Anomaly Pattern Detection for Biosurveillance

Kaustav Das, M.S., Jeff Schneider, Ph.D., Daniel B. Neill, Ph.D.

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

OBJECTIVE

We propose a new method for detecting patterns of disease cases that correspond to emerging outbreaks. Our Anomaly Pattern Detector (APD) first uses a "local anomaly detector" to identify individually anomalous records and then searches over subsets of the data to detect self-similar patterns of anomalies.

BACKGROUND

Traditional anomaly detection techniques [1] examine each data record individually and find records that are anomalous with respect to the historical distribution of data. However, disease outbreaks typically create many related data records (e.g. multiple patients visiting nearby hospitals with similar symptoms), and may not be evident by examination of any single record alone. Methods such as "What's Strange About Recent Events" (WSARE) [2] detect anomalous patterns by finding differences in the relative counts of records matching particular rules for the current (test) and historical (training) datasets. However, an outbreak may create a relatively small number of anomalous records, and thus have little effect on the total number of records matching any rule. However, if we can distinguish between "normal" records and those that are likely to be anomalous, the outbreak pattern becomes much more evident.

METHODS

We incorporate a local anomaly detector into a rule-based pattern detection method, searching for rules that correspond to a higher than expected proportion of detected anomalies. We first apply the conditional anomaly detector [1] to identify each individual record as normal or anomalous. We then consider all rules of the form R: A = a_j, where A is a subset of up to two attributes and a_j is an assignment of attribute values (e.g. Gender = Male and Syndrome = GI). In order to determine if the subset of the test data corresponding to rule R has an unexpectedly high concentration of anomalies, we compare it to the corresponding subset in the training data. For each rule R, we determine the total number of corresponding records in the test and training datasets ($C(R)_{test}$ and $C(R)_{train}$) and the number of anomalous records in those subsets ($C(R)^{+}_{test}$ and $C(R)^{+}_{train}$). We use a one-sided Fisher's Exact Test on the 2×2 contingency table (Table 1) to test the hypothesis that the proportion of anomalies detected for the given rule R is higher in the test dataset, and report the most significant rules R as anomaly patterns corresponding to potential outbreaks. More details are provided in [3].

	Test	Train
Positives	$C(R)^{+}_{test}$	$C(R)^{+}_{train}$
Negatives	$C(R)_{test} - C(R)^{+}_{test}$	$C(R)_{train} - C(R)^{+}_{train}$

Table 1 – The 2×2 contingency table for APD.

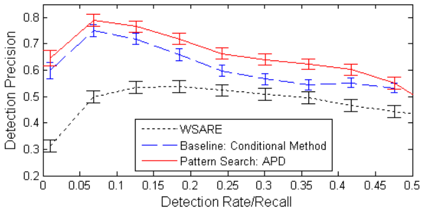

Figure 1 – Precision-Recall curves comparing different methods.

RESULTS

We compared the detection performance of APD with the conditional method [1] and WSARE [2] on simulated anthrax releases [4] injected into real-world Emergency Department case data from Allegheny County, PA. Figure 1 plots the detection precision, i.e. the proportion of detected anomalies that were true anomalies (anthrax cases injected by the simulator), against the detection rate, i.e. the proportion of total true anomalies that were detected. We see significant improvement in detection performance using APD over the baseline method and WSARE.

CONCLUSIONS

By combining a local anomaly detector and a contingency table method similar to WSARE, we achieve higher performance than either component method alone. Our new method is computationally efficient and can be applied to any categorical dataset. In [3], we show significant performance improvements on several other detection tasks including network intrusion detection and finding illicit container shipments.

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8-R01-HK000020-02.

REFERENCES

[1] Das K, Schneider J, Detecting anomalous records in categorical datasets. Proc. Knowledge Discovery and Data Mining, 2007.

[2] Wong W-K, Moore AW, Cooper GF, Wagner M, Rule-based anomaly pattern detection for detecting disease outbreaks. Proc. 18th Natl. Conf. on Artificial Intelligence. MIT Press, 2002.

[3] Das K, Schneider J, Neill DB, Anomaly pattern detection in categorical datasets. Proc. Knowledge Discovery and Data Mining, 2008.

[4] Hogan WR, Cooper GF, et al. The Bayesian aerosol release detector. Stat. Med., 2007, 26: 5225-5252.

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