Diagnosing Outbreaks: Multivariate, Spatio-Temporal Anomaly Detection For Presentations To Hospital Emergency Departments

Bolt, Sarah
CSIRO Mathematics, Informatics and Statistics
Locked Bag 17
North Ryde, NSW, 1670, Australia
E-mail: sarah.bolt@csiro.au

Sparks, Ross
CSIRO Mathematics, Informatics and Statistics

Okugami, Chris
CSIRO Mathematics, Informatics and Statistics

Lind, James
Gold Coast Hospital
Southport, QLD, 4215, AUSTRALIA

Acknowledgements

We thank the Queensland Department of Health for access to the presentation data and the Autralian EHealth Research Centre for their support.

1. Introduction

Every year hospital Emergency Departments (EDs) struggle as the demand on their resources increases. And while predictive tools are already being implemented to assist in forecasting the daily volume of patients (Jessup et al., 2010), they are unable to detect and diagnose when these estimates fall short due to a change in the system. Yet early detection of such changes would help authorities to manage limited health resources and communicate effectively about risk, both in a timely fashion. For example, a new strain of flu might hit the state but affect a particular age group more dramatically and early identification of this group would allow for a shift in resources as well as the possibility of targeted public interventions.

However, a difficulty in developing systems to identify these outliers is distinguishing them from the usual variation observed in these groups. For example, presentations for respiratory infections vary significantly with season and mental health complaints vary greatly with age. Thus any surveillance attempt will require significant energy devoted to understanding typical patterns of behaviour. This is made clear in the survey of outlier detection techniques by Chandola et al (Chandola et al., 2007), where it is observed that most statistical outlier detection techniques can be reduced to two phases:

1. **Training phase:** Determining the probabilistic model from which the data are generated

2. **Testing phase:** Testing if instances are generated by that model or not.

This is certainly true of the EWMA Surveillance Tree method that is the focus of this paper. So we divide the paper accordingly, with one section for each phase. After providing a more detailed

problem specification and some background in Section 2, we approach the phases in reverse, outlining the key features of the testing phase in Section 3. Then in Section 4, we present some of the detail associated with the modelling necessary for the specific task of ED surveillance. Throughout the paper our goals are twofold: we want to both give the reader an understanding of the key components of the EWMA Surveillance Tree method and provide some detail about its application to the problem of ED Surveillance.

2. Background

We begin with a more formal description of the problem and a review of current methodologies. In this project we are given non-identifiable data for each ED presentation across the state of Queensland, Australia. Each record is described by variables in three categories: temporal (time and date of presentation), demographic (age and gender) and presentation type (triage category, disease group based on ICD-10 disease code, departure status and hospital).

In order to consider the data as counts per day we essentially transform it into a large, high dimensional contingency table. In this table, each *cell* is the count of the smallest possible subgroup, that is the number of presentations with a particular disease group, for a particular age, gender, triage category etc. This table is referred to as the *target space* and its dimensions are the *surveillance variables*. Each cell is considered over time and its collective observations are referred to as a *series*.

The goal of the project is then to monitor the behaviour of the cells as new data is added and to detect as soon as possible when this behaviour changes from the expected for any cluster of cells. This problem can thus be classified as prospective, multivariate outbreak surveillance. (Unkel et al., 2011) provides an up-to-date review of prospective disease surveillance methodologies including a section on multivariate detection. They point out one of the key features that these techniques all aim to address: when monitoring many different series they are very likely to be correlated. This means that the simple multiple application of univariate methods to each series is usually inefficient and furthermore, we want to detect groups whose behaviour changes together so for example, their needs can be managed jointly or an intervention targeted effectively.

To date, most methods find outbreaks by essentially performing exhaustive searches in the target space. For example, the extension of the popular spatio-temporal SCAN statistic by Kulldorf defines a test statistic that incorporates an adjustment for multiple testing and then systematically scans the target space applying the test to all windows of the data up to a given fixed size in time and space (Kulldorff et al., 2007). This method has the benefit of being intuitive but has been critisised for being less efficient than some control chart methods (Woodall et al., 2008). However, control chart methods such as the MEWMA control chart method proposed by Joner and Woodall et al (Joner et al., 2008), frequently do not account for underlying changes such as seasonal effects.

There are also some non-parametric approaches such as Wong et al's WSARE (Wong et al., 2003), which compares all possible groups defined by rules of a fixed length with their historic values. This method is extremely computationally demanding as you increase the rule length. However, the latter technique does demonstrate that methods from the machine learning and data mining literatures can be exploited here for their ability to find patterns in high dimensional data sets.

The technique explored in this paper, EWMA Surveillance Trees, combines aspects from all the aforementioned methods. It is inspired by the tree algorithms that are frequently used in machine learning areas for their ability to seek out patterns in high dimensions and incorporates the benefits

of control charts for temporal monitoring by using an EWMA smoothing. All of these contributions are further described in the section that follows.

3. Testing for Outliers with EWMA Surveillance Trees

EWMA Surveillance Trees is a multivariate outlier detection method developped in (Sparks and Okugami, 2009) to monitor numbers of vehicle crashes. At a given time point, the EWMA Surveillance Tree method consists of three major steps for the test of whether the observed data fits the model of expected counts:

1. EWMA based temporal smoothing of observed and expected counts;
2. Growing a Surveillance Tree on departures from expected value in the smoothed counts using a binary recursive partitioning approach;
3. Pruning the Surveillance Tree to reveal outbreaks and control the false alarm rate.

These steps will be outlined in subsections 2.1, 2.2 and 2.3 respectively.

3.1 EWMA Smoothing

Let Y_t be the number of presentations on day t to a cell $\mathbf{x}$. We are given y_t, an observed number of presentations, and from the training phase (see Section 4 below) we also have estimates of the mean $\mu_t = E(Y_t)$ and variance $\sigma_t^2 = Var(Y_t)$.

In order to accumulate the temporal memory needed to detect small changes that persist over time, the Surveillance Trees are built based on an Exponentially Weighted Moving Average (EWMA) of the observed counts. Let $\hat{y}_t$ be the smoothed EWMA of y_t

$$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1} \quad \text{for } t = 1, 2, \ldots \text{ and where } \hat{y}_0 = y_0$$

where α is a suitable constant $0 < \alpha < 1$ that determines how much memory to retain in the average. After applying this smoothing to the observed counts we must now consider its effects on the respective mean and variance, so we consider

$$\hat{\mu}_t = \alpha \mu_t + (1-\alpha)\hat{\mu}_{t-1} \quad \text{for } t = 1, 2, \ldots \text{ and where } \hat{\mu}_0 = \mu_0$$

$$\hat{\sigma}_t^2 = \alpha^2 \sigma_t^2 + (1-\alpha)^2 \hat{\sigma}_{t-1}^2 \quad \text{for } t = 1, 2, \ldots \text{ and where } \hat{\sigma}_0^2 = \sigma_0^2$$

In order to begin the testing phase, we need a measure of how far the smoothed counts depart from the expected. The response variable, z_t, considered in this project is the usual standardisation to a statistic with mean zero and variance one:

$$z_t = \frac{\hat{y}_t - \hat{\mu}_t}{\hat{\sigma}_t}$$

3.2 Growing the Surveillance Tree

The response variable z_t is then used to grow a Surveillance Tree at each time point. The Tree is grown using a binary recursive partitioning approach whose goal is to identify regions in the target space with unusually high departures from expected counts.

The process begins with the whole target space and the focus for each partition is to find a region with (in some sense) an unusually high value of z_t. At each stage of the tree growing process, given

a parent region of the target space, we calculate the value of a test statistic for all possible binary partitions on each surveillance variable[1]. The partition which maximises the test statistic is chosen and the parent region is split on that variable into two offspring, one of which has the unusually high smoothed counts and the other which is simply the remainder of the parent region. The process is then repeated now considering each of the two offspring as parents. Each generation of offspring is grown in the same way and gives rise to a representation of the target space by means of a tree data structure referred to as a *Surveillance Tree*.

The naive test statistic for each partition is simply z_t itself. However we have variables of different types and sizes. For example the variable 'Gender' has only one possible partition whereas the variable 'Age' has over 100. To make the variable selection process equally likely for each variable, we use the same approach as in (Sparks and Okugami, 2009). We generate parametric bootstrap samples from the model of in control behaviour over time and grow Surveillance Trees on these samples. The result is data on the range of values for maximising z_t scores for in-control situations.

This data is used to model, for each variable, the location and spread of in control z_t scores conditional on variables such as the amount of searching, $\hat{\mu}_t$ and z_t in the parent, and $\hat{\mu}_t$ in the node itself. So, if μ^* and $(\sigma^*)^2$ are the respective conditional estimates of mean and variance, then the final test statistic used to choose partitions is

$$z_t^* = \frac{z_t - \mu^*}{\sigma^*}$$

Once partitioning has been completed, then recursive pruning of the terminal nodes commences.

3.3 Pruning the Surveillance Tree

The aim of pruning is to trim away all insignificant nodes. If all nodes in the tree are pruned away for a particular time point then no outbreak is signalled, however if nodes remain after pruning is completed, then an alarm is given. The outbreak is diagnosed by the set of partitioning rules that define the remaining terminal nodes.

Again, the pruning process is given in more detail in (Sparks and Okugami, 2009) but perhaps the most important aspect of the pruning strategy is the one designed to control the false alarm rate. Nodes are pruned recursively starting with the last offspring in the tree. Nodes are pruned if their signal to noise ratio fails to exceed an upper threshold value. This threshold is used to control the false alarm rate.

The three steps outlined above complete the testing phase. However an area of crucial importance when it comes to applying the EWMA Surveillance Tree testing procedure is providing an adequate model of the expected behaviour of the system. The following section explains the modelling approach used for the ED Surveillance problem.

4. Training the Model

In order to apply the EWMA Surveillance Tree test for unusual behaviour we must first develop a model for the expected counts of presentations for all possible subgroups. For example, we need to be able to forecast the expected number of patients on a given day at a particular hospital, with a

[1]Note that for unordered categorical variables (Sparks and Okugami, 2009) provides a method for finding partitions without searching all possible binary splits. The method involves ranking the categories of the variable by their z_t and then treating the variable as if it were ordered

particular disease, of a particular age, etc. It is important to remember the primary goal of the model we seek here: it is not a model for long term forecasting or indeed a model to completely explain a set of data. For surveillance we want to the model to characterise the behaviour of the system when in control and be able to forecast one day ahead for our selected training period.

When trying to characterise the behaviour of such a complex system, it is crucial that we incorporate domain knowledge of known behaviours. In this application, the domain understanding was at two levels. Firstly there is existing work identifying what factors influence the total volume of patients to EDs (Sparks et al., 2010), for example: annual seasonal effects, day of the week contributions, public and school holiday influences, and transitional effects. And at the individual presentation level we also know there are some strong interactions between demographic variables such as age with presentation variables such as triage category.

As well as incorporating this domain knowledge, the other challenges we addressed in this project were:

- The inclusion of predictor variables of different types (nominal, ordered categorical and continuous).
- The sparsity of the system when we consider counts at such a detailed level of classification.
- The importance of not only modelling the mean of the system but also capturing the variation in order to correctly establish unusual cases in the testing phase.
- Lastly, the scale of this problem poses computational challenges. Even storing the counts in memory for this large target space across many time points causes problems.

In trying to address all of these challenges, we employ a 'divide and conquer' approach. Since the domain knowledge of the process of arrivals is at two scales we divide the modelling problem similarly. Rather than one large table to be modelled over time we consider each disease group separately and consider two levels within each group: total counts to be modelled over time and a table aggregated over presentation and patient characteristics assumed to be fixed in time. The details of each of these levels is outlined below.

4.1 A time dependent model for total counts

For each disease group, we firstly develop a transitional regression model for total counts over time. So let Y_t be the total volume of patients to that group on day t, whose expected value λ_t we model as a function of time using either a transitional Poisson or negative binomial model:

$$\log(\lambda_t) = \beta_0 + \sum_{j=1}^{n} \beta_j f_j(t) + \sum_{j=1}^{m} \beta_{j+n} y_{t-j} \tag{1}$$

Where $f_j(t)$ are functions of time such as harmonics, or indicators for day of the week or holidays and where y_{t-j} are lagged, observed counts going back m days.

This high level modelling allows for incorporation of domain knowledge about the timing of presentations and presents few computational demands.

4.2 Expected proportions to cells

Section 4.1 models the total number of presentations for a disease group so we then need a way to allocate these counts to all the cells of the target space. In this project we assume that this allocation

remains constant over time and is independent of the total number of presentations.

To model the allocation of counts to cells we use a Poisson regression tree approach. We sum the data over time for each cell and train a regression tree on these aggregated counts. Let X be the set of all cells to be modelled, the resulting tree gives for each cell, $\mathbf{x} = (x_1, x_2, \ldots) \in X$, an expected count, $\nu(\mathbf{x})$, for the whole training period . These estimates are then used as simple proportions, independent of total volume:

$$p(\mathbf{x}) = \frac{\nu(\mathbf{x})}{\sum_{\xi \in X} \nu(\xi)} \tag{2}$$

While the assumption that this allocation remains constant over both time and total volume is unlikely to hold true for most disease groups, little is known about any systematic changes in this process. Consequently, outliers that are detected in the testing phase will be outliers with respect to this assumption rather than with respect to the 'true' process. This is not really an issue because if the 'true' process is not understood in the domain area then delivering outliers relative to this process would be equally a mystery.

The advantages of using this regression tree approach are that:

- By aggregating the data over time we achieve a computationally significant dimension reduction.
- Variable of different types are easily included and exploited to their full potential.
- Regions of very low or zero frequency are grouped together and are given low (but non-zero) expected values.
- Interactions are naturally included.

4.3 Expected counts to cells

Lastly, for a given cell, disease group and time, the expected number of presentations is then the product of equations 1 and 2:

$$\mu(\mathbf{x}, t) = \lambda_t \times p(\mathbf{x}) \tag{3}$$

This combination of the two models thus allows us to bypass the computational issues associated with such a high dimensional problem while simultaneously allowing for the inclusion of domain knowledge at the two different scales and addressing a number of other problem constraints.

5. Conclusion

We have now given in detail how a model was developed for the ED surveillance problem. These expected values are then used to find regions of outliers in the target space using the EWMA Surveillance Tree method for testing. Results from the application of both phases will be presented in the talk.

However, it is clear from this exposition of the theory how fundamental the dimensionality of the application is to both phases of the outlier detection problem. The core of both the testing and modelling parts are tree based, a method used increasingly frequently across many areas of statistics and informatics to cope with issues of dimensionality. This observation is important since, given the highly multivariate nature of modern data collection, research into multivariate outlier detection methods is likely to be an area of growth in the future across many different fields.

REFERENCES (RÉFERENCES)

Chandola, V., Banerjee, A., and Kumar, V. 2007. Outlier detection: A survey. Technical report, University of Minnesota.

Jessup, M., Wallis, M., Boyle, J., et al. 2010. Implementing an emergency department patient admission predictive tool. *Journal of Health Organization and Management* 24:306–318.

Joner, M. D., Woodall, W. H., Reynolds, M. R., et al. 2008. A one-sided mewma chart for health surveillance. *Quality and Reliability Engineering International* 24:503–518.

Kulldorff, M., Mostashari, F., Duczmal, L., et al. 2007. Multivariate scan statistics for disease surveillance. *Statistics in Medecine* 26:1824–1833.

Sparks, R., Carter, C., Graham, P., et al. 2010. Understanding sources of variation in syndromic surveillance for early warning of natural or intentional disease outbreaks. *IIE Transactions* 42:613–631.

Sparks, R. S. and Okugami, C. 2009. Surveillance trees: early detection of unusually high number of vehicle crashes. *InterStat* .

Unkel, S., Farrington, C. P., Garthwaite, P. H., et al. 2011. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Technical report, The Open University, UK.

Wong, W.-K., Moore, A., Cooper, G., et al. 2003. What's strange about recent events. *Journal of Urban Health* 80:i66–i75. Supplement 1.

Woodall, W. H., Marshall, J. B., Joner Jr., M. D., et al. 2008. On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of the Royal Statistical Society Series A* 171:223–237.