A Robust Expectation-Based Spatial Scan Statistic

Daniel B. Neill and Maheshkumar R. Sabhnani

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

This paper describes a new expectation-based scan statistic that is robust to outliers (individual anomalies at the store level that are not indicative of outbreaks). We apply this method to prospective monitoring of over-the-counter (OTC) drug sales data, and demonstrate that the robust statistic improves timeliness and specificity of outbreak detection.

BACKGROUND

The expectation-based scan statistic [1-2] is a variant of the spatial and space-time scan statistics [3-4] that enables timely and accurate detection of disease outbreaks by accounting for spatial and temporal variation in baseline disease rates. The expectation-based method first infers the expected count for each spatial location by time series analysis, and then finds spatial regions with counts that are significantly higher than expected. Our SSS system [5] is currently using this method for daily, nationwide monitoring of OTC sales data from the National Retail Data Monitor [6].

Our experience with prospective surveillance of OTC sales has revealed that *outliers* (individual stores with counts that are much higher than expected) are a common source of false positives. These outliers are not due to disease outbreaks, but instead reflect a variety of unmodeled events in the OTC data, including data irregularities, bulk purchases, inventory movements, and promotional sales. Because we expect an outbreak to increase counts in multiple stores in the affected area, while only a small proportion of stores are outliers, we can accurately distinguish between these potential causes of an increase in counts.

METHODS

Our robust scan statistic model assumes that all counts c_i are Poisson distributed with mean equal to the product of the (known) expectation b_i and an (unknown) relative risk q_i. The robust statistic compares the null hypothesis H_0 of no clusters to the set of alternative hypotheses $H_1(S)$, each representing a cluster in some region S, using a likelihood ratio statistic. Under H_0, each q_i is equal to 1 with probability $1-\varepsilon$, and equal to some outlier value o_i with probability ε. Under $H_1(S)$, each q_i is equal to q (inside S) or 1 (outside S) with probability $1-\varepsilon$, and equal to some outlier value o_i with probability ε, for some constant $q > 1$.

The probability of outliers ε must be provided in advance; for $\varepsilon = 0$, this statistic is identical to the original expectation-based statistic. Maximum likelihood estimation is used to determine the values of q, the values of each o_i, and whether each count is an outlier under the null and alternative hypotheses. Randomization testing is performed by generating each count according to whether or not it is an outlier under the null. More details of this approach are given in [7].

RESULTS

We compared the robust scan statistic (with a range of ε values from 10^{-10} to .25) to the standard expectation-based scan statistic for semi-synthetic data: simulated respiratory outbreaks injected into real store-level OTC sales data for western Pennsylvania. The robust statistic reduced the proportion of false positives in the baseline data from 66% to 9%, with higher ε corresponding to fewer false positives. For a fixed false positive rate of 1/month, detection power was maximized for an intermediate value of $\varepsilon = .05$, detecting 95.4% of outbreaks in an average of 5.1 days. The non-robust statistic ($\varepsilon = 0$) detected only 62.4% of outbreaks in an average of 5.7 days.

CONCLUSIONS

The robust scan statistic is a useful method for reducing the number of false positives due to outliers, thus increasing our power to detect true outbreaks. We have also developed expectation-based statistics that are robust to small fluctuations in rate and to distributional assumptions respectively; these statistics can also be used to improve detection power in practice.

REFERENCES

[1] Neill DB, Moore AW, Sabhnani MR, Daniel K, Detection of emerging space-time clusters. Proc. 11th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2005, 218-227.

[2] Neill DB, Moore AW, Methods for detecting spatial and spatio-temporal clusters. In Wagner et al., eds., Handbook of Biosurveillance, 2006, 243-254.

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

[4] Kulldorff M, Prospective time-periodic geographical surveillance using a scan statistic. Journal of the Royal Statistical Society A, 2001, 164: 61-72.

[5] Sabhnani MR, Neill DB, et al., Detecting anomalous patterns in pharmacy retail data. Proc. KDD Workshop on Data Mining Methods for Anomaly Detection, 2005.

[6] Wagner MM, Tsui F-C, et al., A national retail data monitor for public health surveillance. Morbidity and Mortality Weekly Report, 2004, 53 (Supplement): 40-42.

[7] 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