Analyzing over-the-counter medication purchases for early detection of epidemics and bio-terrorism attacks

Anna Goldenberg

CALD

Carnegie Mellon University

anya@cs.cmu.edu

Abstract

Early detection of epidemics and bio-terrorism attacks is one of the most vital problems attempting to prevent deaths and provide for the opportunity of early treatment. There are many sources of data for analysis, though they are hard to get access to. However, there is very little real-time data that is readily available in substantial quantities at the first signs of an epidemic or a bio-terrorism attack. The main goal of this project is the investigation of the possibility of an early detection by analyzing trends of consumer purchases of over-the-counter medications. The idea is that by taking a large enough time frame of the regular over the counter medication purchases it might be possible to detect abnormalities early enough so that lives would be saved. Since we are interested in early detection or forecasting of purchasing behavior, time series prediction approaches are to be considered. It is expected that due to the nature of the problem, a combination of one or more techniques would have to be utilized. As a result of the project, it is desirable to have a system that would detect an outlier signal, an epidemic, in the regular over-the-counter medication purchases. The work will be evaluated based on different medical datasets, such as ICD9-code data. It is expected that the evaluation data has a delayed signal of an epidemic or an attack, due to the fact that it takes time for a person to seek medical care, so “shift and smooth” technique will be used to approximate the beginning of an epidemic. This work might have a significant impact on the evolution of early detection and consequently on computer-based surveillance (CBS) systems, since the main approach being explored now is coming from the medical field, such as reporting systems for clinicians, rather than evaluation of nonspecific syndrome data, such as over-the-counter medications.

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. There were many ideas and plans 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 known or an unknown virus is introduced to the environment, the course of its evolution and spread is up to an early detection system. Unfortunately, early detection systems are still in the stage of development.

The investigation of prevention systems is in the hands of the government officials. The development of such a system is a hard task to accomplish. A more feasible but not less desirable task is a creation of an early detection system. Though it seems that the task is simplified, 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? 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.

Several ideas were discussed by a large group of CMU and PITT researchers. It was proposed to try to collect the information from the web and look for spikes in the hits to particular sites, another idea was to use the device designed by BodyMedia (see www.bodymedia.com for more information) in order to track the body measurements to identify abnormalities. Unfortunately, the implementation of ideas requires substantial amounts of information that is hard to get in case of web sources or unavailable due to a small number of participants as in the BodyMedia program. On the other hand, among other valuable sources is the large supermarket dataset that contains the information about over-the-counter medication purchases. It is possible that the data is noisy, but the potential of discovering the first signals of seasonal flu that in turn might be an indication of a bio-terrorism attack (see Goldman 2000) might be enormous.

In this project I propose to perform time series analysis on the purchases of over-the-counter medications to establish regular trends in medication purchase behavior and thus to detect abnormalities early enough to prevent the epidemic from spreading or to provide the early treatment to the victims of the bio- attack. It is natural to assume that the data is highly periodic, for example some of the stores are closed at night, thus the amount of purchases during that time is going to highly drop, therefore daily periodicity can be expected. In fact, it is possible that there are several different periodic signals present in the data such as weekly, seasonal, sale and other periodicities. Each of them can be modeled separately for simplicity and then combined together for an accurate prediction. There are several time series prediction approaches that will be inspected in order to decompose the data signal into periodic sub-signals, including Fast Fourier Transforms and wavelets. I propose to use dynamic recurrent neural network (DRNN) to make predictions for each of the signals the original series was decomposed to.

There are several complications that might arise due to the nonspecific syndrome nature of the dataset used for such an unusual task as prediction of epidemics or bio-terrorism attacks. First of all, people buy over the counter medications not only in cases of illness, but also to replenish their home reserves in case of future sickness. Also, sales and mail-in coupons might be a factor in increased amounts of purchases of the over-the-counter medications. There can be other products that are correlated with sicknesses, for example increased purchase counts of orange juice for cold or flu. All of these are to be inspected during the course of the project. It is expected that this project will result in creation of a system that is based on the real time data and the history of purchases of over the counter medications and possibly a few other products will be able to early detect an abnormality in the purchasing behavior.

2. 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.

Most of the analysis performed on the supermarket data has also had a slightly 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.)

The subject of using nonspecific syndrome datasets such as evaluating the trends in over-the-counter medication purchases to detect epidemics or bio- attacks has not been widely explored up-to-date. However, the problem at hand can be viewed as a time series prediction from noisy data problem with a slight modification. In order to use as much relevant information in the data as possible, it is desired to take advantage of products other than medications such as orange juice, facial tissues, hot tea, honey etc. It was estimated that there are close to 2,000 of such products (accounting for different brands of the same product). It is clear that to inspect all permutations of selected items is too computationally expensive; hence, I propose to mine data for association rules as a first step in data analysis.

Association rules have been introduced in 1993 by Agrawal, Imielinski and Swami. The idea had since received a lot of attention due to the easy intuitive understanding. The algorithms find rules of type X => Y (c%), i.e. whenever a set of items X is present in the basket it is probable that set of items Y is also present with confidence c%. The algorithm will be formally described in the Approaches section. There are many algorithms that have been introduced to mine the association rules. The first and the most commonly used algorithm is *Apriori* introduced by Agrawal at el, 1994. The algorithm is efficient and can be easily adapted to the particular task at hand. There were several modifications to the Apriori algorithm that were introduced later, some of them are AprioriTID, where the algorithm works with data candidates rather than with raw data. Other algorithms include AprioriHybrid, which is the hybrid of Apriori and AprioriTID, DIC that smoothes the bounds between counting and generating candidates. The greater and more detailed overview of association rules mining methods is available in Hipp, et. Al 2000. The authors in the paper have done extensive method validation and comparison and have established that all existing methods perform comparably well and the performance may vary based on the nature of the dataset.

There is a great variety of overview literature published on the subject of time series analysis (TSA). TSA has a long history dating before 1920s, when the forecasting was done by simply extrapolating the series through a global fit in the time domain. With the Yule's invention of autoregressive technique in 1927 and for almost fifty years following his discovery the paradigm was that of linear models driven by noise. (Weigend and Gershenfeld, 1993; Chatfield, 1989) The linear models however tend to be oversimplified and thus fail in predictions for even moderately non-linear realms. Starting 1980s more complicated mixtures of linear models as well as non-linear approaches were explored. Some of them are: state-space reconstruction by time-delay embedding, coming from differential topology and dynamical system fields proposed by Packard et al. (1980), Takens (1981) and Sauer et al. (1991); and neural networks coming from Machine Learning area. Lapedes and Farber (1987) trained their nonlinear network to emulate the relationship between output (x_{t+1}) and inputs ($x_{1\ldots}x_t$) for computer generated time series, and Weigend, Huberman and Rumelhart (1990, 1992) addressed the issue of finding networks of appropriate complexity for predicting observed real-world time series. Both methods have been tried on several different datasets yielding good performance and producing accurate predictions. For the purposes of the project it is desirable to concentrate on the high order interaction models. Neural Network by its nature does not limit the order of the interactions, just the total number of them and learns to select an appropriate combination of inputs, which is more relevant to the project at hand, hence the NN approach will be used for predictions.

The described linear and non-linear methods can be better predictors if instead of forecasting the original data series that maybe rather complicated, they were used to predict several simpler signals. Since data can be assumed to contain several periodic signals, it would be useful to preserve the information in the data by treating those signals differently, i.e. by applying methods described above to each of the components comprising the original data separately. There are several techniques that can be used to decompose an original waveform into a number of components. One of the most common techniques is the Fourier Transform (FT), which in essence decomposes or separates a waveform or function into sinusoids of different frequencies that sum to the original waveform (Brigham, 1988). Several studies have shown however that FT does not perform well when the signal is localized in space (Daubechies, 1992) or cannot be approximated by a sinusoidal wave (such as square wave signal). The improvement on FTs are wavelets that avoid the problem by using dilated and translated wavelets rather than sines and cosines as its basis functions (Shimizu 1995). Since the data in the problem may have some simple periodicities as well as localized signals, we propose to use both Fourier transforms and wavelets to locate those.

3. Approach

I propose to apply the time series prediction methods to the supermarket data to evaluate the over-the-counter medication and other related products purchasing trends with the following modification. As a first stage of the process, I am planning to use association rules in order to find several independent or a combination of products that might contain the most signal. It is then possible to apply Fourier transforms and wavelet technology to decompose the signal into several simpler signals. For prediction purposes, I propose to use Temporal Recursive Neural Network, since it has proved to be a valuable prediction tool in application to many datasets. I plan to compare the predictions by TRNN to the ones obtained by fitting to the sample ACVF. The evaluation of success will be done based on different datasets, such as ICD-9 code data. Since the evaluation datasets might have intrinsic latency due to the medical nature of them, I propose to use a “shift and smooth” technique that is further to be discussed in the approach validation techniques section.

3.1. Association Rules

It is important to notice that the non-specificity of the data to the cause of the problem renders it to be a non-classical TSA problem. The overall data is very noisy, however we should take into account the fact that there are many variations and combinations of the medication and non-medication type products that might have a strong intrinsic signal of abnormal behavior in the data. In fact, the combination of the products containing the most signal might vary based on the geographical location or demographic differences. It would be useful to keep this information in the data. It is quite possible that tracking of individual or a combination of products might give a better signal than the aggregate counts over all products. For, example if there was an expensive medication that cured flu, then during non-flu seasons such medication would be purchased very rarely where as at the height of the flu season it was going to peak high. If such medication existed and the predictions for flu were based solely on it, the signal would have been clearer than if the aggregate counts of all cough, cold and flu medications were taken into account. This problem can be resolved by recursively trying different 'meaningful' combinations of products. I propose to determine the meaningfulness of the combination by association rules. We can assume that all of our products are independent and make our predictions based on the ones or pairs or even triplets that occur more than others based on the independence assumption.

It was earlier discussed that it would be beneficial to use individual and combination of products rather then overall aggregates of data to look for the first appearances of abnormal signal in the supermarket data for the purposes of determining the beginning of epidemics or events of bio-terrorism attacks. The assumption is further confirmed by preliminary data inspection.

I propose to use Apriori association rule algorithm to mine the supermarket health related data for association rules. At this point I would like to formally introduce the idea of association rules:

Let $I=\{x_1,\ldots,x_n\}$ be a set of items
$X \subseteq I$ with $k=|X|$ is a k-itemset
D is a multi-set of subsets of I
$T \in D$ is a transaction.

It is said that $T \in D$ supports an itemset $X \subseteq I$ if $X \subseteq T$ holds. An Association rule is an expression X => Y, where X, Y are itemsets and $X \cap Y = \emptyset$ holds. The confidence of this rule is defined as conf(X=>Y) = supp($X \cup Y$)/supp(X), where supp(X)=$|\{T \in D | X \subseteq T\}|/|D|$. i.e. the support of X is a fraction of T supporting an itemset X with respect to D and supp(X=>Y)=supp($X \cup Y$).

The main challenge of concern is the exponential growth of |I|. It is undesirable to search such a space; it is too computationally expensive and time consuming. The problem of searching the space of association rules is described in detail by Agrawal et al. 1993, 1994 and then later summarized by Hipp et al, 2000.

3.2. Signal Decomposition

I propose to decompose the data using the Fourier Transforms and 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 may be able to predict different aspects of the data whereas the merged signal has these effects embedded in a hard to recover format.

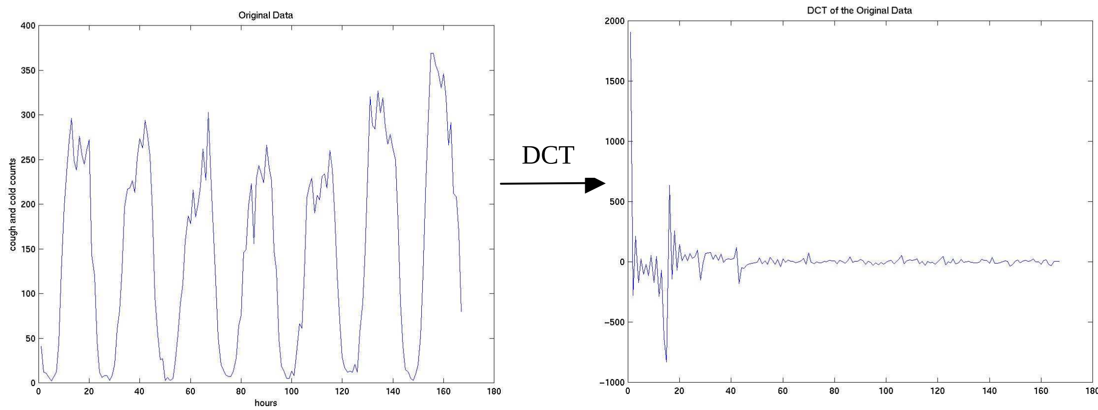
Clearly, when thinking about over-the-counter purchases it is natural to expect daily, monthly, yearly, etc. periodicities. If such periodicities in fact exist in the data, they should be found and approximated by Fourier Transforms. FT decomposes or separates a waveform or function into sinusoids of different frequency that sum to the original waveform. It identifies or distinguishes the different frequency sinusoids and their perspective amplitudes. The FT of f(x) is defined as

$$F(s) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi xs} dx$$

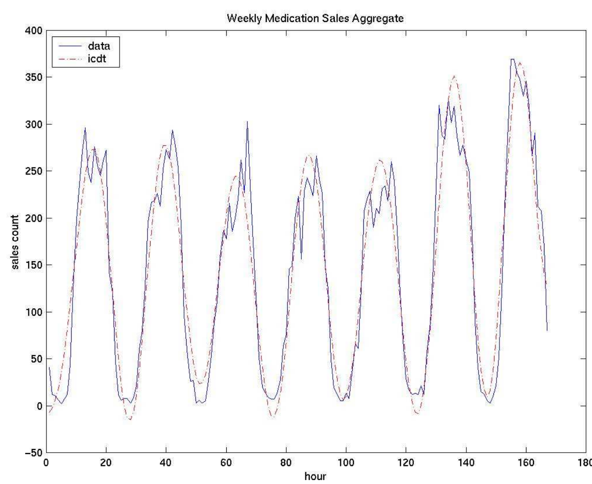
The inverse of an FT is also an FT: $f(x) = \int_{-\infty}^{\infty} F(s) e^{i2\pi xs} ds$, so that $f(x) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x) e^{-i2\pi xs} dx \right] e^{-i2\pi xs} ds$.

as long as the integral exists and any discontinuities are final.

Preliminary results on the week short sample of data show that the assumption of periodicities is valid and that FT can be used for establishing periodicities in the data:



If the inverse of the full decomposition were used to predict the original data, it would be able to recreate the original data with 100% accuracy. However that would mean overfitting, i.e. since the data is noisy, FT has modeled noise as well as the periodic signal. So, it is of interest to concentrate on whether FT captures predicted periodicities. For this purpose the DCT (discrete cosine transform – real part of the FT) signal was analyzed and the frequencies corresponding to weekday – weekend and daily phenomena were extracted (threasholding has also revealed importance of those periodicities):



From the graph it is evident that the periodicities were captured quite well. Since for the trial purposes I decided to concentrate on daily periodicities, the function appears to be smooth. However it is possible to trace hourly periodicities as well by looking at higher periodicity values.

It is evident that FT is a good tool for capturing periodicities. Such conclusions are supported by preliminary results. But it is expected that not all signals that we would like to decompose our data to can be well described by a composition of sine and cosine waves. For example, the effect of positive and 'negative' (competitor sales) must be taken into account and such signal maybe a square shape function: sale is present and the magnitude of the sale vs no-sale event. Since it is desired to retain information about such signals, I propose to use wavelets to capture those.

Wavelet Transforms (WT) are very similar to the FTs. While the FT uses sines and cosines as its basis functions and that being a shortcoming when trying to model square or localized signal type waves, WT uses dilated and translated wavelets as its basis functions. The wavelet has to 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 basic 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). The integral wavelet transform is defined as follows:

$W_\psi f(b,a) = | a |^{1/2} \int_{-\infty}^{\infty} f(x)\psi\left(\frac{x-b}{a}\right)dx$, $f \in L^2(\Re)$, with $a=2^{-j}$, $b=k/2^j$. Though there are no preliminary results yet, the wavelet method of decomposing a signal has been successful in application to similar retail type datasets (Daubechies, 1993).

3.3. Prediction Stage

There are several methodologies to predict data based on the decomposed signals. If the dataset was periodic and linearly decomposable, the prediction would be a simple $x_{t+1} = \sum_{i=1}^{N} \alpha_i s_i(t+1)$, where N is the number of the signal-components and $s_i(t+1)$ is the value of the function modeling signal s_i at time t+1. However, there is no reason to assume linear decomposability of the data. In fact, it was found that most real datasets could not be modeled by linear approximations (Gershenfeld and Weigend, 1993). Hence, it would be natural to select a nonlinear method that makes no assumption about the nature of the data, but rather is a powerful prediction tool independent of the dataset. I propose to use artificial neural network for making predictions of each of the signals decomposed as discussed in 3.2. It is possible to use feedforward backpropagation NN, where the inputs are counts at discrete times 1..t and the output consists of one neuron which is a predicted value x_{t+1}. Such system is easy to implement (various implementations already exist as part of data mining packages). However, it does not take into account the temporal information intrinsic to the data. It has been shown (Back & Tsoi, 1991) that a special type of neural network may give a better performance than a general feedfoward backpropagation network. To be more precise, it is a connectionist network with internal delay lines or a neural network architecture modeling synapses as FIR (Finite Impulse Response) linear filters. The internal dynamics is brought to the basic neural network by replacing each static synaptic weight by a FIR filter:

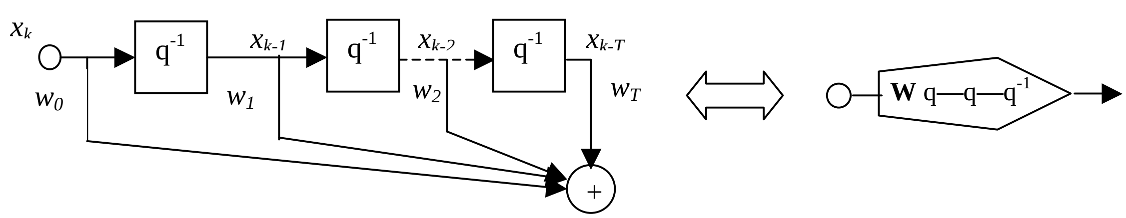


Figure 2. FIR filter model, where q^{-1} represents a unit delay operator $x_{k-1} = q^{-1}x_k$

The approach is to make use of the general feedforward backpropagation NN first, since it is readily available in standard data mining packages and then possibly modify the code to account for temporal factors.

3.3 Approach Validation Techniques

Validation is an important and a challenging part of the project. Fortunately or unfortunately, there is very little data on bio-terrorism attacks available for analysis. There is more data on epidemics however the labels are not clear. From looking at the data it is possible to infer whether the epidemic has occurred during the given time frame. It might even be feasible to make predictions about the peak of an epidemic. However, there is no clarity about the beginning of an epidemic. Since the system proposed in the project is an *early* detection system, it is important to have means of evaluating the early stages of an epidemic in the data long before it reaches the peak. There are several ways to deal with the problem. I propose to use "shift and smooth" technique in application to other datasets in order to validate the project. Several medical datasets are available for analysis. They are accurate and are used in reporting the disease statistic to government agencies. However, they fail in early predictions of bio- attacks and epidemics of quickly spreading epidemics because disease may be long on its way before medical community gets sufficient number of patients to get alarmed. The same reason is valid when making an assumption of the latency in the appearance of the signal in the medical data. The source of the information is nonetheless valid and very accurate. I therefore propose to use this data with a slight modification. The idea is to shift the peak to the left but since we are less certain about peak occurring before the set date, scale the data by the confidence percentage. For example, it can be asserted that the epidemic has happened 2 days prior with only 80% confidence, so the data shaped should be moved 2 days prior to the original location and scaled by a factor of 0.8. We should stop the shifting and scaling process when there is no information available about the possibility of an epidemic happening. The obtained set of functions – shapes, should then be smoothed over and the new function should be used to evaluate the early detection system proposed in the project. Due to the diversity of datasets available (school absenteeism, lab results, respiratory counts, etc), it is expected that there would be enough information in order to get an accurate validation system.

It is also possible that the data reveals patterns of abnormality that can be attributed to an epidemic, i.e. unusual high fluctuation of purchasing behavior when no other factors, such as sales, release of a new ad, are present. Then, the cross validation method can be applied. The two-year data can be split into one year training set and one year testing set, so the system will be validated on the second year testing set.

4. Work Plan

Week 1: applying association rules to the datasets. Looking for interesting patterns based on location.
Use an existing Apriori implementation, tailoring it to the particulars of the dataset if needed.
Week 2: applying Fourier Transforms and Wavelets to decompose signals. Use Matlab or Splus.
Week 3: using neural network to make predictions for each of the signals and generate a global prediction
Make the system recurrent, so that the predictions could be made as far into the future as needed
Use existing code for NN with modifications to incorporate the delay information.
Week 4: validate the system and write the report

5. Summary

This project proposes an alternative to conventional Computer Based Surveillance Systems (CBS), by concentrating more on the non-specific syndrome data such as over-the-counter medication purchases. The idea is based on the time series prediction analysis of the series of medication sales to recover the signal of epidemics or possibly the bio-terrorist attack. The difference between the common time series analysis application and the proposed approach is in integration of association rules and the signal decomposition technique. As a result of the project, it is desirable to have a system that would reliably detect a signal of abnormality in the regular over-the-counter medication and other related products purchases. This work will have a significant impact on the evolution of early detection systems, due to the uniqueness of its nature.

References

[1] Torok T.J., Tauxe R.V., Wise R.P., Livengood J.R., Sokolow R., Mauvais S., et al. A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. JAMA;278:389-95, 1997.

[2] Montgomery, A. Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. 16, 4, 315-337, 1997

[3] Weigend, A.S. and Gershenfeld, N.A., "Time Series Prediction: Forecasting the Future and Understanding the Past", Addison Wesley, 1993.

[4] Markus, C., Gupta, S., Porter, J. C., Staelin, R., and Wittink, D.R. Using Market-Level Data to Understand Promotional Effects in a Nonlinear Model, 34 (August), 322-334

[5] Boatwright P., McCulloch R., and Rossi P.E. Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model, 1999

[6] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000

[7] Chatfield, C. What is the best Method in Forecasting? J. Appl. Stat. 15: 19-38,1988

[8] Goldman B, Inhalation Anthrax in Cecil Textbook of Medicine. Saunders Company. New York. 2000

[9] Stephenson J. Pentagon-funded research takes aim at agents of biological warfare. JAMA;278:373-375, 1997

[10] Associated Press. Lab meets with residents over bacteria release. New York, NY: Associated Press; July 9, 1999.

[11] Jacobs P. Attack of the killer microbes. The Pentagon is funding research into novel ways to thwart bioterrorists. Advances could pay o commercially as well. Los Angeles Times. August 19, 1999.

[12] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), Washington, USA, May 1993.

[13] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994

[14] Ramakrishnan Srikant and Rakesh Agrawal. Mining Generalized Association Rules. In Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, September 1995.

[15]S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Proc. Of the ACM SIGMOD Int’l Conf. On Management of Data (ACM SIGMOD ‘97), 1997

[16] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Series in Applied Mathematics 61, SIAM, Philadelphia, 1992

[17]Brigham, E. Oren, 1988, The fast Fourier Transform and Its Applications, Englewood Cliffs, NJ: Prentice-Hall, Inc., 448pp.

[18]Shimizu K., Signal Detection Using the Wavelet Transform, 1995.

[19] Back, A. & Tsoi, A. C. FIR and IIR synapses, a new neural network architecture for time-series modelling', Neural Computation 3, 375—385, 1991.

[20]H. Mao, R. McLaren, "Modelling and Control of Dynamic Nonlinear Systems Using DRNN", Proc. 1996 Intern. Neural Network Society Annual Meeting, San Diego, California, pp. 137-142, 1996.

[21] Packard, N.H., J.P. Crutchfeld, J.D. Farmer, and R.S. Shaw. Geometry from a Time Series. Phys Rev Lett 45(9): 712-716, 1980.

[22] Takens, F. Detecting Strange Attractors in Turbulence. In Dynamical Systems and Turbulence. Lecture notes in Mathematics, Vol 898, 336-381. Warwick, 1980.

[23] Sauer. T., J.A. Yorke and M. Casdagli. Embedology Journal of Statistical Physics 65(3/4):579-616, 1991.

[24] Lapedes, A. and R. Farber. Nonlinear Signal Processing Using Neural Networks. Technical Report No. LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, NM.

[25] Weigend, A.S., B.A. Huberman, and D. E. Rumelhart. Predicting the Future: A Connectionist Approach. Intl. J. Neur. Sys 1: 193-209, 1990.