistical approaches.
- The descriptor constitutes an input to several SVMs that classify it as normal or anomalous. Each SVM puts a different emphasis on different parts of the descriptor, e.g. GI SVM puts a high emphasis on about 250 numbers in the descriptor related to GI subsyndromes.
- Each of SVM produces a decision whether the data presented to it is normal or anomalous. The Decision Fusion Center combines decisions from the individual classifiers and produces the final decision (alert/ no alert) presented to the epidemiologist.

Initial Results

The training was performed on ESSENCE data with disease outbreaks removed. Testing was performed on both normal ESSENCE data and ESSENCE data with superimposed simulated outbreaks (3-7 day long). Initial results show 94.7% specificity and 54.8% sensitivity of the SVM method. This compares very favorably with pulling univariate EWMA results (94% specificity for 29% sensitivity, and 68.9% specificity for 54.8% sensitivity).

Conclusions

A multivariate SVM-based approach for disease outbreak detection shows promising initial results for reducing false alarms abundant in purely statistical methods.

References

[1] Vapnik, V. N., "Statistical Learning Theory", John Wiley & Sons, 1998.

[2] Schölkopf, S., Burges, C. J. C., Smola, A. J., "Advances in Kernel Methods: Support Vector Learning", MIT Press, Cambridge, MA, 1999.

[3] Burkom H., Elbert Y., Magruder S.F., Najmi A.H., Peter W., Thompson M.W. "Developments in the Roles, Features, and Evaluation of Alerting Algorithms for Disease Outbreak Monitoring", Johns Hopkins APL Technical Digest, Vol. 27, No. 4, 2008.

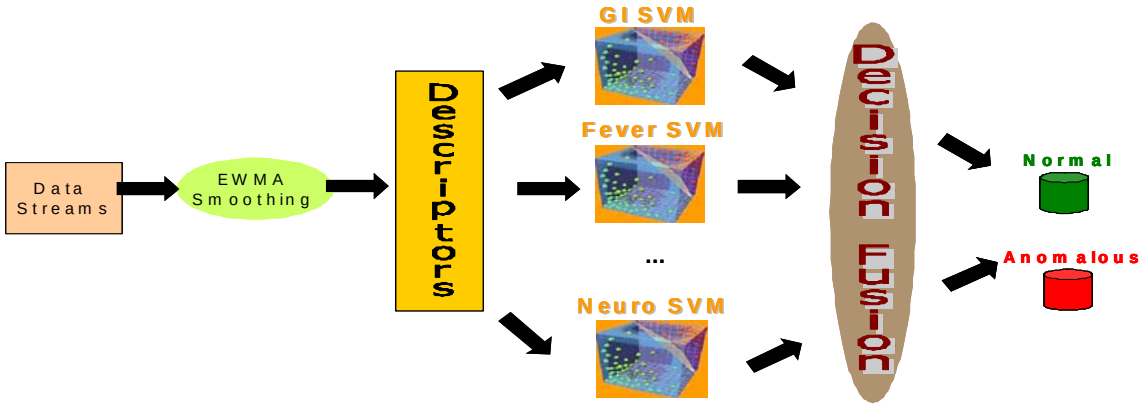

Figure 1. Architecture of SVM-Based Disease Outbreak Detection system