Comparing Syndromic Surveillances using Two Aspects: Emergency and Telehealth Data Sources

El Sayed Mahmoud and David Calvert

Computing and Information Science, University of Guelph, Guelph, Ontario, Canada

Objective

The objective of this paper is to examine the utility of Emergency Department and Telehealth data for Syndromic Surveillance. This works attempts to minimize false outbreak detection. It also demonstrates that these two data sources contain independent information which is useful for outbreak detection.

Background

For this work, an aberration is recognized as a change in distribution or frequency of data relative to the historical or recent data. Many analytical models can be used to do this such as rule based anomaly pattern detection, computer modeling and Artificial Neural Networks [1]. This work will compare back propagation (BP) and the support vector machine (SVM). These are two learning algorithms use the historical data of the outbreaks as a training set to build a model of this data. A sliding window of several days is used with the sequence data to train the neural networks (NN). Emergency Department (ED) and Telehealth (TH) data is used in this work. The utility of these two datasets are examined.

Methods

The ED and TH datasets contain the daily counts of complaints related to respiratory illness. Each dataset contains 441 days starting from 01/04/2004. A data simulator was used to generate a hundred years of data based on the statistical characteristics of the original data. Two types of outbreak patterns were injected into the training set at a rate of one outbreak per year. These simulated fast and slow outbreaks. The BP network was trained for 10000 iterations. A threshold was used to identify the outbreaks. Various thresholds were tested to reduce the number of false positives. The accuracy of detecting the outbreaks in the test set and time to detect the outbreak are calculated. The SVM was tested using the same datasets.

Statistics and clustering were used to determine if these datasets contain complimentary information. The two datasets were normalized based on the mean and the standard deviations, clustered to fifteen clusters. Each cluster covers 30 days. The average normalized count per cluster was calculated then graphed on two dimensional charts to show overlap between these clusters.

Results

For BP the average accuracy of 100 experiments using the Telehealth data was 99.73% and 43% using the Emergency data for the same period. The average time to detect the outbreak using the Telehealth data was 13.7 days and 17.6 days using the Emergency data. See figure (1).

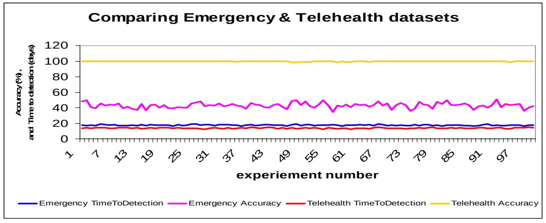
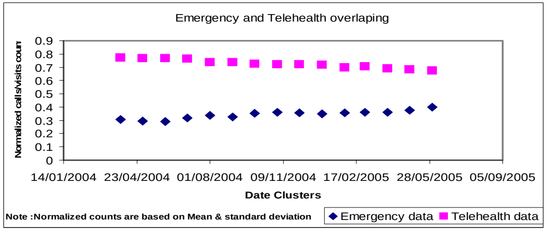


Figure1:comparing emergency & Telehealth data

For SVM the accuracy using TH dataset was 96.14% and 65.71% using the ED dataset for same period. No overlap was observed between the clusters of the two datasets as shown in figure (2).



Note :Normalized counts are based on Mean & standard deviation

Figure2:clusters of Emergency & Telehealth data

Conclusions

In these experiments, the TH data is more accurate and sensitive than the ED data for detecting outbreaks. This due to characteristics of the TH data. It is less variable than the ED and therefore easier to analyze and achieve consistent results.

These two data sources contain independent information which is useful for syndromic surveillance, The TH data contains complaints which are not in ED and the reverse is true. Figure2 demonstrates there is no overlap observed between these two datasets.

References

[1]Detection of disease outbreaks in pharmaceutical sales: neural networks and threshold algorithms, Guthrie, G.; Stacey, D.A.; Calvert, D.; Edge, V. Neural Networks, 2005. IJCNN.

[2] Ben Y Reis and Kenneth D Mandl , 2003,Time series modeling for syndromic surveillance.

[3] Artificial Intelligence A Modern Approach 2nd edition –Start Russell & Peter Norvig- ch(20)

[4] Mitchell, T.. Machine Learning. McGraw-Hill, New York, NY ,1997.