Analyzing Over-the-Counter Medication Sales for Early Detection of Epidemics and Bio-Terrorism Attacks

Anna Goldenberg ANYA@CS.CMU.EDU
Center for Automated Learning and Discovery, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

Abstract

Early detection of epidemics and bio-terrorist attacks is a vital problem that has concerned people for ages. Current techniques for early detection concentrate on analyzing symptom-specific data originating from emergency room visits, lab tests, etc. However, most people do not call emergency services immediately but start by taking pain killers, looking for symptoms on the web, calling friends, etc. By the time the medical community has received a substantial amount of calls to get concerned - it may be too late. Preliminary studies showed evidence of correlation of flu with non-specific syndrome data. In this paper we provide a framework and discuss preliminary results of using over-the-counter medication sales to predict epidemics and/or bio-terrorism attacks.

1. Introduction

The problem of epidemics and bio-terrorism attacks is related to public health and therefore is one of the most vital but also controversial problems. The possibility of early detection of epidemics has been researched for centuries. Many ideas and plans were implemented in order to prevent or detect epidemics as early as possible. However, even though vaccines to prevent the known severe diseases were developed, when the new strain of a known or an unknown virus is introduced to the environment, the course of prevention of its evolution and spread is up to an early detection system. Unfortunately, early detection systems are still in the stage of development.

Though it might seem that the task of creation of an early detection system is a simplified version of designing a prevention system, it remains a non-trivial problem. Several important questions are to be answered: what are the sources that first notice the increase of the public health threat? Where is the information that could hint the first signals of sickness in a large group of people? At the present moment the concentration is on making inferences from medical records, emergency room calls, etc. But, most people do not call emergency services immediately. They start by taking painkillers, looking for symptoms on the web, calling friends, etc. By the time, the medical community has received a substantial amount of calls to get concerned - it may be too late.

In this paper we show that it is feasible to use non-specific syndrome data, such as over-the-counter medication sales, for early detection of epidemics and bio-terrorism attacks. We suggest and fully develop a model that can be implemented in the regular supermarket or a pharmacy store. The method consists of a combination of several data manipulation and time series analysis and forecasting techniques to increase the detection accuracy. Each of the stages of the model is described in detail in the Approach section. The framework provided in the paper can be extended to other non-specific syndrome data, provided that there is prior evidence of correlation with the symptom specific data.

To prove the validity of the model it was tested on the aggregated daily counts of cough medication sales. The results are encouraging. It was shown that the system performs well under both scenarios: when the epidemic was known to occur in the near future (within 30 days) and when there was no known abnormality.

The flu predictions were validated based on the Center for Disease Control and Prevention (CDC) influenza weekly reports compiled from observations of existing surveillance systems. [1] Results suggest that the propsed method might revolutionize the conventional approach to the Computer Based Surveillance Systems (CBS).

2. Approach

The task of creation of an early detection system is far from trivial. It is complicated more by the fact that the data being non-specific to the task has a lot of detail and noise that make the design of the system challenging. To deal with the complexities of the task and intricacies of data a multilevel process is proposed:

[1] The weekly influenza surveillance reports are available at the CDC site http://www.cdc.gov/ncidod/diseases/flu/fluvirus.htm

1. Prepare the data by normalizing the counts and removing the known periodicities
2. Create a smoothed model of the data
3. Decompose the model into several components
4. Make predictions for each of the components
5. Combine the predictions and compare to the incoming real-time data.
6. Raise an abnormality flag if the data is outside the confidence band determined by a threshold function or adjust the model to reflect the new data, if the data flow is regular.

It is important to notice that we are trying to predict the smoothed model rather than the real-time data therefore we do not expect the prediction to be as close to the data as possible, rather to be within the predetermined confidence band from the smoothed model. It is however important that the smoothed model is predicted as accurately as possible, in order to keep the reliability of the system high. The threshold function can be an arbitrary constant depending on the estimated model accuracy or function of cost of an action committed as a response to and based on the number of raised flags.

Each of the stages of the model is discussed in more details in the following section.

2.1 Model Description

2.1.1 DATA PREPROCESSING

It is natural to assume that the data is highly periodic, for example, majority of the supermarket stores are closed at night, so even though bigger stores maybe open, the total amount of purchases during that time is going to drop significantly, therefore daily periodicity can be expected. It is also a fact that there are a lot more sales on weekends than during the week, so we propose to normalize the data by the total sales to dampen the week/weekend effect. According to Mark Airhart (1997), there are more flu/cold related sicknesses in winter as opposed to summer, therefore seasonal periodicity can be expected. To account for the known periodicities in the data, we propose to remove those frequencies as part of data preprocessing. It can be argued that if the epidemic occurred in seven-day type patterns, then by removing seven-day periodicities we would be removing the signal, however the probability of such signals is very low.

2.1.2 SMOOTHED MODEL CREATION

To achieve the goal of creating a reliable early detection system, we have to provide substantial confidence level in the predictions made by the system. Therefore, we are not interested in predicting the data perfectly as it was already mentioned above. The signal appears to be very noisy and even with the most complicated non-linear prediction tools the reliability of the prediction would not be very high. In addition to that, if the model were to fit the training data perfectly, i.e. all of the data available up to the moment were taken into account then the model would be overfitting the data, taking into account random noise. Since our goal is to predict the future sales as reliably as possible, taking into account the specified confidence band in order to insure high confidence in our abnormality detection, we propose to smooth or de-noise the data. And make predictions for the smoothed model.

There exist many techniques for smoothing and de-noising. Since the medication sales data depends on time and is highly periodic, standard Times Series Analysis (TSA) techniques can be used. For the purposes of this paper we are proposing the Discrete Cosine Transformation (DCT) smoothing. It can be argued that there are more powerful de-noising techniques, such as wavelet de-noising, but since the purpose of smoothing is not to model the localized signal but rather to create as detailed as possible but still a general trend, DCT seems to be a more appropriate choice.

Discrete Cosine Transform computes the cosine transformation of an input vector. Mathematically, the unitary DCT of an input sequence x is

$$y(k) = w(k)\sum_{n=1}^{N} x(n)\cos\frac{\pi(2n-1)(k-1)}{2N}, \quad k=1,\dots,N$$

where $w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k=1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases}$, N – length of x

and x and y are the same size. DCT is very closely related to Discrete Fourier Transform (DFT). In fact, the only differences are that the cosine transform approximates the function using cosines and not complex cosine-sine functions, working on the real scale. This allows DCT to have better energy compaction properties, with just a few of the transform coefficients representing the majority of the energy in the sequence. A sequence can often be reconstructed very accurately from only a few DCT coefficients, which is a very useful property for applications requiring data reduction, as it is the case in this project. The result of the DCT transform is a vector of numbers (coefficients calculated according to the formula specified above). To reconstruct the series, the inverse discrete cosine transform (IDCT) has to be used, mathematically described as follows:

$$x(n) = \sum_{k=1}^{N} w(k)\,y(k)\cos\frac{\pi(2n-1)(k-1)}{2N}, \quad n=1,\dots,N$$

and w(k) is defined as before.

We have the power of selecting the smoothed model of arbitrary coarseness by setting a threshold on the coefficients that we would like to ignore, for example, set to zero all coefficients that are below 0.1 Naturally, if we set the threshold to 0, the original vector would be reconstructed with 100% accuracy. The selection of threshold can be based on the error measure, such as root mean square error between the original data and the smoothed model, as we are doing in this paper.

2.1.3 Decomposition

In order to preserve the accuracy of the detection system, the smoothed model would still have to be fairly complicated and hard to predict. In order to correct for this complication in the model, it is suggested to decompose the smoothed model using wavelets to obtain several components of the original signal to then use them separately for prediction. The reason behind such an idea is the assumption that each individual signal reflects different aspects of the data whereas the merged signal has these effects embedded in a hard-to-recover format.

Wavelets were originally introduced in 1930s by several groups of scientists, coming from different research areas from physics, mathematics and other fields and has been developed further to become what is used now in the mathematical packages in 1990s (Graps, 1995). The Discrete Wavelet Transform (DWT), that is used in the system is similar to DFT or DCT. Both of them are linear operations that generate a data structure that contains segments of various lengths in the time domain, usually filling and transforming it into a different data vector in the frequency domain. They also have similar mathematical properties.

The wavelet coefficients must fulfill the following requirement: $\int_{-\infty}^{\infty} |f(x)|^2 \, ds < \infty$, i.e. it is required that wave decays to zero at $\pm\infty$. The basis or mother wavelet is dilated by j and translated by k as follows to form a basis: $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, where j, k are integer values. The wavelet transform uses these basis vectors as Fourier Transform uses its sin(kt) and cos(kt). They span the model domain at different resolutions. The analyzing wavelet is used in the following equation:

$$W_\psi f(b,a) = |a|^{1/2} \int_{-\infty}^{\infty} f(x)\psi\left(\frac{x-b}{a}\right)dx,$$

where $f \in L^2(\Re)$, with $a=2^{-j}$, $b=k/2^j$.[2]

There are several major differences, however, that make DWT a better choice for the particular task of decomposition. First of all, the resolution can vary, for example, in order to isolate signal discontinuities, it would be useful to use very short basis functions. On the other hand, to obtain the detailed frequency analysis, long low-frequency basis functions have to be utilized. This kind of variation is not possible in FFT or DCT since by definition a single window of particular width has to be used for all frequencies in the window FT analysis. It is also important to note that wavelets can have an infinite set of basis functions whereas Fourier analysis takes advantage of just sine and cosine functions. This feature helps wavelets take advantage of more information that the model might have than the FT or DCT could.

2.1.4 Component Predictions

Now, that the model has been decomposed and is represented by a set of wavelet components, the forecasting technique should be applied to each of them. The prediction of sales time-window can vary depending on the number of days desired to be predicted. It should be noted that the window should not be too large, in order to have high confidence in the forecasted data. For example, the experiments show that the predictions are not reliable if the prediction time frame exceeds 30 days, given 365 days of training data. That window size is large enough to make a reliable claim of sales counts for the upcoming month, which is sufficient for the purposes of early detection, considering that the model gets more refined and accurate with time as more information becomes available.

There are many tools available for forecasting. Since the model is decomposed into wavelet components that are fairly smooth, it is in fact plausible to make accurate predictions using linear models. For example, given enough data (450 days of sales history), the predictions for 30 days in the future were validated using the hold out set and were 98% accurate on average for all five wavelet components, using a standard technique of autoregressive (AR) forecasting. The covariance method was used to fit an AR model of order of the size of the prediction time-window. This method minimizes the forward prediction error in the least-squares sense. The AR models can be described by the following equation:

$x_t = \sum_{i=1}^{n} a_i x_{t-i} + z_t$, where $z_t \sim WN(0,\sigma^2)$, $|a_i| < 1$ and z_t is uncorrelated with x_s for each $s < t$.[3]

2.1.5 Final Prediction and Comparison

Results of the forecasting obtained in the previous step are now combined together in order to obtain the final prediction. Direct summation of the values is equivalent to the forecast of the model the components were derived from, i.e. the smoothed model. The accuracy of the prediction made can be assessed based on the confidence in the smoothed model calculated in the Smoothed Model step in combination with the prior information on the prediction accuracy of the forecasting tools.

In real time settings, the data from new sales would start coming in and can be compared directly to the prediction of the smoothed model. If the data falls within the

---

[2] For more information about wavelets, please refer to Amara Graps (1995).

[3] There is a great variety of literature on linear models, such as Brockwell and Davis (1996), Weigend and Gershenfeld (1993), etc.

confidence band, or even more precisely, is under the threshold obtained from the confidence in prediction and various cost functions set forth by the authorities, the data is declared to have the “normal flow”, i.e. the abnormality flag is not triggered. In this case new data becomes part of the training set and the model is adjusted to retain the previous accuracy level if needed. If, however, the sales are higher than expected, the abnormality flag is raised and depending on the magnitude of the difference either the warning or a request to notify authorities of immediate action is generated. In this case the model is not adjusted.

Since the outcomes of the system can be of great significance to the nation, it is expected that some posterior analysis such as epidemiological evaluation would be requested by the government.

2.2 Evaluation and Validation

During the initial or training stage of the system before implementation in the real time setting, it is suggested that the models are tested with the hold out sets. Steps as described above should be carried out and the hold out set should be used to verify the validity of the model and give preliminary estimates of accuracy and reliability of the system. It is expected that the precision and robustness of the system will increase as more data becomes available.

3. Results

The most important result of the project is the evidence of support of the idea of using non-specific syndrome data such as over-the-counter medication sales for early detection of epidemics of flu, obtained from the experiments (see Figures 1 and 2). Fortunately or not, there is no validation data available for detection of bio-agent releases, so no accurate conclusions can be drawn from the experiments regarding bio-terrorism attacks. However, the framework is general enough, not to dismiss the possibility of correct identification. The model should presumably work at least for agents that can be modeled with flu data, such as inhalational anthrax.

The framework proposed in the project is a complicated system that is a product of integrating several different methodologies. Each of the steps can be performed using different techniques conceivably achieving similar results. When designing and performing experiments, special attention was paid to providing valid grounds for the necessity of each of the stages and reasons for applying the particular methods suggested.

3.1 Data Description and Experiments Setup

The data used in the experiments are the counts of supermarket over-the-counter cough medication sales aggregated by day over the area of the Allegheny County with zip-code starting with 152. Two different subsets were selected for training. The sizes were 176 and 365 data points (days) respectively. The reason behind such selections is that the first dataset is on the verge of flu epidemic occurrence, i.e. there is data available from the CDC agency with information about the flu epidemic that has occurred within the 30-day time frame after the last point in the training dataset. As expected several of the flags were raised (see Figure 1). On the other hand, when working with the second dataset, no flags were expected to be raised since there was no significant abnormal activity. Thus, the datasets were selected to reveal several characteristics of the system. The occurrence of flu epidemics was judged based on the weekly surveillance CDC reports.

3.2 Results of applying complete model to the datasets

Since the first dataset was relatively small, the model chosen to fit the data was rather smooth:

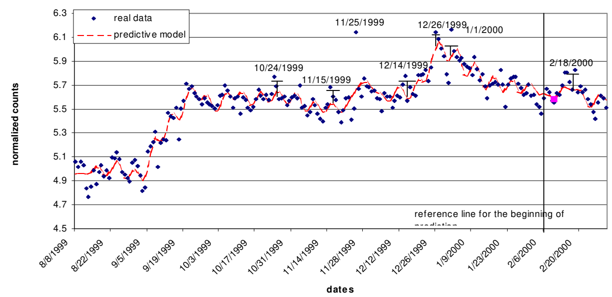

Figure 1. First dataset of 176 data points plotted with predictive smoothed model and outliers. The outliers have date stamp next to them. Reference line identifies the prediction zone. Error bars are based on the one- sided 98% confidence band.

Scripts were written for matlab in order to automate the smoothing model selection based on desired parameters for the model. For the purpose of illustration, the selection of the model was determined by the RMS error between the model and real data and also the number of false alarms. The set of parameters can be expanded to accommodate requirements set forth by the government agencies. The threshold for the FFT values was selected to be 0.2, in the particular case, which led to 139 frequencies to be set to zero. The overall model has made only one mistake in identifying the epidemic dates, i.e. it raised one false alarm and six positive flags correctly, including the one in the prediction zone.

The purpose of the second trial was to show that the data can be predicted accurately enough not to cause extraneous false alarms. In reality there were no spikes occurring during the prediction time (30 days starting beginning of August 2000) and the model has predicted real data with 98% accuracy.

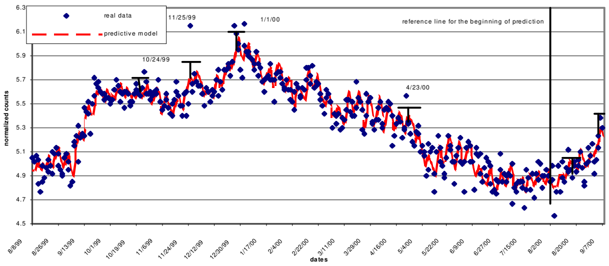

Figure 2. Second dataset of 365 points plotted with predictive smoothed model and outliers. Reference line identifies the prediction zone. Error bars are based on one-sided 98% confidence band.

As it can be seen from the graph, even the questionable points still do lie within the desired 98% confidence band. The model in the second case is more precise. Since more data was available when building the model, there were more features that required larger frequencies and roughly similar amount of noise has remained. Again, the predictions were correct, however in addition to one false positive alarm, there was one false negative. This can be explained that by providing more information during the model creation stage, it tends to overfit more, producing the false negatives. If the threshold value (the accuracy level of the confidence band) depended on cost of action taken when a warning is issued, it would have been expected to have several negative alarms, to reduce the costs of false positives. The precision of the system however, can be easily controlled based on the constraints placed on the system by the appropriate authorities and can be incorporated into the set of parameters that are used to calculate thresholds.

It is important to notice that validation of the system is not an easy task. Surveillance systems in place can only produce reports in weekly aggregates and are usually regional rather than local. The results of the particular project, therefore, had to be evaluated with help of an expert in epidemiology.

4. Previous Work

Sponsored by Defense Advanced Research Projects Agency, DARPA, there have arisen many initiatives for creating early detection systems, most of them however are concerned with slightly different issues, such as developing a prototype database and Web site to give clinicians the opportunity to report suspicious trends of possible bio- terrorist events; assessing the current capacity of hospitals and health systems to respond to a bio-terrorist attack; or evaluate and improve linkages between the medical care, public health, and emergency preparedness systems to improve detection of and response to a bio-terrorist event. Since the "bio-terrorism awareness" project has been open relatively recently and most of the initiatives do not overlap, there is not much history in developing systems monitoring real-time data in order to prevent bio-terrorist attacks and epidemics.

In fact, the subject of using nonspecific syndrome datasets such as evaluating the trends in over-the-counter medication purchases to detect epidemics or bio- attacks is not known to have been researched up-to-date. The Center for Disease Control and Prevention (CDC) keeps track of the most recent developments in the area of early-detection surveillance systems. The information on the systems that are available and comply with government standards is available from their site at any time. So far, the major developments have been done only in area of symptom-specific systems that by nature can not provide information fast enough especially in the situation of a release of a bio agent that can cause death in a large group of population within a week. An example of such a bio-agent is anthrax (Goldman, 2000).

Most of the analysis performed on the supermarket data has also had a different purpose. The research, often sponsored by the private corporations owning the particular supermarket chain, has concentrated on developing pricing strategies and analyzing the pricing sensitivity (see Boatwright, McCulloch and Rossi, 1999; Montgomery 1997, etc.) , rather than modeling base flow of every day sales for non-price related anomaly detection.

5. Summary

The problem of early detection of epidemics and bio-terrorism attacks has been researched for centuries. There are surveillance systems in place that evaluate epidemiological data received from hospitals and laboratories. However, these systems are not as “early” as it is desired. Initiatives have been raised to research the issue and steps are being taken in many directions in order to model an “early detection” system. Preliminary experiments conducted give strong evidence that the non-specific syndrome based systems that have not been looked at before, may indeed be accurate enough to have a right to be considered. Complete framework has been developed that can serve as a prototype for such a system. It has been created to be flexible enough, to be able to other datasets that have time series properties and satisfy the condition for correlation with the particular epidemic that is to be predicted. The condition is only necessary to provide more confidence in the accuracy of the system. However, it is not necessary for the general purpose one. Given a time window enough for the basis signal to be established, any spikes can be examined for the purposes of detecting an abnormality.

6. Discussion and Future Work

It is apparent that the task at hand requires a highly accurate system. Numerous stages to the system are introduced in order to increase the reliability and make a claim about the lower bound on accuracy that can be achieved. The idea of smoothing the raw data in order to make the predictions more reliable and the fact that the current accuracy of the model can be explicitly calculated based on the previously seen data, gives a significant increase in confidence when making claims about the accuracy and reliability of the system. The idea of an absolute low bound on accuracy that can be achieved by smoothing the original time series will be researched further. If proved valid, it can be extended to many other applications involving time series data. It is also desirable to verify whether the mixing of stages would give an increase in the performance of the system. Intuitively it seems reasonable that better results can be achieved by using a combination of products or categories. Verification of this idea however, requires more data. The model should also be tested with the known periodicities removed to see if a significant improvement in accuracy can be achieved.

Acknowledgements

We would like to thank Rich Caruana, Galit Schmueli and Sebastian Thrun for their invaluable help with participating in brainstorming sessions and assisting in making the project stronger.

References

Aha, D. W. (1990). *A study of instance-based algorithms for supervised learning tasks: Mathematical, empirical, and psychological evaluations.* Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine.

Goldman B. (2000). *Inhalation Anthrax in Cecil Textbook of Medicine*. New York: Saunders Company.

Stephenson J. (1997). *Pentagon-funded research takes aim at agents of biological warfare.* JAMA;278:(373-375).

Airhart, M. (1997). Why do we get the flu most often in the winter? Are viruses more virulent in cold weather? *Scientific American*, Ask Experts section.

Weigend, A. and Gershenfeld, N. (1993). *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison Wesley.

Brockwell, P. and Davis, R. (1996) *Introduction to Time Series and Forecasting* Springer.

Chatfield, C. (1998) What is the best Method in Forecasting? Journal of Applied Statistics. 15:(19-38).

Brigham, E. (1988), *The Fast Fourier Transform and Its Applications*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Daubechies, I. (1992). *Ten Lectures on Wavelets* CBMS-NSF Series in Applied Mathematics 61, Philadelphia: SIAM

Graps, A. (1995). An Introduction to Wavelets. *IEEE Computational Science and Engineering*. Los Alamitos, CA: IEEE Computer Society.

Shimizu K. (1995). *Signal Detection Using the Wavelet Transform*

Nijs, V., Deimpe, M., Steenkamp JB. and Hanssens, D. (2001). *The Category Demand Effects of Price Promotions* Under Review at Marketing Science, second round.

Montgomery, A. (1997). *Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data*.

Boatwright P., McCulloch R., and Rossi P. (1999). *Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model*. (315-337).