Perspectives

Machine learning and data mining for epidemic surveillance

Social networking and search engine data confer real-time advantages

The population-level pattern-based nature of epidemiological research makes it well suited for computational work in general, and for machine learning in particular. The social nature of disease spread makes recent trends in social media computing specifically amenable to epidemiological research, but can computational techniques be reliable indicators and predictors of communicable disease?

Previous approaches to epidemiology have had to rely on self-report mechanisms (eg, online health surveillance at www.flutracking.net), or on reports from health care services, such as the United Kingdom Health Protection Agency (HPA), the United States Centers for Disease Control (CDC) and the European Centre for Disease Prevention and Control. The recent growth in the use of social media means that people now volunteer large amounts of information on a real-time and location-specific basis. Along with this, advances in computational intelligence in the form of machine-learning and data-mining techniques have proven useful in knowledge discovery and predictions in other domains.

So, how best can we make use of and interpret that vast quantity of information? The information is of two main kinds: that provided by users (status updates and microblogging, such as on Facebook and Twitter); and requests for information using search engines such as Google and Yahoo. Patterns found in this information can be interpreted by applying machine-learning and data-mining techniques.

Roughly speaking, machine learning and data mining are used to make predictions based on patterns learned from data and discovering patterns in data. In the words of Tom Mitchell, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".[1]

Researchers from the University of Bristol in the UK have been using Twitter to investigate the possibility of tracking influenza spread. They collected about 160 000 tweets per day over 24 weeks from the 54 most populated areas in the UK, in which they sought, for example, reports of sore throat, fever or headache. Reports from the HPA (based on general practitioner consultations per 100 000 citizens resulting in influenza diagnoses) were used as the "gold standard" basis for disease activity when learning to predict influenza rates. Learning and discovering patterns resulted in the ability to predict HPA flu rates with about 90% accuracy.[2]

> "Despite the impressive results, these approaches are not a panacea for epidemic surveillance"

The University of Iowa used a similar approach, whereby public sentiment about pandemic (H1N1) 2009 influenza and actual disease activity was tracked using Twitter posts. Using about a million influenza-related tweets and the CDC's reported data, machine learning was used to construct a predictive model. Even though the model does not predict disease activity, it was able to estimate activity in real-time, reflecting the data in the official CDC reports. The advantage is that "real-time" is typically 1–2 weeks ahead of CDC reports.[3]

People do not use use Twitter to discuss their health — many use search engines to research symptoms, treatments, spread of diseases, and so on. Google conducted an experiment on monitoring search queries related to influenza. An analysis of 5 years' worth of log-files in combination with available CDC information was used to construct a regression model for influenza surveillance, which contains the top 45 search terms. This model was used to predict the spread of influenza in the 2007–2008 influenza season. As with the University of Iowa example, the system was able to consistently estimate the percentage of the population with influenza 1–2 weeks ahead of official CDC reports.[4]

I have highlighted just some of the emerging work on applying clever algorithms and large amounts of computational power to vast amounts of information generated by users of search engines and social media. Despite the impressive results, these approaches are not a panacea for epidemic surveillance. There are challenges, some shared with existing techniques and some that are unique. People discussing their health on Twitter are not representative of the general population, and Twitter use is not uniform across time and geography. Demographic data provided by traditional surveillance cannot (yet) be supplied by search queries. Additionally, it is difficult to separate discussion about epidemics from actual cases.

Regardless of current shortcomings, these approaches will prove to be important parts of modern medicine.

Anders Kofod-Petersen
DrScient,
Senior Research Scientist,[1] and Adjunct Associate Professor[2]

1 SINTEF, Trondheim, Norway.
2 Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway.

anderpe@idi.ntnu.no

doi: 10.5694/mja12.10283

Competing interests: No relevant disclosures.

Provenance: Commissioned; not externally peer reviewed.

1 Mitchell T. Machine learning. New York: McGraw Hill, 1997.
2 Lampos V, Cristianini N. Tracking the flu pandemic by monitoring the social web. Proceedings of the 2nd International Workshop on Cognitive Information Processing; 2010 Jun 14-16; Elba, Italy: Institute of Electrical and Electronics Engineers Xplore digital library, 2010 14 Oct: 411-416. doi: 10.1109/CIP.2010.5604088.
3 Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the US during the Influenza A H1N1 Pandemic. *PLoS One* 2011; 6: e19467.
4 Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature* 2009; 457: 1012-1014.