Online Temporal Clustering for Outbreak Detection

Tomas Singliar[1], Denver Dash[2]
[1]University of Pittsburgh, [2]Intel Research

OBJECTIVE

This paper describes a method of detecting a slowly-growing signal in a large population, based on clustering the population into subgroups more homogeneous in their infectious agent susceptibility.

BACKGROUND

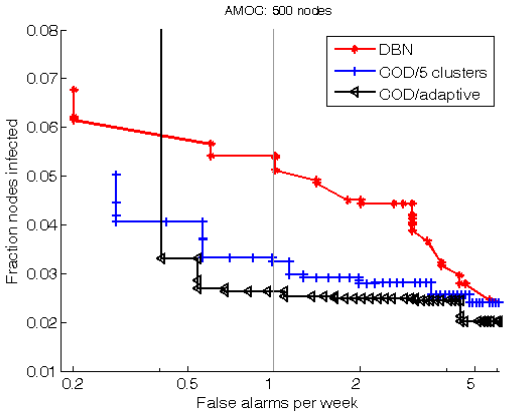
We hypothesize that epidemics around their onset tend to affect primarily a well-defined subgroup of the overall population that is for some reason particularly susceptible. While the vulnerable cohort is often well described for many human diseases, this is not the case for instance when we wish to detect a novel computer virus. Clustering may be used to define the subgroups that will be tested for over-density of symptom occurrence [1]. The clustering slowly changes in response to changes in the population.

METHODS

With each member of the population, we associate a noisy binary indicator of infection, e.g. whether a person has sought medical care with a possibly epidemic symptoms. The indicator status is communicated periodically to a global detector, which clusters the population according to sequences of past indicator values, and other attributes when available. The clustering is based on a graphical probability model with one latent variable which is the cluster membership indicator. The variables in the graphical model are defined so that each variable's value is derived from a fixed time interval of the day. This helps distinguish between outbreak and an expected upswing of background activity. Other periods of background activity (e.g. a week) can be used if more appropriate. The activity level (average number of positive indications) in each cluster is computed and the hypothesis is tested that the most active cluster is in fact significantly more active than the rest of the population.
The method is tested on the task of detecting an intrusion of a stealthy computer worm into an enterprise network [2]. Testing is performed on a simulator running with real background traffic and worm traffic from a parametric model.

RESULTS

We define the target (largest acceptable) false positive rate (FPR) at 1 per week. The proposed method achieves about 40% decrease of infection penetration when operating at the target FPR, as compared with the best detector available at the time, based on a Dynamic Bayesian Network (DBN). Detection performance improves with increasing network size. Computational complexity allows a single desktop computer to run the detection for an enterprise-sized network.



Activity monitoring operating characteristic (AMOC) curves for the previous detector (DBN) and two versions of the proposed method. The curves capture the tradeoff between false positives and time-to-detection resulting from operating the detector at different sensitivity levels. Fraction of nodes infected at detection is an analogue for time-to-detection that has the advantage of being normalized for infection spread speed.

CONCLUSIONS

Dividing the population into subgroups according to susceptibility increases the signal-to-noise ratio and can lead to detection performance boost. The results are relevant to disease outbreak, worm intrusion and other detection problems, such as trend detection in marketing. More ways to exploit the partitioning idea computationally are being studied.

REFERENCES

[1] Singliar, T. and D. Dash. COD: Online Temporal Clustering for Outbreak Detection. In Proceedings of the 22nd Conference on Artificial Intelligence, AAAI, 2007

[2] Dash, D.; B. Kveton.; J. M. Agosta; E. M. Schooler; J. Chandrashekar.; A. Bachrach; and A. Newman. When gossip is good: Distributed probabilistic inference for detection of slow network intrusions. In Proceedings of the 21st Conference on Artificial Intelligence, AAAI, 2006

Further information:
Tomas Singliar, tomas@cs.pitt.edu
Denver Dash, denver.h.dash@intel.com