A Multivariate Bayesian Scan Statistic

Daniel B. Neill[1], Andrew W. Moore[1], Gregory F. Cooper[2]

[1]*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213*

[2]*Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15213*

OBJECTIVE

This paper develops a new method for multivariate spatial cluster detection, the "multivariate Bayesian scan statistic" (MBSS). MBSS combines information from multiple data streams in a Bayesian framework, enabling faster and more accurate outbreak detection.

BACKGROUND

The Bayesian spatial scan statistic [1] is an extension of Kulldorff's spatial scan statistic [2] that combines observations and prior knowledge in a principled Bayesian framework. We have demonstrated that the Bayesian method has several advantages over the standard frequentist approach, including higher detection power, faster computation, and easier visualization and calibration. Here we generalize the Bayesian spatial scan to multivariate data, enabling us to combine our priors with observations of multiple data streams and to characterize outbreaks by modeling and differentiating between multiple potential causes.

METHODS

In the MBSS framework, we are given a set of outbreak types $O = \{O_k\}$ and a set of data streams $D = \{D_m\}$. The outbreak types may be either specific illnesses (influenza, anthrax, etc.) or non-specific syndromes. The data streams may include sources such as emergency department (ED) visits, with each stream representing a different chief complaint type, and over-the-counter (OTC) drug sales, with each stream representing a different product group. We are also given a set of spatial regions S to search, where each S contains some subset of the spatial locations s_i. Finally, we are given the count $c_{i,m}^t$ for each spatial location s_i at each time step t for each data stream D_m. Our goal is to compute the posterior probability $Pr(H_1(S, O_k) \mid D)$ that each outbreak type O_k has affected each spatial region S, as well as the probability $Pr(H_0 \mid D)$ that no outbreak has occurred.

To do so, we combine the prior probability of an outbreak in each spatial region with the likelihood of the multivariate data using Bayes' Theorem: $Pr(H_1(S, O_k) \mid D) = Pr(D \mid H_1(S, O_k))\, Pr(H_1(S, O_k) \mid O_k)\, Pr(O_k) / Pr(D)$, and $Pr(H_0 \mid D) = Pr(D \mid H_0)\, Pr(H_0) / Pr(D)$. In these equations, $Pr(H_0)$ is the prior probability of the null hypothesis (no outbreaks) and $Pr(O_k)$ is the prior probability of outbreak type O_k. The probability $Pr(H_1(S, O_k) \mid O_k)$ is the prior probability that outbreak type O_k will affect a given spatial region S. This distribution can be different for different outbreak types: for example, the affected area for a water-borne illness can be predicted based on water distribution information. To compute the data likelihood given the null hypothesis H_0 or an alternative hypothesis $H_1(S, O_k)$, we use a Gamma-Poisson model (as in the univariate Bayesian statistic) for each stream D_m. We typically assume that streams are conditionally independent given the outbreak type, affected region, and outbreak parameters. The parameter priors for each data stream are learned from the time series of past counts, and the effects of each outbreak type on each stream are modeled using expert knowledge. More details are given in [3].

RESULTS

We compared MBSS to the univariate Bayesian approach on semi-synthetic data: simulated outbreaks injected simultaneously into three streams of real OTC sales data (cough/cold, fever, and thermometers) for Allegheny County, Pennsylvania. At a fixed false positive rate of 1/month, MBSS detected outbreaks in an average of 1.59 days, as compared to 2.23, 2.59, and 1.99 days respectively for univariate Bayesian detectors monitoring each of the three data streams. Additionally, MBSS was able to accurately distinguish between different types of outbreak when they had distinct spatial patterns or when they had substantially different effects on one or more streams.

CONCLUSIONS

MBSS increases detection power by combining evidence from multiple data streams and enables us to distinguish between multiple outbreak types. We are currently developing models of several potential outbreak types (e.g. influenza, anthrax) in order to better detect and distinguish between these outbreaks. We are also developing models of other (non-outbreak) causes of a detected cluster. These models will allow us to discriminate between clusters that are due to outbreaks and those due to other irrelevant causes.

This work was supported by NSF grant IIS-0325581.

REFERENCES

[1] Neill DB, Moore AW, Cooper GF, A Bayesian spatial scan statistic. Neural Information Processing Systems 18, 2006, in press.

[2] Kulldorff M, A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997, 26(6): 1481-1496.

[3] Neill DB, Detection of spatial and spatio-temporal clusters. Ph.D. thesis, Carnegie Mellon University.

Further Information:
Daniel B. Neill, neill@cs.cmu.edu
www.cs.cmu.edu/~neill and www.autonlab.org